DAMA-DMBOK

СВОД ЗНАНИЙ ПО УПРАВЛЕНИЮ ДАННЫМИ

ВТОРОЕ ИЗДАНИЕ

Посвящается памяти
ПАТРИЦИИ КУПОЛИ
(25.05.1948 — 28.07.2015),
посвятившей жизнь делу управления данными
и внесшей неоценимый вклад в настоящую публикацию

DAMA-DMBOK

DATA MANAGEMENT BODY OF KNOWLEDGE
SECOND EDITION

DAMA International

Technics Publications
BASKING RIDGE, NEW JERSEY

DAMA-DMBOK

СВОД ЗНАНИЙ ПО УПРАВЛЕНИЮ ДАННЫМИ
ВТОРОЕ ИЗДАНИЕ

DAMA International

Издательство «ОЛИМП–БИЗНЕС»
МОСКВА, 2020

УДК 004.6:087.7
ББК 92
D17

Научное редактирование осуществляли:
— коллектив специалистов компании «Юнидата» под общим руководством
Руководителя направления Методологии компании «Юнидата»
к. т. н. Николая Владимировича Скворцова
и Chief technical officer компании «Юнидата» Алексея Владимировича Цырюльникова;
— Генеральный директор компании BSSG Юрий Борисович Клочко
и Руководитель направления консалтинга компании BSSG, к. т. н. Андрей Сергеевич Ларионов

D17 DAMA-DMBOK : Свод знаний по управлению данными. Второе издание / Dama International [пер. с англ. Г. Агафонова]. — Москва : Олимп–Бизнес, 2020. — 828 с.: ил.

ISBN 978-5-9693-0404-8

Главная задача книги — определить набор руководящих принципов и описать их применение в функциональных областях управления данными. Издание всесторонне описывает проблемы, возникающие в процессе управления данными, и предлагает способы их решения. В нем подробно описаны широко принятые практики, методы и приемы, функции, роли, результаты и метрики.

«DAMA-DMBOK: Свод знаний по управлению данными. Второе издание» предоставляет специалистам по управлению данными, ИТ-специалистам, руководителям, преподавателям и исследователям обширный материал для совершенствования работы с информационными активами и корпоративными данными.

УДК 004.6:087.7
ББК 92

Оригинальный дизайн обложки: Lorena Molinari

A member of: BPR Business Publishers Roundtable.com

ISBN 978-5-9693-0404-8

Оглавление

Список иллюстраций

Список таблиц

Вступительное слово компании «Юнидата»

В мире российской технической литературы практически отсутствуют полноценные описания современных методик управления данными в масштабах организации или объединения организаций. И это при том, что отечественный рынок ИТ является одним из самых развитых. У нас много талантливых специалистов в ИТ и других областях, которые занимаются непосредственно работой с данными: создают, преобразуют, используют и повышают их ценность. Каждый такой специалист работает в определенном отделе, департаменте, бизнес-единице, и очень часто возникают ситуации, когда информационные потоки между разными подразделениями синхронизированы плохо.

Отсутствие практических рекомендаций и методологий в области управления данными с точки зрения развития организации создает вполне ощутимую системную проблему. Многообразие используемых данных ведет к неконтролируемым инцидентам, связанным с их низким качеством (неактуальность, ошибки, разрозненность), постоянному росту затрат на ИТ-инфраструктуру и серьезному кадровому голоду.

Наша компания долгие годы занимается проблематикой управления данными на отечественном и международных рынках. Изучив весь ассортимент методических материалов в этой области, предлагаемых международными профессиональными сообществами, и детально проанализировав применимость документов подобного рода в российской практике, мы не случайно остановили свой выбор именно на методологии DMBOK2 международной ассоциации DAMA, с которой сотрудничаем уже много лет.

К DMBOK2 у любого специалиста по управлению данными особое отношение. Именно это руководство стало настоящей настольной книгой для многих наших коллег. Оно представляет собой одну из наиболее удачных попыток описания современных методов, которые хорошо себя зарекомендовали при внедрении практик управления данными как в зарубежных, так и в российских организациях.

Мне очень приятно, что сейчас вы держите в руках первое российское издание DMBOK2. Работа по переводу проделана поистине титаническая — ведь устоявшейся терминологии на русском языке просто не существует. Именно поэтому «Юнидата», одной из весомых задач которой является популяризация сферы

управления данными, взялась за этот проект. Отдельное спасибо хотим сказать всем экспертам сообщества Unidata Community. Это открытая площадка, на которой специалисты в области управления данными делятся мнениями относительно самых актуальных вопросов отрасли, обсуждают новые материалы, терминологию и пр. Приглашаем вас вступить в ряды этого сообщества: заходите на сайт community.unidata-platform.ru и регистрируйтесь!

В заключение добавлю: мы прекрасно осознаём, что варианты перевода многих терминов могут у кого-то вызвать определенные вопросы; но перед вами — первая редакция, и мы планируем продолжать работу над новыми редакциями этого фундаментального издания.

С уважением,
Сергей Кузнецов,
Генеральный директор компании «Юнидата»

Вступительное слово компании BSSG

Руководство данными (Data Governance) в качестве новой корпоративной функции только-только набирает обороты в российских компаниях, открывших ее для себя в контексте реализации амбициозных программ цифровой трансформации. В то же время парадигма руководства данными существует уже более 30 лет, и за это время крупнейшие международные корпорации успели не только внедрить ее у себя, но и накопить значительный практический опыт в реализации ее отдельных элементов. Так почему же российские компании так долго не проявляли интереса к управлению данными на корпоративном уровне?

Полагаем, тому есть две причины. Во-первых, до последнего времени вопросы управления данными практически не регулировались в России на законодательном уровне; а во-вторых, что гораздо важнее, на протяжении многих лет все вопросы управления данными в корпоративном сегменте воспринимались как прикладная задача ИТ-функции, которая с трудом воспринималась как ценность для бизнеса.

Сегодня можно с уверенностью заявить, что обе эти причины потеряли актуальность. Революция больших данных показала, что крупные организации могут извлечь огромную выгоду из информации, созданной и накопленной за много лет. Стало очевидным, что без умелого руководства корпоративные озера данных превращаются в цифровые болота. Кроме того, регулирование в области защиты информации привело к тому, что руководство данными стало фактически обязательным.

Эта история началась в середине 1970-х с появлением в Лос-Анджелесе первого неформального кружка специалистов по данным. Раз в неделю энтузиасты стали собираться за обедом для обсуждения трудностей, возникающих в области электронной обработки данных (Electronic Data Processing). Принято считать, что благодаря этим собраниям и зародилась профессиональная ассоциация специалистов по управлению данными, которую в 1980 году официально зарегистрировали как добровольную международную некоммерческую организацию — DAMA (Data Management Association) International. Именно эта организация с 1980 года начала систематизировать знания и опыт различных организаций в области управления данными и в 2009 году выпустила первое сводное руководство по управлению данными — DAMA-DMBOK (Data Management Body of Knowledge).

Появление DMBOK стало важнейшей вехой в истории индустрии данных, поскольку

позволило перейти к единой организационной модели. Руководство DMBOK сформулировало основные виды деятельности по управлению данными, определило роли и функции участников, описало типовые процессы, что в итоге позволило интегрировать распоряжение данными в общую систему управления корпорацией.

Книга стала основным ресурсом для подготовки специалистов к международной сертификации на степень CDMP (Certified Data Management Professional), что привело к формированию новых профессий: менеджер по управлению данными, распорядитель данных (Data Steward) и директор по данным (CDO).

За 10 лет существования «Свода знаний» вышло его второе издание — DMBOK2, которое сфокусировалось на важнейших и актуальнейших задачах: этике управления данными и культуре их использования. Массовые нарушения в области сбора и хранения данных, а также многочисленные утечки информации наносят серьезный ущерб репутации таких гигантов, как Facebook и Alphabet, и снижают их рыночную стоимость. Руководство этих и подобных им компаний наконец-то осознало тот факт, что данные являются не только источником обогащения, но и огромным риском.

Мы гордимся тем, что наша компания выступила одним из инициаторов проекта по переводу на русский язык и изданию DMBOK в России. Реализуя множество проектов разного масштаба в области управления корпоративными данными в российских и международных компаниях последние 15 лет, мы можем отметить одну очень важную тенденцию. Сегодня компании-лидеры переходят от локального решения проблем к системной и долгосрочной работе с данными на общекорпоративном уровне, оперируя ими как ценнейшим активом. Консолидируются усилия по внедрению единых каталогов данных, снабженных бизнес-глоссарием для связывания знаний о предметной области в бизнес-слое со знаниями о данных в информационных системах-источниках. Это позволит детально картировать потоки информации от сырых данных в транзакционных системах через кластеры больших данных до потребителей в BI-системах. Ведь эффективная работа AI-алгоритмов и реализация принципов SSA (Self Service Analytics) для быстрого и во многих случаях автоматизированного принятия бизнес-решений в условиях многокритериальной и быстроменяющейся рыночной среды возможны только на единственной версии правды.

DMBOK позволяет лидерам цифровой трансформации быстро настраивать Data-мышление, чтобы удерживать правильные фокусы для ускоренного формирования Data-инфраструктуры в компании и предоставлять Data-сервисы для устойчивого развития бизнеса в цифровую эпоху. Кроме того, Data-компетенции становятся ключевыми для специалистов практически во всех бизнес-функциях.

Мы уверены, что эта книга будет полезна как начинающим специалистам, так и опытным руководителям по данным (CDO), которые уже появились в отечественных компаниях и реализуют прорывные проекты по преумножению нового вида ценности для бизнеса — упорядоченных корпоративных данных.

Юрий Клочко,
Генеральный директор компании BSSG

Андрей Ларионов,
к. т. н., MBA, Руководитель направления консалтинга компании BSSG

Предисловие

DAMA International рада представить второе издание Руководства DAMA к своду знаний по управлению данными (DAMA-DMBOK2). Со времени публикации первого издания (2009 г.) в области управления данными произошли значительные изменения. Руководство данными (Data Governance) стало стандартной практикой для многих организаций, новые технологии обеспечили возможность сбора и использования «больших данных» (неструктурированных или полуструктурированных данных в самых различных форматах), а осознание важности соблюдения этических принципов при обращении с данными растет вместе с нашей способностью исследовать и применять огромные объемы данных и информации, производимых в процессе повседневной деятельности.

Эти перемены волнуют и захватывают. При этом они выдвигают всё новые и новые требования к нашей профессии. В ответ на меняющуюся ситуацию DAMA провела пересмотр своей рамочной структуры управления данными (колеса DAMA — DAMA Wheel), добавив ряд дополнительных деталей и уточнений, а также расширив круг вопросов, охватываемых DMBOK.

- Доработаны и обновлены контекстные диаграммы по всем областям знаний (Knowledge Areas).
- Тема интеграции и интероперабельности данных добавлена в качестве новой области знаний, чтобы подчеркнуть важность этого аспекта управления данными (глава 8).
- Этике обращения с данными в новом издании посвящена отдельная глава в силу признания всевозрастающей необходимости этического подхода к управлению данными, какую бы из его областей мы ни рассматривали (глава 2).
- Руководство данными описано и как отдельная функция (глава 3), и в контексте его взаимоотношений с каждой из остальных областей знаний.
- Аналогичный подход применен в отношении управления организационными изменениями. Эта тема вынесена в самостоятельную главу (глава 17), а также дополнительно рассматривается в главах, посвященных отдельным областям знаний.
- Новые главы, посвященные большим данным и науке о данных (глава 14), а также оценке зрелости управления данными (глава 15), помогут организациям лучше понять, куда они стремятся, и предлагают инструменты для достижения намеченных целей.

- Второе издание также включает набор заново сформулированных принципов управления данными, следование которым предоставляет организациям возможность управлять своими данными эффективно и извлекать ценность из имеющихся в их распоряжении информационных активов (глава 1).

Мы надеемся, что DMBOK2 послужит полезным ресурсом и руководством для профессионалов в области управления данными во всем мире. При этом мы в полной мере отдаем себе отчет в том, что наша книга — лишь отправная точка. Реальный прогресс достижим лишь по мере накопления опыта практического применения изложенных здесь идей. Но ведь DAMA для того и существует, чтобы помогать своим членам непрерывно учиться, делясь идеями, результатами осмысления тенденций и анализа проблем, а также используемыми решениями.

Сью Гьюенс,
Президент DAMA International

Лаура Себастьян-Коулман,
Главный редактор изданий DAMA International

Справка

DAMA (Data Management Association) International — основанная в 1980 г. в Лос-Анджелесе (США) и действующая с 1988 г. в статусе международной некоммерческой организации добровольная профессиональная ассоциация специалистов по управлению данными (dama.org).

Управление данными

1. ВВЕДЕНИЕ

Многие организации отдают себе отчет в том, что имеющиеся в их распоряжении данные — жизненно важный корпоративный актив. Данные и информация могут обеспечить им более глубокое представление о собственных клиентах, продуктах и услугах, что, в свою очередь, помогает в развитии инноваций и достижении стратегических целей. Вопреки такому признанию, пока лишь немногим компаниям удается эффективно управлять данными как активом и непрерывно извлекать из них практическую пользу (Evans and Price, 2012). Извлечение выгоды не происходит случайным образом — для этого требуются решительность и целеустремленность, планирование и координация, а также приверженность, управление и лидерство.

Управление данными (*Data Management*) — это разработка, выполнение и контроль выполнения политик, программ и практик предоставления, проверки, защиты и повышения ценности данных и информационных активов на протяжении всего их жизненного цикла.

Профессионал в области управления данными — любое лицо, рабочая деятельность которого направлена на обеспечение какого-либо аспекта управления данными в интересах достижения стратегических целей организации — от технической поддержки управления данными на протяжении их жизненного цикла до контроля соблюдения правил доступа к данным и их использования. Профессионалы в области управления данными формально могут занимать самые различные должности — от чисто технических (например, администратор баз данных, сетевой администратор, программист и т. п.) до стратегически важных для бизнеса (например, распорядитель данных, информационный стратег, директор по данным и т. п.).

Спектр деятельности в сфере управления данными широк и разнообразен. Сюда относится всё, начиная с принятия логически обоснованных решений о том, как извлечь стратегическую ценность из имеющихся или предполагаемых к получению данных, до технического развертывания и обслуживания баз данных. Таким образом, управление данными требует как технических, так и управленческих (или деловых) навыков. В связи с этим ответственность за управление данными распределяется между менеджерами и специалистами по ИТ. И представителям двух этих категорий необходим общий язык для эффективного сотрудничества по обеспечению организации высококачественными данными, соответствующими ее стратегическим целям и задачам.

Данные и информация — это не только лишь активы, в которые организации вкладываются в расчете на получение выгоды в отдаленной перспективе. Они жизненно необходимы организациям для обеспечения каждодневной операционной деятельности. Именно потому данные и информацию и называют «валютой», «животворящей кровью» и даже «новой нефтью» информационной экономики[1]. Вне зависимости от того, получает организация отдачу от аналитики или нет, без данных никакого бизнеса не построить.

В помощь профессионалам в области управления данными DAMA International (международная Ассоциация управления данными) и выпустила данную книгу — второе издание *Руководства DAMA к своду знаний по управлению данными* (*DMBOK2*). За основу взято первое издание 2009 года, заложившее фундамент научно обоснованного развития и созревания профессии.

В настоящей главе излагаются общие принципы управления данными, обсуждаются трудности с практической реализацией этих принципов и предлагаются подходы к преодолению трудностей. В ней также описана рамочная структура управления данными DAMA (DAMA Data Management Framework)[2], предоставляющая контекст для работы, выполняемой профессионалами в области управления данными в рамках различных областей знаний по управлению данными.

1.1 Бизнес-драйверы

Обладание информацией и знаниями — ключ к преимуществу над конкурентами. Организации, располагающие надежными, высококачественными данными о своих клиентах, продуктах, услугах и операциях, способны принимать более эффективные решения, нежели те, которые подобными данными не располагают или не имеют гарантий их достоверности. Неумение управлять данными сродни неумению распоряжаться капиталом. Результатом становятся убытки и упущенные возможности. Главным драйвером управления данными выступает возможность получать выгоду от информационных ресурсов, имеющихся в распоряжении организации. А потому управлять данными следует столь же эффективно, как финансовыми и материальными активами.

---

[1] Поищите в Google, например, «данные как новая валюта» или «данные как новая нефть», — и убедитесь сами, как много публикаций посвящено этим понятиям. — *Здесь и далее, если не указано иное, в сносках даны примечания авторов из DAMA International.*

[2] Термин «framework» (и его общее значение) хорошо известен отечественным специалистам в сфере ИТ и других профессиональных областях. При этом любые варианты его перевода находят своих критиков.

[в] данном издании «framework» чаще всего переводится как «рамочная структура». Такой же перевод используется, в частности, в ГОСТ Р ИСО 19439-2008 «Интеграция предприятия. Основа моделирования предприятия». В отдельных случаях используются уже устоявшиеся альтернативы (например, «модель Захмана»).

[в] отличие от других возможных вариантов (например, «основа», «среда», «модель»), термин «рамочная структура» в большинстве случаев применяется в специальной литературе именно для перевода слова «framework». Таким образом, читатель может четко определить, что в оригинале имеется в виду именно это понятие, а не «foundation», «environment» или «model», и более точно уяснить заложенный авторами смысл. — *Примеч. науч. ред.*

1.2 Цели

В рамках отдельно взятой организации управление данными преследует следующие цели.

- Выявление и обслуживание информационных потребностей организации и заинтересованных в ее развитии сторон, включая акционеров, клиентов, сотрудников и деловых партнеров.
- Сбор, хранение, защита и обеспечение целостности данных.
- Обеспечение качества данных и информации.
- Обеспечение конфиденциальности и неразглашения данных, касающихся всех заинтересованных сторон.
- Предотвращение несанкционированного доступа к данным и информации, а также их искажения, подтасовки или нецелевого использования.
- Обеспечение эффективного использования данных, что генерировало бы дополнительную выгоду предприятию.

2. ОСНОВНЫЕ ПОНЯТИЯ И КОНЦЕПЦИИ

2.1 Данные

Устоявшиеся определения понятия *данные* подчеркивают их роль в представлении фактов об окружающем мире[1]. В сфере информационных технологий *данные* также понимаются как информация, сохраненная в цифровой форме (хотя в реальности данные не ограничиваются исключительно оцифрованной информацией, а принципы управления данными в равной степени применимы и к бумажным архивам, и к электронным базам данных). В связи с тем, что сегодня мы в состоянии фиксировать в электронном виде массу всевозможной информации, мы стали называть «данными» многие вещи, которые ранее так не назывались: например, имена, адреса, дни рождения, съеденный в субботу обед или недавно приобретенную книгу.

Подобные факты об отдельных людях можно агрегировать, анализировать, использовать в коммерческих целях, в нуждах здравоохранения или для оказания влияния на публичную политику. Более того, растущие технологические возможности регистрации и замеров параметров

[1] В «Новом оксфордском словаре американского английского» (*New Oxford American Dictionary*) данные определяются как «факты и статистика, собранные вместе для анализа». Американское общество качества (ASQ) определяет данные как «подборку фактов» и описывает два типа численных данных: (1) измеримые, переменные или варьируемые и (2) счетные или качественные. Международная организация по стандартизации (ISO) определяет данные как «многократно интерпретируемое представление информации, пригодное для передачи, интерпретации или обработки формализованным образом» (ГОСТ Р ИСО/МЭК 11179-1-2010). Последнее определение явным образом указывает на электронный характер данных и подразумевает (вполне справедливо), что данные требуют стандартизации, поскольку управляются посредством информационно-технологических систем. Однако в данном определении не упомянуты ни трудности непротиворечивой и однозначной формализации данных, передаваемых между несовместимыми системами, ни понятие неструктурированных данных.

самых разных событий и проявлений человеческой жизнедеятельности (от отзвуков Большого взрыва до пульса отдельного человека), равно как сбора, хранения и анализа электронных версий всего того, что ранее данными вовсе не считалось (видео, фото, аудио, документов), скоро превысят наши способности по объединению этих данных в пригодную для практического использования информацию[1]. Чтобы использовать преимущества, которые открывает изобилие разнообразных данных, и не растеряться при столкновении с объемом и скоростью их поступления, требуются надежные и масштабируемые практики управления данными.

Большинство людей полагает, что данные, как собрание фактов, отражают истинную картину мира, а потому должны согласовываться между собой. Но ведь и «факты» сами по себе вещь упрямая и далеко не всегда простая, однозначная и прямолинейная. Чего в таком случае ждать от данных, которые есть всего лишь средство представления фактов? Данные заслоняют и замещают собой реальные вещи, которые описывают (Chisholm, 2010). Данные — это одновременно интерпретация представляемых ими предметов и, сами по себе, предметы, подлежащие интерпретации (Sebastian-Coleman, 2013). Иными словами, в отрыве от контекста данные становятся бессмысленными. Для понимания их смысла необходимо знать контекст, который, можно сказать, служит репрезентативной системой данных, а такая система должна обязательно включать общепринятую терминологию и набор связей между компонентами. Если принятые в такой системе соглашения нам известны, мы способны интерпретировать данные в ее рамках[2]. Подобные соглашения часто документируются отдельно в виде так называемых метаданных.

Однако, поскольку люди часто выбирают различные представления одних и тех же понятий, способы представления одного и того же перманентно множатся. Таким образом, одни и те же данные облекаются в различные формы. Для примера вообразите себе, сколько различных форматов имеется для представления столь элементарных и неоспоримых данных, как календарные даты. А теперь можно предположить, что творится с понятиями более сложными по структуре (такими, к примеру, как «потребитель» или «продукт»), — тут и глубина, и уровень детализации информации, нуждающейся в представлении, далеко не очевидны. Это усложняет процесс представления данных, а со временем приводит и к тому, что процедуры управления данными обрастают массой нюансов (см. главу 10).

Даже в пределах одной организации одна и та же идея может облекаться в различные формы представления. Отсюда потребность и в архитектуре данных, и в моделировании, и в руководстве и распоряжении данными, и в управлении метаданными, и в управлении качеством данных, — всё это реально необходимо, чтобы помочь людям правильно понимать и использовать данные. На уровне же взаимодействия между несколькими организациями проблемы, обусловленные неоднозначными представлениями, многократно множатся, что свидетельствует о необходимости единых стандартов представления данных и управления ими, без которых обеспечить их единообразие и сопоставимость нереально.

---

[1] http://ubm.io/2c4yPOJ; http://bit.ly/1rOQkt1

[2] Подробнее о конструируемости данных см.: Kent, *Data and Reality* (2012); Devlin, *Business Unintelligence* (2013).

Управлять данными организациям нужно было всегда, но технологические изменения новейшего времени не просто неизмеримо расширили спектр необходимых работ в этой области, но и в корне изменили само человеческое представление о том, что такое данные. Такие изменения теперь позволяют организациям использовать данные по-новому — для разработки продуктов, распространения информации, получения знаний и приумножения успехов в целом. Бурное развитие технологий и обусловленное им лавинообразное нарастание человеческой способности производить, собирать и извлекать данные для осмысления пропорционально повысило и потребность в эффективном управлении ими.

2.2 Данные и информация

На описание взаимосвязи между данными и информацией было израсходовано много чернил, в результате чего данные окрестили «информационным сырьем», а информацию — «данными в контексте»[1]. Для описания взаимоотношений между ними часто используют четырехуровневую пирамиду, в основании (фундаменте) которой лежат данные, а выше идут ярусы «информация», «знание» и «мудрость» (на самой вершине). Для осознания необходимости качественного управления данными образ такой пирамиды полезен, но столь упрощенное представление на практике влечет за собой ряд трудностей.

- Во-первых, оно основано на гипотезе об объективном существовании данных. Но данных как таковых в природе попросту не существует. Данные создаются.
- Во-вторых, линейная последовательность преобразования данных в мудрость через информацию и знание игнорирует неоспоримый факт необходимости обладать знанием о том, как, для начала, хотя бы создавать данные, а затем перерабатывать их в информацию и т. д.
- В-третьих, априори подразумевается, что данные и информация — вещи друг от друга отдельные и независимые, в то время как в действительности оба понятия тесно переплетены и по отдельности друг от друга (в чистом виде) не существуют. Данные — это форма информации, а информация — это форма данных.

Внутри организации полезно проводить четкую границу между информацией и данными хотя бы в целях более ясного донесения требований и ожиданий по различным направлениям практической работы до различных заинтересованных аудиторий. (Пример: «Предлагаем ознакомиться с отчетом о продажах за минувший квартал [информация]. Он составлен на основе данных нашего информационного хранилища [данные]. В следующем квартале эти результаты [данные] будут использованы для создания сравнительных показателей нашей работы по отношению к предыдущему кварталу [информация]».) Признание различий между данными и информацией, а также целями, для которых они могут быть использованы, служит основой стержневого постулата об управлении данными: предметом управления являются и данные, и информация;

[1] English (1999); DAMA (2009).

при этом качество и того и другого возрастает лишь при согласованном управлении ими с учетом потребностей конечных потребителей. Поэтому в дальнейшем в DMBOK термины «информация» и «данные» используются как взаимозаменяемые синонимы.

2.3 Данные как актив организации

Под *активом* понимается имеющийся в собственности или контролируемый экономический ресурс, содержащий в себе или производящий ценность. При этом он может быть конвертирован в деньги. Представление о данных как об активе предприятия на сегодняшний день вполне устоялось, а вот понимание того, как управлять данными как активом, всё еще находится на стадии формирования. В начале 1990-х годов в ряде организаций ставили под сомнение возможность конвертации в денежные активы деловой репутации (гудвилла), а сегодня ее стоимость — стандартная строка в отчете о прибылях и убытках компании. Аналогичным образом и монетизация данных всё чаще переходит в разряд реальных источников пополнения бюджета организаций. Вероятно, и она вскоре окажется в ряду прочих доходных статей публикуемых финансовых отчетов (см. главу 3).

В наши дни организации полагаются на свои информационные активы как на реальный ресурс повышения эффективности и оптимизации работы. Компании используют данные, чтобы лучше понимать своих клиентов и их нужды, создавать новые продукты и услуги, а также повышать операционную эффективность за счет снижения издержек и минимизации рисков. Госучреждения, учебные заведения и некоммерческие организации также нуждаются в высококачественных данных для успешного ведения как своей текущей деятельности, так и планирования развития в кратко-, средне- и долгосрочной перспективе. Чем больше организации зависят от данных, тем отчетливее определяется роль информации как стратегически значимого актива.

Многие организации определяют себя как «управляемые на основе данных» (data-driven). У бизнеса, нацеленного на сохранение конкурентоспособности, нет иного выбора, кроме отказа от принятия решений (подсказанного «внутренним чутьем» или инстинктами руководителей) и перехода к применению аналитики в поисках действенных решений. Управление на основе данных подразумевает безусловное признание необходимости эффективного управления данными, вкупе с профессиональной выучкой и дисциплиной посредством оптимального сплава навыков высокоуровневого руководства бизнесом и технического опыта.

Более того, нарастающие темпы изменений в бизнес-среде в наши дни прямо указывают на то, что назревшие перемены в области управления данными отныне носят характер обязательных, а не факультативных. «Цифровой прорыв» (digital disruption) стал нормой, и для адекватной реакции на него бизнес должен кооперироваться с техническими специалистами по управлению данными с целью совместного создания информационных решений, в равной мере соответствующих потребностям как самого бизнеса, так и его партнеров по каждому направлению. Они обязаны совместно планировать получение и управление данными, которые, по их общему пониманию, нужны для реализации бизнес-стратегии. А в дополнение к этому им необходимо осваивать все новые способы использования данных, чтобы извлекать из них максимальную пользу.

2.4 Принципы управления данными

Управление данными имеет ряд общих характеристик с управлением другими активами (см. рис. 1). Оно требует знания информационных активов, имеющихся в распоряжении организации, а также задач, которые могут быть решены с их помощью. Кроме того, необходимо уметь определять наилучшие методы использования данных для достижения целей организации.

Как и в любых других процессах, связанных с менеджментом, в управлении данными должны сбалансированно учитываться как стратегические цели, так и текущие операционные задачи. Для нахождения и соблюдения оптимального баланса рекомендуется следовать нижеизложенному своду принципов, которые отражают наиболее характерные особенности управления данными и служат проводником в этом процессе.

- **Данные — актив с уникальными свойствами.** Будучи нематериальным ресурсом, данные требуют соответствующего подхода к управлению ими, во многом отличного от подхода, применяемого к денежным или иным материальным ресурсам. Наиболее очевидной отличительной особенностью данных как актива является их неисчерпаемость в том смысле, что мы используем их, не расходуя и не поглощая имеющихся запасов, в отличие от финансовых и материальных ресурсов.
- **Ценность данных может и должна выражаться в экономических терминах.** Коль скоро мы называем данные активом, значит, они имеют объективную стоимость. Методики количественной и качественной оценки ценности данных уже существуют; правда, до их стандартизации дело пока не дошло. Организациям, желающим принимать более эффективные решения в отношении имеющихся в их распоряжении данных, следует разрабатывать как можно более последовательные и унифицированные подходы к объективной оценке их ценности. Кроме того, не лишним будет также оценивать в денежном выражении и издержки, обусловленные низким качеством данных, и дополнительный экономический эффект от реализации мер по обеспечению их высокого качества.
- **Управление данными подразумевает управление качеством данных.** Первоочередная задача управления данными — обеспечивать пригодность данных для их использования по прямому назначению. Чтобы управлять качеством, организации должны удостовериться в том, что правильно понимают требования всех заинтересованных сторон к качеству данных, и оценивать данные согласно этим требованиям.
- **Для управления данными необходимы метаданные.** Для управления любым активом нужны данные об этом активе (численность сотрудников, номера счетов и т. д.). Данные, используемые для управления другими данными, принято называть *метаданными*. Поскольку данные — вещь неосязаемая, для понимания того, что они собой представляют и как их использовать, требуются определения и знания, которые и формулируются в виде метаданных. Порождаются метаданные в недрах самых различных процессов, связанных с созданием, обработкой и использованием данных, включая проектирование архитектуры, моделирование,

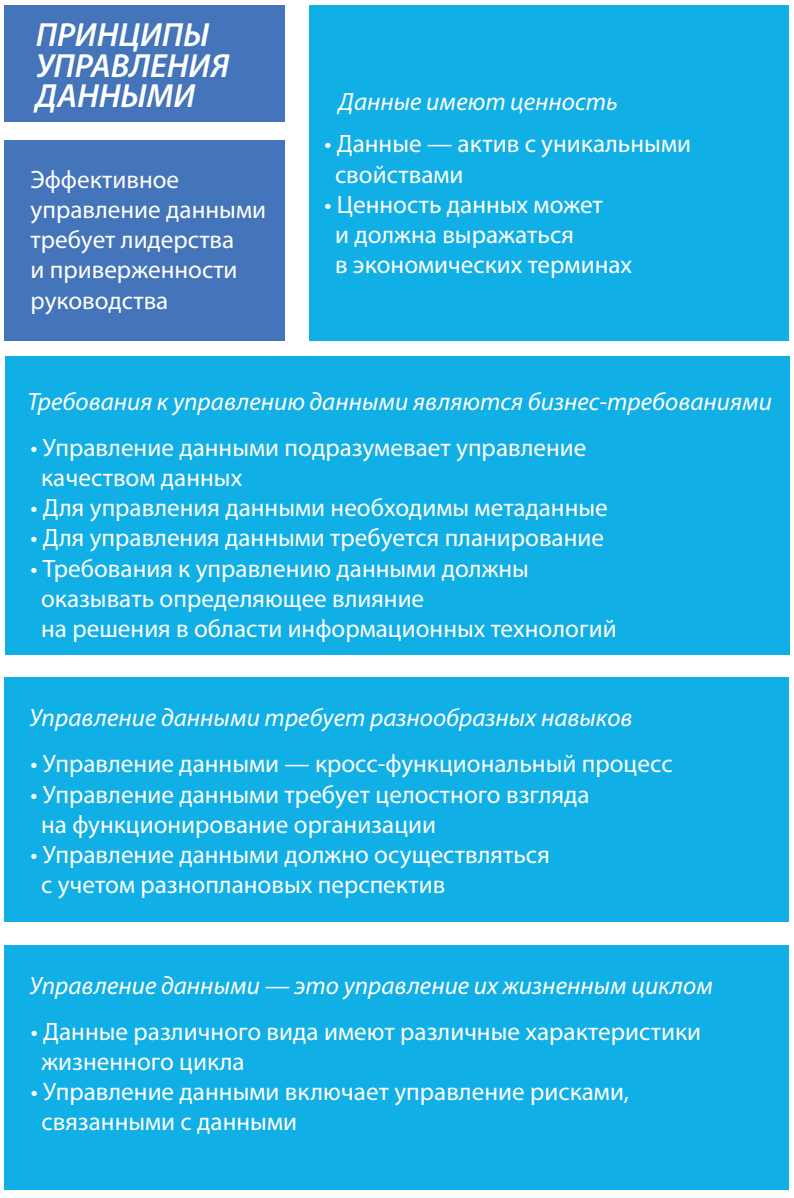


Рисунок 1. Принципы управления данными

распоряжение, руководство и управление качеством, разработку систем, текущие информационно-технологические и бизнес-операции, аналитику.

- **Для управления данными требуется планирование.** Даже в небольших организациях встречаются крайне сложные по структуре ландшафты технологических и бизнес-процессов. Данные создаются в различных местах и перераспределяются по различным участкам с целью последующего использования. Для координации работы и согласования конечных результатов

управление данными должно вестись тщательно спланированным образом — как с точки зрения архитектуры, так и с точки зрения процессов и процедур.

- **Управление данными — кросс-функциональный процесс, требующий широкого спектра знаний и навыков.** Одной-единственной команде не по силам управлять всеми данными организации. Этот процесс требует самых разных технических и нетехнических навыков, а также способности к сотрудничеству между функциональными подразделениями.
- **Управление данными требует целостного взгляда на функционирование организации.** Помимо локальных вариантов применения, управление данными должно осуществляться в масштабах всей организации: таково непременное условие его эффективности. Этим, в частности, и обусловлено тесное переплетение функций управления данными и руководства данными.
- **Управление данными должно осуществляться с учетом разноплановых перспектив.** Данные — субстанция изменчивая. Управление ими должно непрерывно эволюционировать, чтобы не отставать в развитии от новых способов создания и использования данных, а также от нужд конечных потребителей.
- **Управление данными — это управление их жизненным циклом.** У данных есть жизненный цикл, а потому управление ими требует управления их жизненным циклом. Поскольку данные множатся, порождая всё новые данные, жизненный цикл у них бывает крайне сложным и запутанным. Таким образом, практики управления данными должны учитывать все нюансы жизненного цикла данных.
- **Данные различного вида имеют различные характеристики жизненного цикла.** Как следствие, и требования к управлению данными различного вида отличаются друг от друга. Практики управления данными должны распознавать эти отличия и быть достаточно гибкими, чтобы соответствовать разнообразным типам требований к жизненному циклу.
- **Управление данными включает управление рисками, связанными с данными.** Являясь активом, данные также представляют и угрозу для организации, в частности из-за рисков их утери, хищения, нецелевого использования или злоупотребления. Организации должны учитывать и этические аспекты использования имеющихся у них данных. Управление рисками, связанными с данными, должно вестись на протяжении всего жизненного цикла данных.
- **Требования к управлению данными должны оказывать определяющее влияние на решения в области информационных технологий.** Данные и управление данными тесно переплетены с информационными технологиями и управлением ИТ. Управление данными должно быть организовано таким образом, чтобы информационные технологии обслуживали, а не определяли стратегические потребности организации в данных.
- **Эффективное управление данными требует лидерства и приверженности руководства.** Управление данными — сложный комплекс процессов, для эффективной реализации и работы которого требуются координация, сотрудничество и приверженность. Для достижения подобного эффекта одних лишь навыков управления недостаточно: нужны еще и ясное ви́дение, и устремленность к цели, а это всё — производные от лидерства и приверженности руководства.

2.5 Проблемы управления данными

Поскольку функция управления данными обладает определенными особенностями ввиду свойств данных как таковых, следование сформулированным выше принципам управления данными осложняется характерными проблемами. Детальному обсуждению этих проблем посвящены разделы 2.5.1–2.5.13. Обратите внимание на то, что многие из них носят комплексный характер, поскольку касаются соблюдения двух и более принципов.

2.5.1 Данные отличаются от других активов[1]

Материальные активы можно продемонстрировать, потрогать, переместить. В любой момент времени каждый из них находится в определенном месте. К примеру, финансовые активы учитываются в балансовой ведомости. Тогда как данные — нечто иное. Они неосязаемы и эфемерны, но при этом долговечны и износоустойчивы, однако со временем могут обесцениваться или, напротив, дорожать. Данные легко скопировать и отправить куда угодно. Однако в случае утери или уничтожения данных воспроизвести или восстановить их очень непросто. Поскольку данные не расходуются во время пользования, их можно задействовать многократно и даже похищать без каких-либо внешних признаков убыли. Данные динамичны и допускают многоцелевое применение. Одними и теми же данными могут одновременно пользоваться множество людей, что невозможно в отношении материальных или финансовых активов. Многие способы использования данных порождают еще больше данных. Большинству организаций приходится управлять всевозрастающими объемами данных и связями между наборами данных.

Все эти отличия весьма затрудняют денежную оценку данных. А незнание их ценности в денежном выражении, в свою очередь, затрудняет измерение вклада данных в успех организации. Те же отличия приводят к возникновению и других вопросов, сказывающихся на управлении данными: в частности, вызывают затруднения с определением собственников данных, учетом накопленных организацией массивов данных, защитой данных от злоупотреблений, управлением рисками, обусловленными избыточностью данных, определением и обеспечением соблюдения стандартов качества данных.

Несмотря на затруднительность объективного определения ценности данных, большинство людей отдают себе отчет в том, что данные — это реальная ценность. Данные любой организации — вещь уникальная сама по себе. И в случае утери организацией своих уникальных данных (таких, как списки заказчиков, реестры товарных запасов, история претензий и т. п.) заменить их будет нечем, а восстановить невозможно или непомерно дорого. Ведь данные — еще и средство самопознания организации, это метаресурс, описывающий другие ресурсы и активы. В этой роли данные служат фундаментом системы представлений организации о себе самой.

Данные и информация критически важны для ведения бизнеса как внутри организации, так и между организациями. Большинство операционных бизнес-транзакций предусматривают

[1] Настоящий раздел содержит заимствования из следующих публикаций: Redman, Thomas. *Data Quality for the Information Age* (1996), p. 41–42, 232–236; *Data Driven* (2008), Chapter One, «The Wondrous and Perilous Properties of Data and Information».

информационный обмен. Чаще всего он происходит в электронной форме и протоколируется. И по журналам обмена данными можно не только отследить хронологию и содержание собственно сообщений и потоков данных, но и получить практически исчерпывающую информацию о характере деятельности и порядке функционирования организации.

Именно по причине огромной значимости роли данных в жизни любой организации они и нуждаются в бережном обращении и надежном управлении.

2.5.2 Определение ценности данных

Под *ценностью* принято понимать разницу между затратами на создание или приобретение вещи и полученной от нее выгоды. В некоторых случаях, например с акциями, рассчитать их реальную ценность не составляет труда: это разница между ценой, по которой акции проданы, и ценой, по которой они были ранее приобретены. С данными всё обстоит значительно сложнее, поскольку ни затраты на их получение, ни выгоды от обладания ими рассчитать столь же просто не удастся за неимением стандартных методик.

Поскольку данные каждой организации уникальны и присущи только ей, оценку их стоимости нужно начинать с определения общих категорий затрат и экономических выгод, которые можно систематически и непротиворечиво использовать в рамках организации. Вот лишь некоторые примеры подобных категорий[1]:

- Затраты на получение и хранение данных.
- Затраты на восстановление данных в случае утери.
- Потери организации из-за отсутствия нужных данных.
- Затраты на минимизацию риска и потенциальные убытки, обусловленные рисками, связанными с данными.
- Затраты на совершенствование структуры и повышение качества данных.
- Дополнительные выгоды и преимущества за счет обладания данными более высокого качества.
- Цена, которую конкуренты готовы заплатить за данные.
- Стоимость данных в случае их продажи.
- Ожидаемые дополнительные доходы от инновационного использования данных.

Главная трудность с оценкой данных как актива заключается в том, что ценность данных зависит от контекста (то, что имеет высокую ценность с точки зрения одной организации, может не представлять ни малейшей ценности для другой), а нередко — и от времени оценки (то, что имело ценность вчера, назавтра может обесцениться). Тем не менее внутри отдельно взятой организации данные определенного типа могут иметь непреходящую ценность. Возьмем, к примеру,

[1] В процессе подготовки DMBOK2 к печати обнаружилось и вполне объективное средство оценки стоимости данных: сетевой червь-вымогатель WannaCry атаковал (по состоянию на 17 мая 2017 г.) компьютеры свыше 100 000 организаций в 150 странах мира. С помощью этого вредоносного ПО злоумышленники захватывали в заложники файлы с данными пользователей и требовали выкуп за их разблокировку (см.: http://bit.ly/2tNoyQ7).

информацию о постоянных клиентах компании. Чем больше данных о постоянном клиенте накапливается, тем ценнее становится информация о нем, поскольку появляется всё больше сведений о характере его поведения.

Применительно к управлению данными установить порядок финансовой оценки ценности данных безусловно необходимо, поскольку организация обязательно должна иметь четкое представление об информационных активах, которыми она располагает, в финансовых терминах, чтобы принимать решения относительно того, как ими распоряжаться, грамотно и последовательно. Навешивание ценников на данные — первый шаг на пути к оценке реального экономического эффекта от различных направлений деятельности по управлению ими[1]. Процедуру денежной оценки данных также можно использовать в качестве одного из средств управления изменениями. Попросить специалистов в области управления данными и ключевых лиц, в чьих интересах они работают, лучше понять финансовую подоплеку того, чем они занимаются, — верное средство помочь организации научиться по-настоящему ценить собственные данные и переосмыслить подход к управлению ими.

2.5.3 Качество данных

Гарантировать высокое качество данных — наиглавнейшая функция управления ими. Организации управляют собственными данными, потому что хотят их использовать. Если же они не могут полагаться на достоверность данных и их пригодность для использования в соответствии с потребностями бизнеса, то и все усилия, направленные на их сбор, хранение, защиту и управление доступом, были потрачены впустую. Потому для обеспечения соответствия данных потребностям бизнеса организациям нужно работать в тесном контакте с конечными потребителями данных, чтобы четко определить их нужды и совместно установить ключевые характеристики их качества.

Во многом по причине того, что работа с данными традиционно ассоциировалась с информационными технологиями, управление качеством данных исторически оказалось отодвинутым на задний план. ИТ-командам часто бывает глубоко безразлично, что за данные хранятся в создаваемых ими системах, поскольку главное для них — чтобы сами системы работали. Вероятно, тот, кто впервые заметил «мусор на входе — мусор на выходе», был программистом, и он, без сомнения, хотел оставить всё как есть. Вот только те, кому эти данные нужны для использования, не могут себе позволить столь пренебрежительного отношения к их качеству. Пользователи, как правило, исходят из того, что имеющиеся в их распоряжении данные надежны и достоверны, пока сама жизнь не заставляет их в этом усомниться. А вот единожды потеряв доверие к данным, восстановить его бывает непросто.

Большинство практических применений данных так или иначе связаны с их изучением с целью последующего применения знаний для создания чего-то ценного. Примеры включают изучение поведения потребителей с целью совершенствования продуктов или услуг, оценку эффективности работы организации или изучение рыночных тенденций с целью выработки более

[1] Примеры с разборами можно найти в работе: Aiken and Billings, *Monetizing Data Management* (2014).

эффективной бизнес-стратегии, и т. д. Некачественные данные ожидаемо приведут к неверным решениям и негативным последствиям.

Не менее важно и то, что низкое качество данных чревато не только недополученной прибылью, но и прямыми убытками. По различным экспертным оценкам, организации тратят от 10% до 30% дохода на устранение последствий проблем, обусловленных низким качеством данных. По оценке IBM, только в США совокупные издержки из-за низкого качества данных в 2016 году составили 3,1 трлн долларов[1]. Многие издержки, образующиеся из-за данных низкого качества, косвенны и поэтому трудноизмеримы. Другие же, такие как штрафы, являются прямыми и легко подсчитываемыми. Среди основных источников и статей издержек, которые являются следствием низкого качества данных, стоит отметить:

- брак и переделки;
- временные решения и скрытые доработки;
- неэффективную организацию и/или низкую производительность труда;
- внутриорганизационные конфликты;
- низкую удовлетворенность работников;
- разочарование, неудовлетворенность клиентов;
- упущенные возможности, в том числе из-за утраты способности к инновациям;
- непредвиденные расходы на устранение несоответствий или штрафы;
- репутационные издержки.

Противопоставленные им выгоды от высококачественных данных:

- рост удовлетворенности клиентов;
- повышение производительности;
- снижение рисков;
- способность эффективно использовать представляющиеся возможности;
- повышение дохода;
- получение конкурентных преимуществ за счет глубокого понимания клиентов, продуктов, процессов и возможностей.

Все вышеперечисленные издержки и выгоды подразумевают, что обеспечение качества данных — не единоразовая задача. Для производства высококачественных данных требуются планирование, приверженность и в целом образ мышления, ориентированный на встраивание качества во все процессы и системы. Все функции управления данными сказываются на их качестве, а потому все они и отвечают за него в процессе своего выполнения (см. главу 13).

[1] *Источник*: Redman, Thomas. «Bad Data Costs U. S. $3 Trillion per Year». Harvard Business Review. 22 September 2016; https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year

2.5.4 Планирование перехода к улучшенным данным

Как уже упоминалось во введении к настоящей главе, извлечение выгоды из данных не происходит случайным образом. Для ее получения требуется всестороннее планирование. Прежде всего организациям нужно отдавать себе отчет в том, что они вполне способны контролировать процессы получения и создания данных. Лишь начав рассматривать данные в качестве создаваемого ими же продукта, они смогут принимать оптимальные решения по управлению данными на протяжении всего их жизненного цикла. Такие решения требуют системного мышления, поскольку затрагивают следующие аспекты:

- каким образом данные связывают между собой раздельные со всех остальных точек зрения бизнес-процессы;
- связи между бизнес-процессами и задействованными в них технологиями;
- архитектуру информационных систем и массивы данных, которые порождаются и используются в этих системах;
- способы использования данных для совершенствования стратегии организации.

Планирование перехода к улучшенным данным требует стратегического подхода к архитектуре, моделированию и другим функциям проектирования. Кроме того, не обойтись и без стратегического взаимодействия между руководством бизнеса и ИТ. И, конечно же, планирование улучшения данных всецело зависит от способности эффективно реализовывать отдельно взятые проекты.

Трудность заключается в том, что обычно подобные работы ведутся, во-первых, под административным давлением, а во-вторых, в условиях постоянного дефицита времени и денег, — и всё это никак не способствует повышению качества планирования. Поэтому организации должны четко разграничивать долгосрочные и краткосрочные цели и сбалансированно распределять ресурсы, выделяемые на их достижение в рамках реализации своих стратегических планов. Полная ясность относительно приоритетов и компромиссов устраняет многие препятствия с пути к нахождению оптимальных решений.

2.5.5 Метаданные и управление данными

Для управления данными как активом организациям требуются надежные метаданные. В этом контексте метаданные включают не только описываемые в главе 12 бизнес-, технические и операционные метаданные, но и метаданные, встроенные в архитектуру данных, модели данных, требования безопасности данных, стандарты интеграции данных и процессы их обработки (см. главы 4–11).

Метаданные описывают, какими данными располагает организация, что они отражают, как классифицируются, откуда получены, как перемещаются внутри организации, как видоизменяются при использовании, кому доступны, а кому нет, а также каков уровень их качества. Данные — абстрактны. Без определений и сопроводительных описаний контекста понять их невозможно. Следовательно, именно метаданные делают доступными для понимания и сами данные, и их жизненный цикл, и сложные системы, содержащие или использующие эти данные.

Проблема заключается в том, что метаданные являются разновидностью данных, а значит, также нуждаются в управлении. Если управление данными в организации в целом поставлено плохо, управлением метаданными в такой организации, как правило, не занимаются вовсе. Часто именно с налаживания управления метаданными начинается приведение в порядок и совершенствование системы управления данными в целом.

2.5.6 Управление данными как кросс-функциональный процесс

Управление данными — процесс сложный. Данными в различных отделах организации управляют команды специалистов, отвечающих за различные операции на различных фазах жизненного цикла данных. Управление данными требует навыков проектирования для планирования создания систем, высокопрофессиональных технических умений для администрирования аппаратного комплекса и создания программного обеспечения, специальных знаний в области анализа данных для проработки проблемных вопросов, аналитических навыков для интерпретации данных, языковых навыков для достижения консенсуса относительно определений и моделей, и, наконец, стратегического мышления для выявления возможностей улучшения обслуживания клиентов и достижения поставленных целей.

Трудность заключается в том, чтобы собрать команду из людей со столь разнообразными навыками и точками зрения — и подвести их к общему пониманию того, как им воспользоваться умениями каждого, чтобы наладить совместную работу, направленную на достижение общих целей.

2.5.7 Целостный взгляд на функционирование организации

Управление данными требует понимания объемов и спектра предметных областей данных, которыми располагает организация. Данные — одна из ее «горизонталей». Потоки данных свободно движутся между функциями продаж, маркетинга, операционного управления… По крайней мере, это должно быть именно так. В реальности же не только сама организация располагает уникальными данными: зачастую уникальными оказываются и данные какого-нибудь отдела или другого структурного подразделения внутри организации. Поскольку данные зачастую рассматриваются в качестве всего лишь побочного продукта операционных процессов (например, записи о продажах в базе данных появляются как побочный продукт процесса продаж), порой никто и не планирует их использование вне текущих нужд.

Даже в пределах одной и той же организации данные могут быть несопоставимы по самым разным аспектам. Данные создаются на множестве участков внутри организации. При этом в разных отделах могут по-разному формализовать представление одного и того же понятия (например, «клиент», «продукт» или «поставщик»). Любой, кому доводилось участвовать в реализации проектов по интеграции данных или управлению основными данными, подтвердит, что неявные (а порою и вопиющие) разночтения в репрезентации создают серьезные трудности для управления данными в масштабах организации. В то же время заинтересованным лицам из числа неспециалистов представляется само собой разумеющимся, что данные в пределах одной и той же организации непременно логически согласованы между собой, а потому они видят

основную цель управления данными лишь в том, чтобы просто объединить их вместе (в общепринятом понимании), так чтобы они могли использоваться широким кругом потребителей данных.

Одна из главных причин растущего осознания важности руководства данными как раз и заключается в необходимости помочь организациям в принятии решений о том, как наладить беспрепятственный обмен данными по горизонтали, через межфункциональные вертикальные барьеры (см. главу 3).

2.5.8 Учет различных перспектив

Сегодня организации используют не только данные, создаваемые внутри, но и данные из всевозможных внешних источников. Поэтому им нужно принимать во внимание и учитывать различные нормативно-правовые требования, действующие как в силу национального законодательства, так и в рамках отраслевого регулирования. Создатели данных часто забывают о том, что впоследствии этими данными может воспользоваться кто-то еще. Знание потенциальных направлений применения полученных данных позволяет лучше спланировать управление их жизненным циклом, а заодно и дополнительные меры по повышению их качества. Не следует забывать и о возможности злоупотребления данными. Учет этого риска снижает вероятность его материализации.

2.5.9 Жизненный цикл данных

Как и у любого другого актива, у данных есть свой жизненный цикл. Для эффективного управления информационными активами организации необходимо понимание и планирование их жизненного цикла. Хорошее управление данными подразумевает стратегическое управление на основе ви́дения того, каким образом организация будет распоряжаться своими данными в долгосрочной перспективе. Стратегически ориентированная организация определит не только требования к содержанию своих данных, но и требования к управлению ими. Последние включают политики и ожидаемый порядок использования, обеспечения качества, контроля и защиты данных; подход к архитектуре и проектированию данных в масштабах предприятия; устойчивый и последовательный подход к развитию инфраструктуры и разработке программного обеспечения.

Жизненный цикл данных в своей основе такой же, как и жизненный цикл продукта. Не следует путать его с жизненным циклом разработки систем. Концептуально жизненный цикл данных описывается весьма просто (см. рис. 2). Он включает процессы, которые создают или получают данные; процессы, которые осуществляют их перемещение, преобразование, хранение, а также обеспечивают обслуживание данных и предоставление совместного доступа к ним; процессы использования или применения данных, а также процессы, обеспечивающие их ликвидацию[1]. На протяжении всего их жизненного цикла данные могут очищаться, преобразовываться, подвергаться слиянию, улучшаться или агрегироваться. По мере использования или улучшения данных часто создаются новые данные, а потому жизненный цикл включает внутренние

[1] Подробнее о жизненном цикле продукта и данных см.: McGilvray (2008); English (1999).

итерации, которые на диаграмме не отражены. Данные редко бывают статичными, и в управлении ими задействован целый ряд взаимосвязанных процессов, выстроенных вдоль всего жизненного цикла данных.

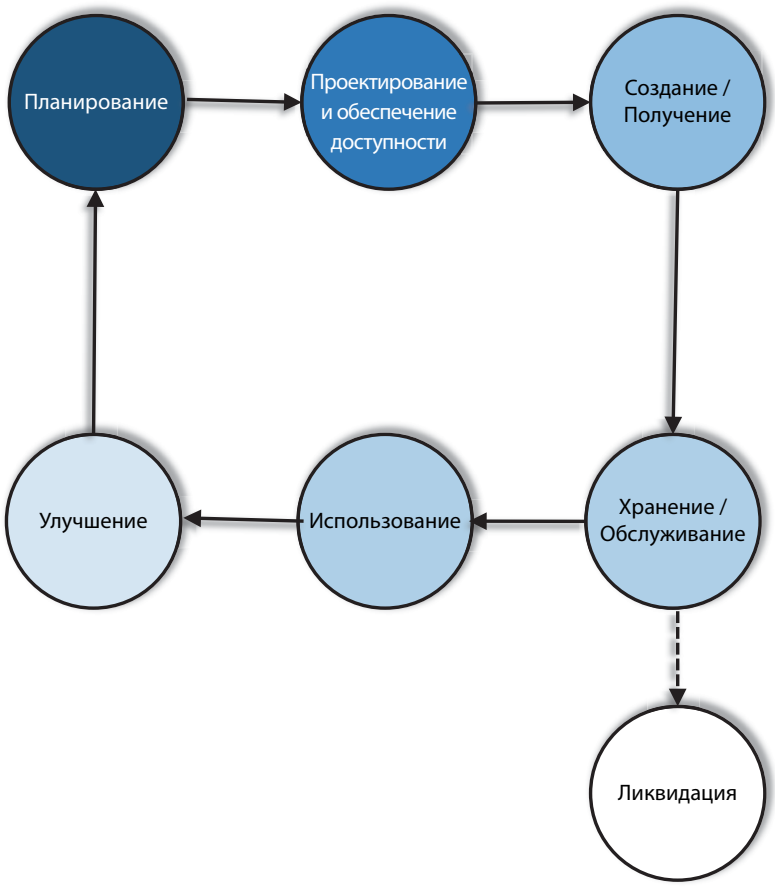


Рисунок 2. Ключевые работы, проводимые в рамках жизненного цикла данных

Специфика жизненного цикла данных в отдельно взятой организации может оказаться весьма запутанной, поскольку данные имеют не только жизненный цикл, они имеют еще и *происхождение* (*lineage*), то есть путь, по которому они движутся от места своего возникновения до места использования, иногда называемый *цепочкой данных* (*data chain*).

Для понимания точного происхождения данных в этом случае приходится документировать первоисточники всех наборов данных, а затем еще и каждое их перемещение, каждое преобразование при передаче из системы в систему, где к ним предоставляется доступ и где они используются. Жизненный цикл и происхождение данных содержат пересечения, и разобраться в них можно, только учитывая все взаимосвязи. Чем лучше организация понимает жизненный цикл и происхождение собственных данных, тем лучше она готова ими управлять.

Фокусировка внимания на жизненном цикле данных при управлении ими имеет ряд важных следствий.

- **Создание и использование — важнейшие элементы жизненного цикла данных.** Управлять данными нужно с четким пониманием того, как они созданы или получены и как используются. Создание данных стоит денег. И окупаются эти затраты лишь в том случае, если полученные данные действительно имеют ценность, то есть активно используются или применяются (см. главы 5, 6, 8, 11 и 14).
- **Управление качеством данных должно осуществляться на протяжении всего их жизненного цикла.** Управление качеством данных — центральное звено управления данными. Данные низкого качества несут лишь издержки и риск вместо выгоды и пользы. Организации нередко сталкиваются с проблемами при управлении качеством данных по той причине, что, как уже отмечалось, данные часто создаются как побочный продукт операционных процессов и явно сформулированные стандарты качества данных не вводятся. Поскольку качество данных может пострадать на любом этапе их жизненного цикла, меры по обеспечению качества следует планировать в расчете на весь цикл жизни данных (см. главу 13).
- **Управление качеством метаданных должно осуществляться на протяжении всего их жизненного цикла.** Поскольку метаданные являются разновидностью данных и поскольку организации опираются на них при управлении прочими данными, подход к управлению качеством метаданных должен быть таким же, как и к управлению качеством других данных (см. главу 12).
- **Управление безопасностью данных должно осуществляться на протяжении всего их жизненного цикла.** К функциям управления данными относится и обеспечение информационной безопасности, и снижение рисков, связанных с данными. Данные, на которые распространяются требования безопасности, должны быть защищены с момента создания вплоть до их ликвидации (см. главу 7).
- **Основные усилия по управлению данными следует фокусировать на критически важных данных.** Организации создают массу данных, значительная часть которых в действительности никогда не используется. Управлять всеми без исключения наборами данных попросту невозможно. Управление жизненным циклом требует фокусировки на критически важных для организации данных и минимизации излишних, устаревших, тривиальных данных[1] (Aiken, 2014).

2.5.10 Различные виды данных

Управление данными осложняется тем фактом, что существуют различные виды данных, которые выдвигают различные требования к управлению жизненным циклом. Любая система управления нуждается в четкой классификации объектов управления. Данные можно классифицировать либо по их видам (например, транзакционные, справочные, основные данные и метаданные;

[1] В англоязычных источниках по отношению к излишним, устаревшим и тривиальным данным часто используются акроним ROT (Redundant, Obsolete, Trivial) и термин «data ROT». — *Примеч. науч. ред.*

или, в качестве альтернативы, данные о категориях, ресурсах, событиях и деталях транзакций), либо по их содержанию (например, области данных, предметные области и т. п.), либо по формату и уровню защиты. Также данные можно классифицировать по способу хранения, размещению хранилищ, порядку доступа и другим параметрам (см. главы 5 и 10).

Поскольку данным различных видов соответствуют различные требования, присущи различные риски и отведены различные роли в организации, многие инструменты управления данными всецело сфокусированы на различных аспектах классификации и контроля (Bryce, 2005). Например, основные данные имеют иное назначение и области применения, нежели транзакционные данные, — соответственно, и требования к управлению данными двух этих видов предъявляются различные (см. главы 9, 10, 12 и 14).

2.5.11 Данные и риск

Данные — не только ценный актив, но и источник риска. Некачественные (неточные, неполные или устаревшие) данные, бесспорно, представляют угрозу для организации по той простой причине, что не являются верными. Но есть и другая категория рисков, обусловленных возможностью неправильной трактовки данных, а также вероятностью злоупотреблений.

Наибольшую ценность для организации представляют данные высшего качества — доступные, релевантные, полные, точные, непротиворечивые, своевременные, удобные для использования, осмысленные и понятные. Однако многие важные решения нам приходится принимать на фоне информационных разрывов — разницы между тем, что мы реально знаем, и тем, что нам нужно было бы знать для вынесения по-настоящему обоснованных и эффективных решений. Эти информационные разрывы вынуждают предприятия брать на себя ответственность за риски, которая потенциально может самым пагубным образом ударить по эффективности и рентабельности операционной деятельности. Тем не менее организации, отдающие себе отчет в ценности высококачественных данных, могут предпринимать конкретные шаги по повышению качества и удобства использования данных и информации, не выходя при этом за рамки нормативно-правового поля и деловой этики.

Всевозрастающая роль информации как актива организаций во всех без исключения секторах экономики привлекла повышенное внимание законодателей и регулирующих органов к потенциальным злоупотреблениям и лазейкам для ее неправомерного использования. Все инициативы по данному вопросу — начиная с закона Сарбейнса — Оксли[1] в США (устанавливающего строгий контроль за точностью и обоснованностью данных о финансовых транзакциях, отображаемых в балансовых ведомостях) и Директивы Solvency II[2] (регламентирующей порядок подтверждения подлинности данных вплоть до первоисточника и гарантий качества данных, заложенных в основу моделей оценки риска и достаточности собственных средств в страховом секторе) и заканчивая лавиной принятых за последнее десятилетие правил и норм защиты персональных и конфиденциальных данных (жестко регламентирующих их получение и обработку в самых разных

[1] Sarbanes–Oxley Act of 2002 (Pub. L. 107–204, 116 Stat. 745, enacted July 30, 2002).

[2] Solvency II Directive (2009/138/EC).

отраслях и юрисдикциях) — ясно указывают на то, что, пока мы всё еще ждем, когда же практика ведения бухгалтерского учета придет к отражению в балансовой ведомости информации в качестве актива, регулирующая среда всё чаще и неотвратимее требует внесения ее в реестр рисков и принятия соответствующих мер по их смягчению или взятию под контроль.

Аналогичным образом и потребители, всё яснее сознавая, как именно используются данные о них, ожидают от организаций не просто более удобного обслуживания и эффективной работы, но и гарантий защиты конфиденциальной информации и невмешательства в их частную жизнь. А это подразумевает, что круг лиц, стратегически заинтересованных в эффективном управлении данными, часто не ограничивается одними лишь профессионалами и может оказаться значительно шире, чем мы традиционно привыкли полагать (см. главы 2 и 7).

К сожалению, ввиду слаборазвитого управления данными в организации объем затрат на устранение ущербов при материализации рисков, связанных с данными, может внезапно вырасти и повлиять на финансовые результаты. Когда подобные риски не взяты под должный контроль, акционеры «голосуют ногами», избавляясь от своих долей, регуляторы наказывают компанию штрафами и/или запретами на виды деятельности, а потребители «голосуют кошельком», отдавая предпочтение конкурентам.

2.5.12 Управление данными и информационные технологии

Как отмечалось во введении к главе и подчеркивается постоянно, управление данными включает широчайший спектр разнообразных действий и мер как технического, так и управленческого характера. Поскольку большинство данных сегодня хранится в электронном виде, тактика управления ими сильно зависит от информационных технологий. Концепция управления данными с самого начала была связана с управлением информационными технологиями, и эта унаследованная взаимосвязь сохраняется. Однако во многих организациях растет напряжение вследствие кажущихся противоречий между стремлением к созданию новых технологий и стремлением к большей надежности данных, в то время как это вещи взаимодополняющие.

Для успешного управления данными требуется взвешенное принятие обоснованных решений в отношении применяемых технологий и их надежная реализация, но тут важно помнить о том, что нельзя ставить знак равенства между управлением технологиями и управлением данными. Организациям нужно понимать влияние технологий на данные, дабы избежать соблазна принимать решения относительно данных исходя из особенностей и ограничений тех или иных технологий. На самом деле всё должно обстоять с точностью до наоборот: требования к данным, согласованные с бизнес-стратегией, призваны служить основой для принятия решений относительно того, какие технологии использовать и как.

2.5.13 Эффективное управление данными требует лидерства и приверженности руководства

Опубликованный в 2017 году «*Лидерский манифест о данных*» отмечает, что «лучшие возможности для органичного роста организации заложены в данных». Хотя в большинстве организаций признают за данными статус актива, компании всё еще далеки от того, чтобы называть себя

«управляемыми на основе данных» (data-driven). Более того, большинство из них даже не представляют, какими данными располагают и какие именно данные имеют решающее значение для их бизнеса. Организации продолжают не видеть разницы между данными и информационными технологиями и плохо управляют как тем, так и другим. Они не подходят к данным стратегически и недооценивают работу по управлению данными. Такое положение дел усугубляет проблемы управления данными и подчеркивает критически важный фактор потенциального успеха организации: лидерство и приверженность руководства, помноженные на вовлечение всех без исключения сотрудников на всех уровнях организации [1].

Обозначенные здесь проблемы ясно дают понять, что управление данными — дело нелегкое и непростое. А поскольку мало где оно поставлено по-настоящему хорошо, подавляющее большинство организаций обладает массой неиспользованных возможностей. Чтобы их реализовать, требуются определенное ви́дение, планирование и желание меняться (см. главы 15–17).

Популяризация роли директора по данным (CDO[2]) обусловлена признанием того, что управление данными представляет уникальные вызовы и успешное управление данными должно рассматриваться как «управляемое бизнесом» (business-driven), а не «управляемое ИТ» (IT-driven). CDO может возглавлять инициативы по управлению данными и помогать организации получать максимальную отдачу от ее информационных активов, а также создавать на их основе конкурентные преимущества. Но руководство подобными инициативами — не единственная функция CDO. Ему или ей следует также выступать в роли лидера культурных изменений, которые позволят организации подходить к своим данным с более развитых стратегических позиций.

2.6 Стратегия управления данными

Под стратегией понимается совокупность выборов и решений, определяющих направление деятельности по достижению высокоуровневых целей. В шахматах, к примеру, под стратегией понимают упорядоченную последовательность ходов, ведущих к желаемому результату — победе (мату) или выживанию (ничьей). Соответственно, *стратегический план* — это высокоуровневое представление последовательности действий, направленных на достижение высокоуровневых целей.

Стратегия работы с данными должна предусматривать бизнес-планы использования информации для получения конкурентных преимуществ и реализации целей организации. Она должна исходить из понимания информационных потребностей, заложенных в бизнес-стратегии, а именно: какие данные нужны организации, как она будет получать эти данные, управлять ими, обеспечивать их надежность и достоверность на протяжении всего жизненного цикла, каким образом их использовать.

Обычно стратегия работы с данными дополняется поддерживающей ее более детальной (программной) стратегией управления данными — иными словами, планом обслуживания

[1] Полный текст «Лидерского манифеста о данных» (*The Leader's Data Manifesto*) см.: http://bit.ly/2sQhcy7

[2] В данном издании в качестве сокращения термина «директор по данным» используется устоявшаяся английская аббревиатура CDO (*сокр.* от Chief Data Officer). Аналогичные соглашения используются в отношении названий других должностей (за исключением тех случаев, когда существует устоявшееся русское сокращение). — *Прим. науч. ред.*

и повышения качества данных, обеспечения их целостности, регулирования доступа, защиты и минимизации известных и предполагаемых рисков. Стратегия также должна предусматривать меры, направленные на решение известных проблем в области управления данными.

Во многих организациях стратегия управления данными находится в ведении и реализуется под началом CDO (при этом она принимается командой по руководству данными при поддержке Совета по руководству данными). Часто CDO составляет первоначальные проекты стратегий работы с данными и управления данными даже до того, как сформирован Совет по руководству данными, с целью добиться решимости вышестоящего руководства поддерживать организацию деятельности по распоряжению и руководству данными.

Стратегия управления данными должна включать следующие компоненты:

- убедительно изложенное ви́дение управления данными;
- краткое экономическое обоснование стратегии с избранными примерами;
- руководящие принципы, ценности и перспективы с позиции управления;
- миссию и долгосрочные цели по основным направлениям управления данными;
- предлагаемые показатели успешности управления данными;
- краткосрочные (на 12–24 месяца) задачи программы управления данными, которые должны быть конкретными, измеримыми, практически значимыми, реалистичными и привязанными к точным срокам (согласно принципу SMART: specific, measurable, actionable, realistic, time-bound);
- описания ролей и организационных систем управления данными, включая распределение обязанностей и прав принятия решений;
- описания компонентов и инициатив программы управления данными;
- приоритетную программу работ с объемами и сроками выполнения;
- первоначальный вариант дорожной карты реализации с разбивкой по проектам и мероприятиям.

Результаты стратегического планирования управления данными включают следующее.

- **Положение об управлении данными.** Общее ви́дение, экономическое обоснование, цели, руководящие принципы, показатели успешности и ключевые факторы успеха, известные риски, операционная модель и т. д.
- **Описание содержания программы управления данными.** Цели и задачи на ближайший период планирования (обычно на три года), роли, организационные системы, ответственные за направления и решение задач.
- **Дорожную карту внедрения управления данными.** Конкретные программы, проекты, распределение задач и сроки получения результатов (см. главу 15).

Стратегия управления данными должна охватывать все области знаний, входящие в рамочную структуру управления данными DAMA и имеющие отношение к организации (см. рис. 5, а также разделы 3.3 и 4).

3. РАМОЧНЫЕ СТРУКТУРЫ УПРАВЛЕНИЯ ДАННЫМИ

Управление данными включает ряд взаимозависимых функций, каждая из которых имеет собственные цели, направления деятельности и сферы ответственности. Профессионалам в области управления данными необходимо учитывать весь комплекс проблем, сопутствующих процессу извлечения выгоды из абстрактного актива организации, тщательно уравновешивая стратегические и операционные цели, специфические бизнес-требования и технические требования, задачи по минимизации рисков и обеспечению нормативно-правового соответствия. Дополнительные сложности связаны с необходимостью приведения к общему знаменателю противоречивых представлений относительно того, что отражают те или иные данные и можно ли их считать высококачественными.

Очень многое требует отслеживания, именно поэтому и полезно иметь рамочную структуру, которая помогает составить всестороннее представление об управлении данными и разобраться в составных частях и связях между ними. Функции управления данными взаимозависимы и нуждаются в согласовании. Чтобы организация могла извлекать выгоду из своих информационных активов, люди, отвечающие за различные аспекты работы с данными, должны сотрудничать между собой.

Разработанные на разных уровнях абстракции, рамочные структуры позволяют ознакомиться с широким диапазоном подходов к управлению данными. Рассмотрение под различными углами помогает получить более глубокое представление о подоплеке процессов управления данными и использовать его для прояснения стратегии, разработки дорожных карт, оптимизации командного взаимодействия и согласования функций.

Идеи и концепции, представленные в DMBOK2, будут по-разному применяться различными организациями. Подход к управлению данными в каждом конкретном случае зависит от таких ключевых факторов, как отрасль, совокупность используемых данных, корпоративная культура, уровень зрелости, стратегия, ви́дение и конкретные задачи, решаемые организацией. Описываемые далее в этом разделе рамочные структуры можно уподобить нескольким линзам, через которые можно рассматривать управление данными с целью применения концепций, излагаемых в настоящем руководстве.

- Две первые рамочные структуры — модель стратегического выравнивания и амстердамская информационная модель — отражают высокоуровневые взаимосвязи, которые оказывают влияние на то, как организации управляют собственными данными.
- Рамочная структура DAMA-DMBOK (включая колесо, шестиугольник и контекстную диаграмму DAMA) описывает области знаний по управлению данными согласно определениям DAMA и наглядно объясняет их место в рамках концепции DMBOK.
- Две последние рамочные структуры берут колесо DAMA в качестве отправной точки и перегруппировывают его элементы в таком порядке, который обеспечивает лучшее понимание взаимосвязей между ними.

3.1 Модель стратегического выравнивания

Модель стратегического выравнивания (Henderson and Venkatraman, 1999) абстрактно представляет фундаментальные движущие силы, действующие в рамках любого подхода к управлению данными. В центре модели — связь между данными и информацией. Информация чаще всего ассоциируется с бизнес-стратегией и операционным использованием данных. Данные ассоциируются с информационными технологиями и процессами, которые поддерживают управление системами, делающими данные доступными для использования. Соответственно, по периметру центрального блока «информация/данные» находятся четыре фундаментальные области принятия стратегических решений: стратегия бизнеса; стратегия развития информационных технологий; организационная инфраструктура и процессы; ИТ-инфраструктура и процессы.

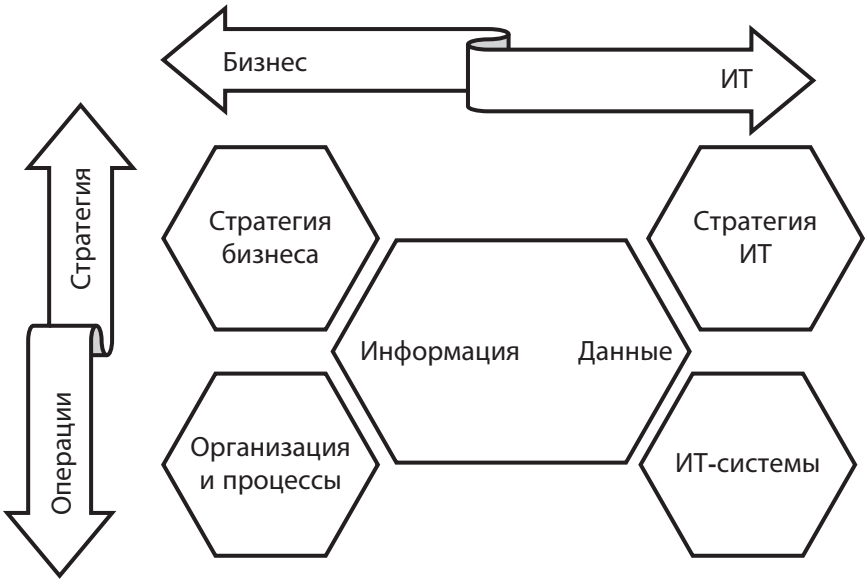


Рисунок 3. Модель стратегического выравнивания Хендерсона — Венкатрамана[1]

На рисунке 3 представлена предельно упрощенная схема модели стратегического выравнивания. В более детализированной версии каждый из четырех шестиугольников имеет дополнительные измерения. В частности, в рамках стратегий (как бизнеса, так и ИТ), нужно дополнительно учитывать объемы решаемых задач, распределение компетенций и иерархию управления. На операционном же уровне следует учитывать инфраструктуру, процессы и навыки. Сопоставляя различные элементы своей деятельности с моделью и определяя связи между ними, организации начинают лучше понимать, как им согласовать и увязать между собой стратегически, а также интегрировать функционально всевозможные части и компоненты, соединив их в единое целое.

[1] Упрощенная версия публикуется с согласия авторов модели.

Даже самое высокоуровневое графическое представление модели вполне может принести пользу организации в плане понимания организационных факторов влияния на принятие решений, касающихся управления данными.

3.2 Амстердамская информационная модель

Амстердамская информационная модель, как и модель стратегического выравнивания, упорядочивает и сопоставляет различные уровни и компоненты управления бизнесом и информационными технологиями (Abcouwer, Maes, and Truijens, 1997)[1]. Известная также под названием «девятиклеточной» (9-cell), амстердамская модель выделяет промежуточный слой, который фокусируется на структуре и тактике, включая планирование и архитектуру. Кроме того, она обращает внимание на необходимость информационного обмена (отражен на рис. 4 в виде вертикальной колонки, объединяющей руководство информацией и качество данных).

Рисунок 4. Амстердамская информационная модель[2]

[1] См. также: Business IT Alignment Blog, *The Amsterdam Information Model (AIM) 9-Cells* (опубликовано 18.12.2010, https://businessitalignment.wordpress.com/tag/amsterdam-information-model/); *Frameworks for IT Management*, Chapter 13. Van Haren Publishing, 2006 (http://bit.ly/2sq2Ow1).

[2] Переработанная версия публикуется с согласия авторов.

Создатели двух представленных выше рамочных структур детально описывают связи между компонентами как по горизонтали (бизнес ↔ ИТ), так и по вертикали (стратегия ↔ операции).

3.3 Рамочная структура DAMA-DMBOK

Рамочная структура DAMA-DMBOK глубже, чем описанные ранее, рассматривает области знаний, совокупность которых, собственно, и определяет рамки управления данными. Визуально рамочная структура управления данными DAMA описывается тремя диаграммами:

- Колесо DAMA (рис. 5).
- Шестиугольник факторов среды (рис. 6).
- Контекстная диаграмма области знаний (рис. 7).

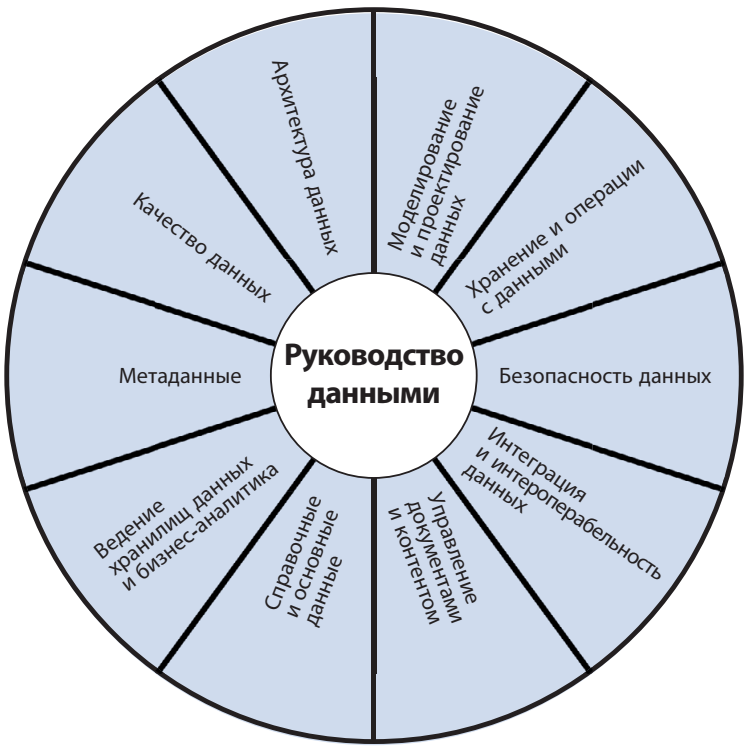


Рисунок 5. Рамочная структура управления данными DAMA-DMBOK2 (колесо DAMA)

Колесо DAMA определяет области знаний по управлению данными. Руководству данными в явном виде отведено центральное место в структуре деятельности по управлению данными, поскольку именно руководство призвано обеспечить согласованность и сбалансированность всех функций. Другие области знаний (архитектура, моделирование и проектирование данных и т. д.) сбалансированно распределены вокруг центра колеса (см. рис. 5). Все эти компоненты необходимы, чтобы функцию управления данными можно было назвать зрелой, но реализовывать их

можно постепенно и в различном порядке, определяемом нуждами организации. Эти области знаний детально рассмотрены в главах 3–13 настоящего издания.

Шестиугольник факторов среды отражает связи между людьми, процессами и технологиями и служит ключом к прочтению контекстных диаграмм DMBOK. В центр помещены цели и принципы, поскольку именно ими необходимо руководствоваться, принимая решения о том, как подойти к исполнению конкретных работ и какие инструменты использовать для эффективного управления данными (см. рис. 6).

Рисунок 6. Шестиугольник факторов среды DAMA

Контекстная диаграмма области знаний (см. рис. 7) описывает отдельные элементы области знаний, включая те, что относятся к людям, процессам и технологиям. За основу контекстных диаграмм взят принцип построения, применяемый в диаграммах SIPOC (suppliers, inputs, processes, outputs, consumers — поставщики, входы, процессы, выходы, потребители), широко используемых в управлении продуктом. В контекстных диаграммах центральное место отводится работам, поскольку именно они дают результаты, удовлетворяющие требованиям тех, кто в этих результатах заинтересован.

Каждая контекстная диаграмма начинается с определения и целей области знаний. Работы, обеспечивающие продвижение к целям, помещены в центральной части диаграммы и распределены по четырем фазам — планирование (П), разработка (Р), операции (О) и контроль (К).

Слева («втекают» в работы) указаны входные материалы и их поставщики. Справа («вытекают» из работ) — результаты работ и их потребители. Наконец, участники работ и их роли определены под работами. В нижней части диаграммы перечисляются инструменты, методы и метрики, относящиеся к рассматриваемой области знаний.

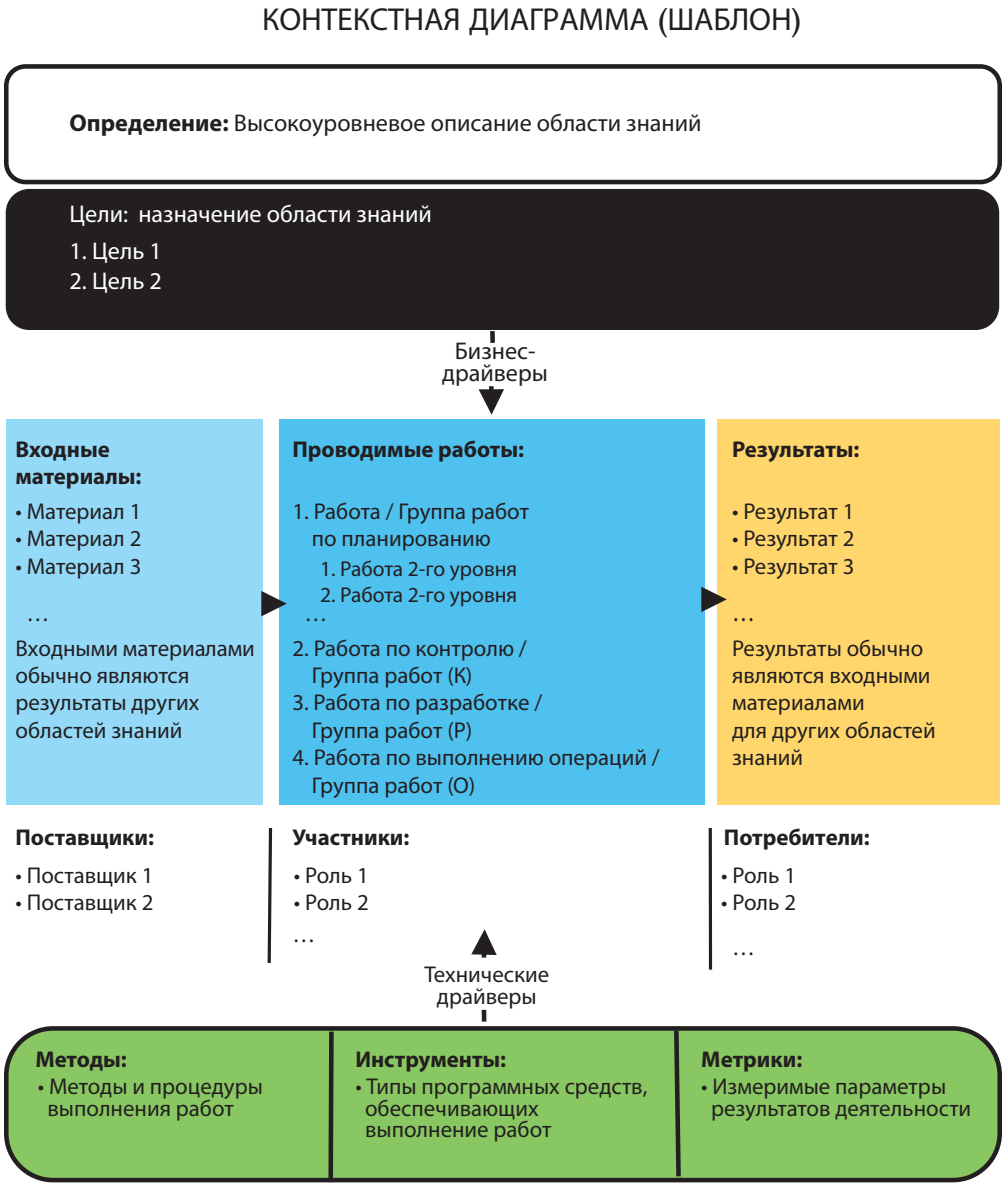


Рисунок 7. Контекстная диаграмма области знаний

Списки в контекстной диаграмме носят иллюстративный и далеко не исчерпывающий характер. На практике они могут дополняться другими пунктами и варьироваться по составу

в зависимости от специфики организаций. В списки ролей включаются лишь самые важные роли. Каждая организация может адаптировать этот шаблон с целью обеспечения соответствия своим потребностям.

Основные блоки в составе контекстной диаграммы включают:

1. **Определение.** Краткое общее определение области знаний.
2. **Цели.** Описание назначения области знаний, а также основополагающих руководящих принципов работы в каждой из них.
3. **Проводимые работы.** Мероприятия и задачи, необходимые для достижения целей данной области знаний. Некоторые работы подразделяются на работы второго уровня, задачи и шаги. Все работы распределяются по четырем категориям: планирование, разработка, выполнение операций и контроль.
   a. **Работы по планированию** определяют стратегический курс и тактику достижения целей управления данными. Работы по планированию повторяются на регулярной основе.
   b. **Работы по разработке** строятся вокруг жизненного цикла разработки систем (system development lifecycle, SDLC), включающего фазы анализа, проектирования, сборки, тестирования, подготовки и развертывания.
   c. **Работы по контролю** включают мероприятия по непрерывной поддержке качества данных, а также целостности, надежности и защищенности систем, обеспечивающих доступ к данным и их использование.
   d. **Работы по выполнению операций** включают мероприятия по поддержке применения, обслуживания и совершенствования систем и процессов, обеспечивающих доступ к данным и их использование.
4. **Входные материалы.** Конкретные вещи, которые необходимы для начала работы в области знаний. Многие работы требуют одних и те же входных материалов — например, знания бизнес-стратегии.
5. **Результаты.** Всё то, что получается в итоге работы, проведенной в рамках области знаний; конкретные вещи, за производство которых отвечает данная функция. Результаты могут как являться самоцелью, так и служить входными материалами для других функций. Некоторые важнейшие результаты создаются совместно несколькими функциями.
6. **Роли и обязанности.** Описывают вклад отдельных лиц и команд в работы, проводимые в данной области знаний. Роли описываются концептуально, основное внимание уделяется группам ролей, которые требуются большинству организаций. Индивидуальные роли определяются в терминах навыков и квалификационных требований. В целях унификации названий ролей за основу была взята модель классификации ИТ-навыков для информационной эры (Skills framework for the information age, SFIA)[1]. Многие роли при этом кросс-функциональны (см. главу 16).

[1] См.: http://bit.ly/2sTusD0

7. **Поставщики.** Лица, ответственные за предоставление входных материалов или обеспечение доступа к ним.
8. **Потребители.** Те, кто получает прямую пользу от основных результатов работ по управлению данными.
9. **Участники.** Лица, непосредственно выполняющие работы в данной области знаний, а также осуществляющие управление работами или задействованные в процессах согласования и утверждения.
10. **Инструменты.** Приложения и иные технологии[1], позволяющие достигать целей данной области знаний.
11. **Методы.** Методы и процедуры, используемые для проведения работ и получения результатов в данной области знаний. К ним относятся, в частности, общие правила и соглашения, рекомендации лучших практик, стандарты и протоколы, а также (там, где это применимо) альтернативные новые подходы к решению задач.
12. **Метрики.** Показатели для измерения или оценки производительности, прогресса, качества, эффективности и других параметров. В этом блоке могут быть указаны как сугубо технические измеримые параметры работ, выполняемых в области знаний, так и более абстрактные обобщенные показатели наподобие улучшения или ценности.

Таким образом, колесо DAMA отражает состав областей знаний на верхнем уровне; шестиугольник на более низком уровне выделяет общие структурные компоненты для всех областей, а контекстные диаграммы позволяют представить детали этих компонентов для каждой области. Однако ни один из элементов рамочной структуры управления данными DAMA не описывает связи между различными областями знаний. В результате усилий, направленных на заполнение данного пробела, появились уточнения и дополнения, рамочной структуры DAMA, которые рассматриваются в двух следующих разделах.

3.4 Пирамида DMBOK (Айкен)

Во многих организациях убежденно говорят о желании извлечь максимальную выгоду из имеющихся в их распоряжении данных и стремятся к тому, чтобы делать всё возможное ради того, чтобы подняться на вершину золотой пирамиды управления данными, откуда открывается доступ к заветным передовым практикам (в частности, к извлечению данных (data mining), аналитике (analytics) и т. п.). Однако эта пирамида является лишь верхушкой сложной многоярусной структуры, возведенной на основательном фундаменте. Большинство организаций, увы, не могут позволить себе роскошь определить стратегию управления данными, прежде чем приступать к самому управлению. Вместо этого они выстраивают стратегию постепенно, в процессе практического управления данными, и делают это путем проб и ошибок в условиях, далеких от идеальных.

[1] DAMA International принципиально не занимается сертификацией или продвижением каких-либо программных либо аппаратных средств, инструментов или решений, равно как и их поставщиков.

Предложенная Питером Айкеном рамочная структура использует функциональные области DMBOK для описания различных ситуаций, в которых оказываются многие организации. С ее помощью организация может определить оптимальный для себя путь вперед — к состоянию, в котором у руководителей появятся надежные данные и процессы, в полной мере соответствующие стратегическим целям организации и способствующие их реализации. Для этого многие организации совершают примерно одну и ту же последовательность шагов (см. рис. 8).

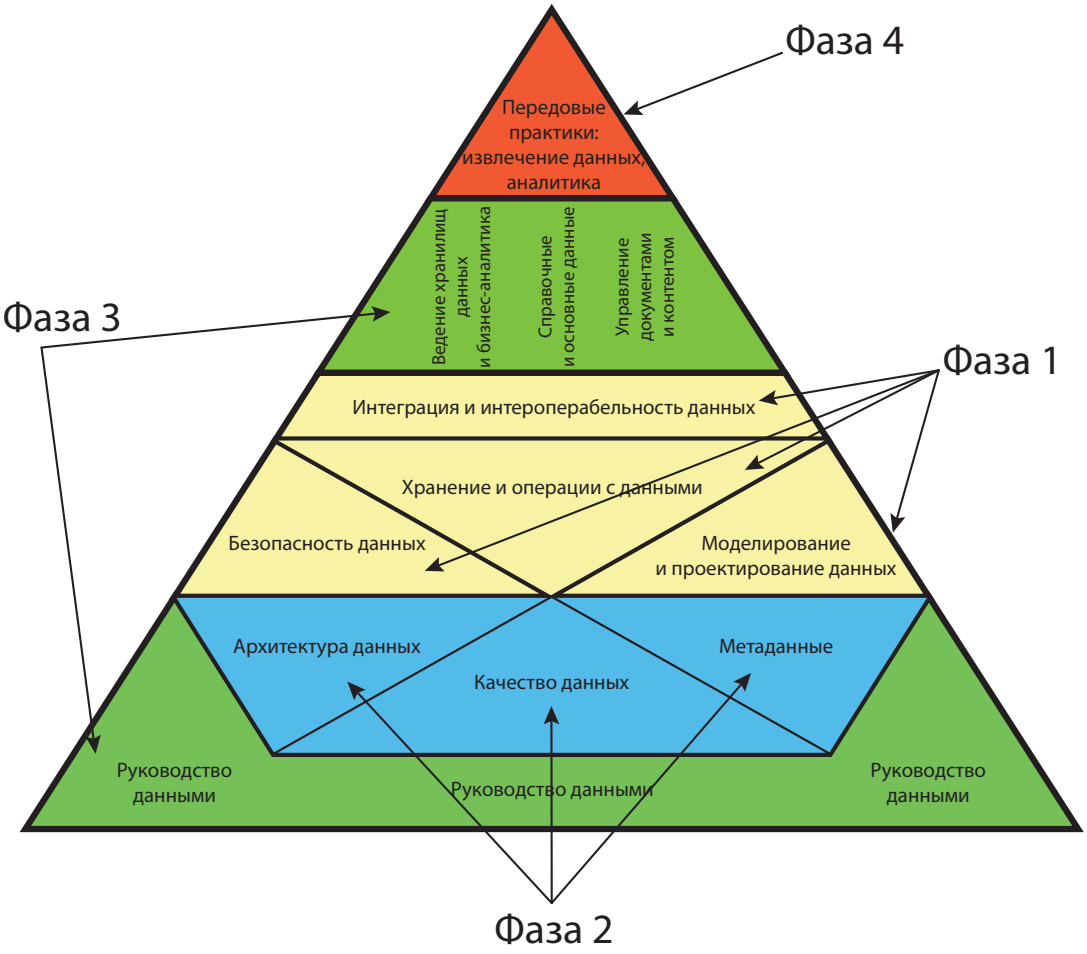


Рисунок 8. Пирамида Айкена: приобретенные или построенные возможности системы управления базами данных[1]

- **Фаза 1.** Организация покупает приложение с функциональной поддержкой возможностей системы управления базами данных. С этого момента она может приступить к моделированию/проектированию, хранению и обеспечению безопасности данных (например, путем выборочного доступа сотрудников к различным категориям данных и функциям). Для обеспечения

[1] Публикуется с разрешения Data BluePrint как владельца авторских прав на «золотую пирамиду Айкена».

функционирования системы в существующей ИТ-среде и работы с имеющимися данными необходимо проработать вопросы интеграции и интероперабельности.

- **Фаза 2.** Начав пользоваться приложением, в организации непременно столкнутся с проблемами качества имеющихся данных, и выяснится, что для его повышения не обойтись без надежных метаданных и логически стройной и непротиворечивой архитектуры данных. Именно они обеспечивают ясное понимание взаимодействия подсистем и обмена данными между ними.
- **Фаза 3.** Для согласованного и дисциплинированного управления качеством данных, метаданными и архитектурой потребуется функция руководства данными, которая обеспечивает структурную поддержку работ по управлению данными. Руководство данными также позволит перейти к реализации стратегических инициатив, таких как управление документами и контентом, управление справочными и основными данными, ведение хранилищ данных и бизнес-аналитика, что и откроет возможность для перехода к продвинутым практикам управления данными.
- **Фаза 4.** Организация в полной мере использует преимущества хорошо управляемых данных и расширяет возможности по осуществлению аналитической деятельности.

Пирамида Айкена, с одной стороны, заимствует информацию об областях знаний из колеса DAMA, а с другой — дополняет ее, показывая взаимосвязи между этими областями. Данные модели нельзя считать взаимозаменяемыми, поскольку на них отражены различные виды взаимозависимостей между областями. В рамочной структуре пирамиды в явном виде просматриваются две движущие идеи. Первая заключается в построении системы на едином прочном фундаменте из блоков, каждый из которых должен занять строго отведенное ему место, чтобы оказывать поддержку другим блокам. Вторая идея, несколько противоречащая первой, предполагает, что эти блоки можно встраивать в произвольном порядке.

3.5 Дальнейшая эволюция рамочной структуры управления данными DAMA

Пирамида Айкена описывает эволюцию организаций в направлении всё более совершенных и эффективных практик управления данными. Альтернативный взгляд на области знаний DAMA позволяет исследовать зависимости между ними. Разработанное Сью Гьюенс (Sue Geuens) новое представление рамочной структуры DAMA (см. рис. 9) отражает тот факт, что в иерархии взаимозависимостей функции бизнес-аналитики занимают место надстройки над всеми прочими функциями и всецело зависят от них. Они напрямую зависят от основных данных, а также решений в области хранилищ данных. А те, в свою очередь, зависимы от предоставляющих им данные систем и приложений. Основой надежности систем и приложений являются надлежащим образом организованные практики проектирования данных, обеспечения их качества, а также интеграции и интероперабельности. Все вышеперечисленные функции строятся на фундаменте руководства данными, которое в данной версии структуры включает следующие области: метаданные, безопасность данных, архитектура данных и справочные данные.

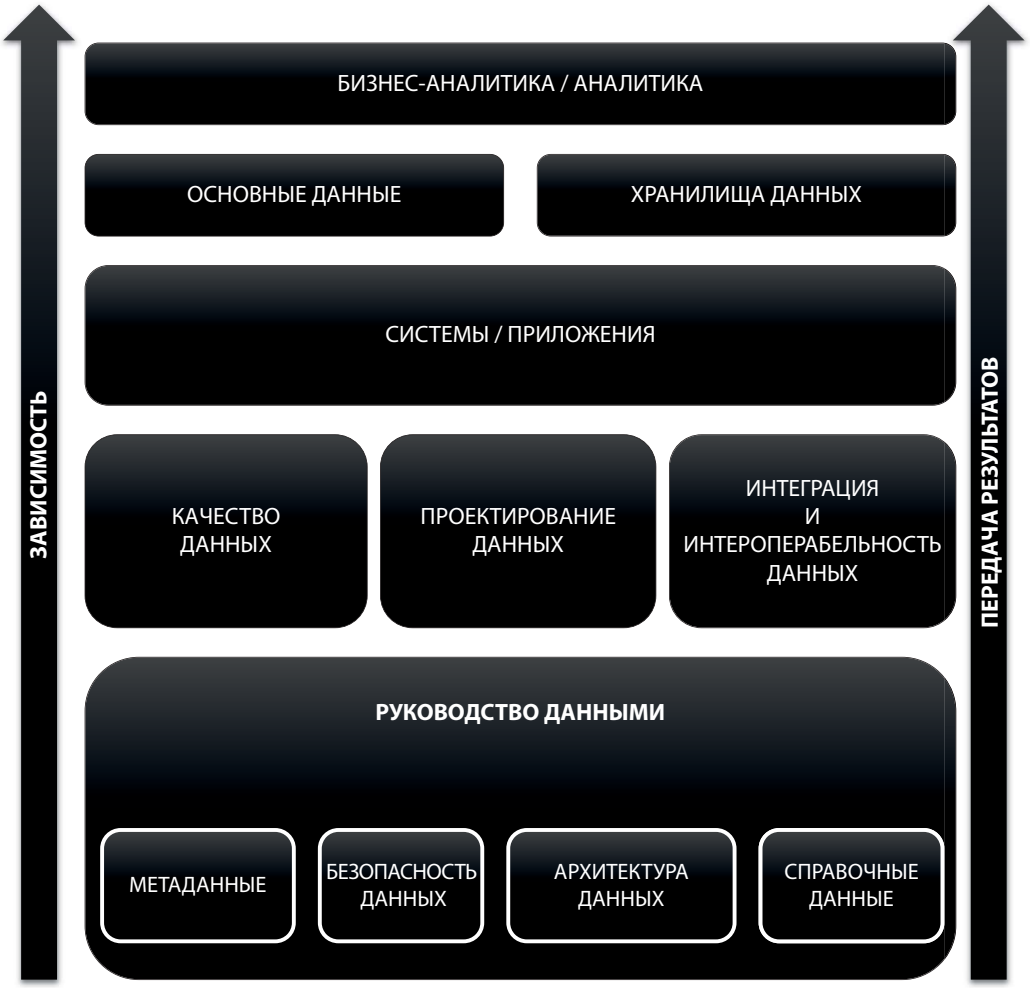


Рисунок 9. Взаимозависимости между функциональными областями рамочной структуры DAMA

Третья альтернатива колесу DAMA (см. рис. 10), подобно пирамиде Айкена, также представляет собой концепцию общей архитектуры областей знаний по управлению данными DAMA и их взаимосвязей. Она включает дополнительные детали, разъясняющие содержание некоторых областей, там, где это необходимо для более четкого понимания этих взаимосвязей.

Эта структура исходит из основной цели управления данными: предоставление организациям возможности извлекать выгоду из их информационных активов так же, как и из любого другого актива. Извлечение выгоды в первую очередь требует управления жизненным циклом данных, поэтому функции управления данными, относящиеся к жизненному циклу, помещены в центральную часть диаграммы. Начинается всё с моделирования и проектирования надежных и качественных данных; затем внедряются процессы и функции, обеспечивающие доступность данных для использования и их обслуживание; и, наконец, осуществляется использование данных в различных типах аналитики, за счет чего их ценность повышается.

Рисунок 10. Рамочная структура функций управления данными DAMA

Отведенный управлению жизненным циклом раздел включает функции по проектированию и операционные функции (моделирование, архитектура, хранение, обработка и т. п.), необходимые для поддержания традиционных способов применения данных (бизнес-аналитика (business intelligence), управление документами и контентом и т. п.). Кроме того, он учитывает и недавно

появившиеся функции — к примеру, хранение больших данных (big data), — которые необходимы для реализации новых возможностей использования данных (наука о данных — data science; предиктивная аналитика — predictive analytics, и т. п.). Когда организации действительно управляют данными как активом, у них появляется возможность извлечения прямой выгоды из своих данных посредством их продажи другим организациям (монетизация данных — data monetization).

Организации, фокусирующие внимание только на тех функциях, которые напрямую связаны с жизненным циклом, не смогут извлечь из своих данных такой же большой выгоды, какую они могли бы получить, осуществляя поддержку жизненного цикла с помощью остальных функций управления данными. Эти функции делятся на основополагающие направления деятельности и деятельность по надзору. Основополагающие направления деятельности (такие, как управление рисками в области данных, управление метаданными и качеством данных) охватывают весь жизненный цикл данных. Они позволяют реализовывать более качественные проектные решения и облегчают использование данных. Если деятельность по этим направлениям хорошо налажена, стоимость обслуживания данных снижается, потребители данных испытывают к ним больше доверия, а возможности их использования существенно расширяются.

Для успешной поддержки производства и использования данных, а также обеспечения осуществления основополагающей деятельности с требуемым уровнем исполнительской дисциплины многие организации устанавливают надзор за управлением данными в форме руководства данными. Программа руководства данными позволяет организации быть «управляемой на основе данных» благодаря наличию стратегии, подкрепленной принципами, политиками и практиками распоряжения данными, которые обеспечивают своевременное выявление и использование возможностей для извлечения выгоды из имеющихся данных. Программа руководства данными должна также предусматривать привлечение процессов управления организационными изменениями для обучения организации и поощрение поведения, ориентированного на стратегическое использование данных. Отсюда следует, что культурные изменения должны охватывать весь спектр деятельности по руководству данными, особенно в период становления практик управления данными.

Рамочную структуру управления данными DAMA также можно графически представить как результат эволюционного развития колеса DAMA. В данном варианте структуры функциональные элементы (направления деятельности) управления данными расположены следующим образом: ключевые элементы, совместно с элементами, относящимися к управлению жизненным циклом и использованию данных, окружены элементами, которые относятся к руководству данными (см. рис. 11).

Ключевые направления деятельности, включая управление метаданными, управление качеством данных, защиту данных и определение структуры данных (архитектура), расположены в верхней и нижней частях рамочной структуры.

Элементы управления жизненным циклом могут определяться как с позиции планирования данных (управление рисками, моделирование, проектирование, управление справочными данными), так и с позиции обеспечения их доступности (управление основными данными, разработка технологий работы с данными, интеграция и интероперабельность данных, ведение хранилищ данных, хранение и операции с данными).

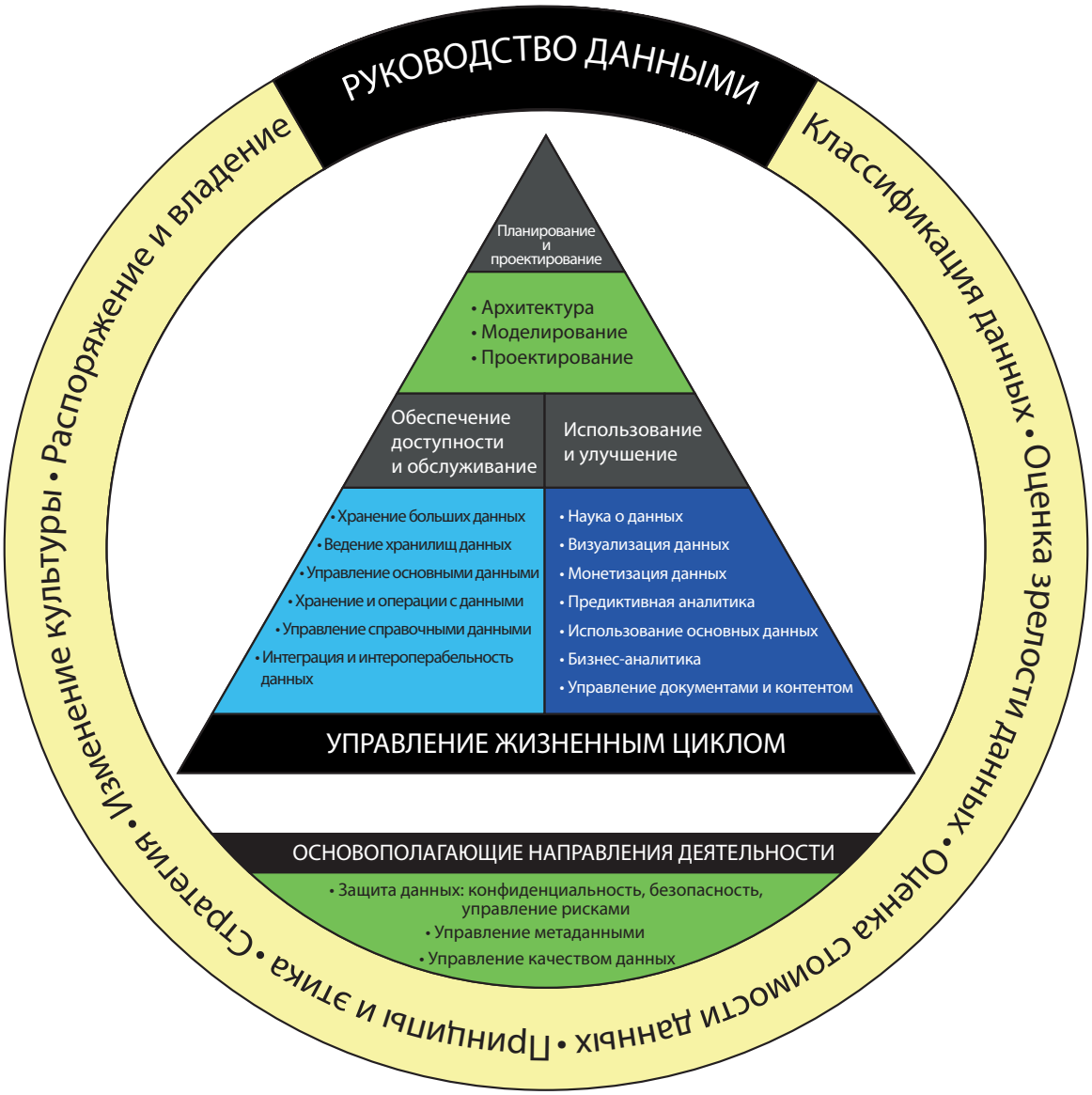


Рисунок 11. Колесо DAMA (развитие)

Использование данных напрямую связано со следующими направлениями деятельности по управлению жизненным циклом: использование основных данных, управление документами и контентом, бизнес-аналитика, наука о данных, предиктивная аналитика, визуализация данных. Многие из этих направлений деятельности порождают новые данные в результате улучшения или углубления понимания существующих данных. Реализацию возможностей по монетизации данных также можно рассматривать как их использование.

Руководство данными обеспечивает надзор и разумные ограничения посредством разработки и реализации стратегии, принципов и политик, а также с помощью практики распоряжения

данными. При этом улучшается согласованность всей деятельности по управлению данными за счет классификации и оценки данных.

Различные варианты визуального представления рамочной структуры управления данными DAMA рассмотрены с целью взглянуть на нее под всевозможными углами и положить начало дискуссии относительно возможных способов применения концепций, представленных в DMBOK. По мере возрастания важности управления данными подобные рамочные структуры становятся полезным инструментом коммуникации как внутри сообщества профессионалов в области управления данными, так и между сообществом и лицами, заинтересованными в нашей деятельности.

4. DAMA и DMBOK

Хотя управление данными сопряжено с рядом проблем, большинство из них не являются новыми. Как минимум с 1980-х годов организации признают, что управление данными — одна из основ их успеха. Как только увеличились наши способности и желание создавать и использовать данные, возросла и потребность в надежных практиках управления ими.

Ассоциация управления данными (DAMA) как раз и была учреждена для того, чтобы ответить на эти вызовы. DMBOK, доступное авторитетное справочное руководство, предназначенное для профессионалов в области управления данными, обеспечивает поддержку миссии DAMA посредством:

- **предоставления функциональной рамочной структуры** для внедрения корпоративных практик управления данными, включая руководящие принципы, широко распространенные практики, методы и приемы, функции, роли, результаты и метрики;
- **создания общего словаря терминов и понятий**, используемых в области управления данными, который служит основой для лучших практик;
- **выполнения роли базового справочного руководства** для подготовки к сдаче экзаменов на получение сертификата профессионала в области управления данными (Certified Data Management Professional, CDMP) и других сертификационных экзаменов.

Справочное руководство DMBOK выстроено вокруг одиннадцати областей знаний рамочной структуры управления данными DAMA-DMBOK (часто называемой колесом DAMA; см. рис. 5). Главы 3–13 посвящены детальному рассмотрению этих областей знаний. Все одиннадцать глав имеют одинаковую структуру:

1. Введение
   - ◇ Бизнес-драйверы
   - ◇ Цели и принципы
   - ◇ Основные понятия и концепции
2. Проводимые работы

3. Инструменты
4. Методы
5. Рекомендации по внедрению
6. Связь с руководством данными
7. Метрики

Каждая из областей знаний описывает состав и контекст характерных для нее работ по управлению данными. При этом фундаментальные цели и принципы управления данными остаются неизменными во всех без исключения областях знаний. Поскольку данные движутся по горизонтали внутри организаций, работы, проводимые в области знаний, пересекаются друг с другом и с прочими функциями[1].

1. **Руководство данными** обеспечивает осуществление руководящей деятельности и надзора в области управления данными, используемыми организацией, посредством создания и внедрения системы прав, обязанностей и отчетности (глава 3).
2. **Архитектура данных** определяет концептуальные решения по управлению информационными активами в соответствии со стратегией организации и устанавливает соответствующие стратегические требования к данным и проектным решениям в области данных (глава 4).
3. **Моделирование и проектирование данных** — процесс выявления, анализа, представления и распространения требований к данным в форме *модели данных* (глава 5).
4. **Хранение и операции с данными** включают проектирование и реализацию решений для хранения, а также сопровождение хранимых данных с целью максимального повышения их ценности. Операции обеспечивают сопровождение данных на протяжении всего их жизненного цикла — начиная с планирования и заканчивая ликвидацией (глава 6).
5. **Безопасность данных** гарантирует их целостность, конфиденциальность и защиту от несанкционированного доступа (глава 7).
6. **Интеграция и интероперабельность данных** включают процессы, относящиеся к обмену данными и консолидации данных как в рамках отдельных хранилищ данных, приложений и организаций, так и между ними (глава 8).
7. **Управление документами и контентом** включает планирование, реализацию и контроль мероприятий по управлению жизненным циклом неструктурированных данных и информации, зафиксированных на всевозможных носителях и в разнообразных форматах; особое внимание уделяется документам, необходимым для подтверждения соблюдения организацией требований нормативно-правового регулирования (глава 9).

[1] В представленных далее описаниях одиннадцати областей знаний их названия следует воспринимать именно как фиксированные названия областей знаний, принятые в DMBOK, а не как названия конкретных процессов или видов данных. Поэтому допускаются выражения типа «Справочные и основные данные включают согласование».
Это замечание распространяется и на остальной текст, поскольку подобные формулировки часто встречаются и в следующих главах. — *Примеч. науч. ред.*

8. **Справочные и основные данные** включают согласование и ведение на постоянной основе критически важных совместно используемых данных в целях обеспечения скоординированного применения всеми системами наиболее точной, полной и актуальной «версии правды» о ключевых бизнес-сущностях (глава 10).
9. **Ведение хранилищ данных и бизнес-аналитика** включают планирование, реализацию и контроль процессов, обеспечивающих управление данными, используемыми для поддержки принятия решений, а также позволяющих специалистам извлекать ценность из данных с помощью анализа и построения отчетов (глава 11).
10. **Метаданные** предусматривают планирование, реализацию и контроль деятельности по обеспечению доступа к высококачественным, интегрированным метаданным, включая определения, модели, описания потоков данных и другую информацию, необходимую для понимания данных, а также систем, используемых для создания, ведения и доступа к ним (глава 12).
11. **Качество данных** включает планирование и внедрение методических решений по управлению качеством данных, обеспечивающих измерение, оценку и повышение качества, включая контроль практической пригодности данных к использованию в организации (глава 13).

Помимо глав, посвященных областям знаний, DAMA-DMBOK включает следующие тематические главы:

- **Этика обращения с данными** описана как занимающая центральное место и играющая важнейшую роль в принятии информированных, социально ответственных решений относительно допустимости сбора тех или иных данных и/или их использования по определенным назначениям. Понимание и соблюдение этических требований обязательно для специалистов по управлению данными на всех этапах их сбора, анализа и использования (глава 2).
- **Большие данные и наука о данных** описывают технологии и бизнес-процессы, которые появляются по мере нарастания способности организаций собирать и анализировать всё более объемные массивы самых разнородных данных в разнообразных областях (глава 14).
- **Оценка зрелости управления данными** очерчивает подход к оценке текущего состояния управления данными и выявлению возможностей для дальнейшего совершенствования процессов (глава 15).
- **Организация управления данными и ролевые ожидания** представляют лучшие практики и приводят полезные соображения относительно оптимальной организации команд по управлению данными и внедрения в организации успешных практик управления данными (глава 16).
- **Управление данными и управление организационными изменениями** описывает рецепты планирования успешного продвижения через культурные изменения, необходимые для внедрения в организации эффективных практик управления данными (глава 17).

Как именно конкретная организация управляет данными, зависит от ее целей, размера, ресурсов, характера деятельности и сложности структуры, равно как и от общего восприятия роли и функций данных в стратегии развития. Большинство организаций не реализуют у себя весь описанный в настоящем руководстве комплекс мер в каждой из областей знаний. Однако понимание максимально широкого контекста управления данными поможет им решить, на какие области следует обратить первоочередное внимание в плане совершенствования работы по каждому из направлений и координации межфункциональных усилий в этой сфере.

5. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Abcouwer, A. W., Maes, R., Truijens, J. «Contouren van een generiek Model voor Informatienmanagement». Primavera Working Paper 97–07, 1997, http://bit.ly/2rV5dLx

Adelman, Sid, Larissa Moss, and Majid Abai. *Data Strategy*. Addison-Wesley Professional, 2005. Print.

Aiken, Peter and Billings, Juanita. *Monetizing Data Management*. Technics Publishing, LLC, 2014. Print.

Aiken, Peter and Harbour, Todd. *Data Strategy and the Enterprise Data Executive*. Technics Publishing, LLC. 2017. Print.

APRA (Australian Prudential Regulation Authority). Prudential Practice Guide CPG 234, Management of Security Risk in Information and Information Technology. May 2013, http://bit.ly/2sAKe2y

APRA (Australian Prudential Regulation Authority). *Prudential Practice Guide CPG 235, Managing Data Risk*. September 2013, http://bit.ly/2sVIFil

Borek, Alexander et al. *Total Information Risk Management: Maximizing the Value of Data and Information Assets*. Morgan Kaufmann, 2013. Print.

Brackett, Michael. *Data Resource Design: Reality Beyond Illusion*. Technics Publishing, LLC. 2014. Print.

Bryce, Tim. *Benefits of a Data Taxonomy*. Blog 2005-07-11, http://bit.ly/2sTeU1U

Chisholm, Malcolm and Roblyn-Lee, Diane. *Definitions in Data Management: A Guide to Fundamental Semantic Metadata*. Design Media, 2008. Print.

Devlin, Barry. *Business Unintelligence*. Technics Publishing, LLC. 2013. Print.

English, Larry. *Improving Data Warehouse and Business Information Quality: Methods For Reducing Costs And Increasing Profits*. John Wiley and Sons, 1999. Print.

Evans, Nina and Price, James. «Barriers to the Effective Deployment of Information Assets: An Executive Management Perspective». *Interdisciplinary Journal of Information, Knowledge, and Management.* Volume 7, 2012. Accessed from http://bit.ly/2sVwvG4

Fisher, Tony. *The Data Asset: How Smart Companies Govern Their Data for Business Success*. Wiley, 2009. Print. Wiley and SAS Business Ser.

Henderson, J. C., Venkatraman, H. «Leveraging information technology for transforming Organizations». *IBM System Journal*. Volume 38, Issue 2.3, 1999 [1993 Reprint], http://bit.ly/2sV86Ay and http://bit.ly/1uW8jMQ

Kent, William. *Data and Reality: A Timeless Perspective on Perceiving and Managing Information in Our Imprecise World*. 3d ed. Technics Publications, LLC, 2012. Print.

Kring, Kenneth L. *Business Strategy Mapping — The Power of Knowing How it All Fits Together*. Langdon Street Press (a division of Hillcrest Publishing Group, Inc.), 2009. Print.

Loh, Steve. *Data-ism: The Revolution Transforming Decision Making, Consumer Behavior, and Almost Everything Else*. HarperBusiness, 2015. Print.

Loshin, David. *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann, 2001. Print.

Maes, R. «A Generic Framework for Information Management». PrimaVera Working Paper 99–02, 1999.

McGilvray, Danette. *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. Morgan Kaufmann, 2008. Print.

McKnight, William. *Information Management: Strategies for Gaining a Competitive Advantage with Data*. Morgan Kaufmann, 2013. Print. The Savvy Manager's Guides.

Moody, Daniel and Walsh, Peter. «Measuring The Value Of Information: An Asset Valuation Approach». *European Conference on Information Systems (ECIS)*, 1999, http://bit.ly/29JucLO

Olson, Jack E. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, 2003. Print.

Redman, Thomas. «Bad Data Costs U.S. $3 Trillion per Year». *Harvard Business Review*. 22 September 2016. Web.

Redman, Thomas. Data Driven: Profiting from Your Most Important Business Asset. *Harvard Business Review Press*. 2008. Print.

Redman, Thomas. *Data Quality: The Field Guide*. Digital Press, 2001. Print.

Reid, Roger, Gareth Fraser-King, and W. David Schwaderer. *Data Lifecycles: Managing Data for Strategic Advantage*. Wiley, 2007. Print.

Rockley, Ann and Charles Cooper. *Managing Enterprise Content: A Unified Content Strategy*. 2nd ed. New Riders, 2012. Print. Voices That Matter.

Sebastian-Coleman, Laura. *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*. Morgan Kaufmann, 2013. Print. The Morgan Kaufmann Series on Business Intelligence.

Simsion, Graeme. *Data Modeling: Theory and Practice*. Technics Publications, LLC, 2007. Print.

Surdak, Christopher. *Data Crush: How the Information Tidal Wave is Driving New Business Opportunities*. AMACOM, 2014. Print.

Waclawski, Janine. *Organization Development: A Data-Driven Approach to Organizational Change*. Pfeiffer, 2001. Print.

White, Stephen. Show Me the Proof: Tools and Strategies to Make Data Work for the Common Core State Standards. 2nd ed. Advanced Learning Press, 2011. Print.

Этика обращения с данными

1. ВВЕДЕНИЕ

Согласно простейшему определению, *этика* — это принципы поведения, основанные на понимании того, что такое хорошо и что такое плохо. Этические принципы часто строятся вокруг идей о справедливости, уважении, ответственности, честности, целостности, надежности, прозрачности и доверии. Соответственно, этика обращения с данными занимается определением правил получения, хранения, управления, использования и ликвидации данных, которые в полной мере соответствуют общечеловеческим этическим принципам и нормам. Этичное обращение с данными — непременное условие долгосрочного успеха организации, желающей извлекать выгоду для себя из имеющихся в ее распоряжении информационных активов. Ведь неэтичное обращение с данными может привести к потере репутации и клиентов, поскольку подвергает риску людей, данные которых раскрыты. В некоторых случаях неэтичное обращение с данными считается незаконным[1]. В конце концов, для профессионалов в области управления данными и организаций, на которые они работают, этичное обращение с данными — часть их социальной ответственности.

Этика обращения с данными включает достаточно сложно устроенный комплекс правил, но все они выстроены вокруг трех ключевых понятий.

- **Воздействие на людей.** Поскольку данные отражают индивидуальные характеристики и особенности людей и используются для принятия решений, которые влияют на человеческие жизни, крайне важно управлять их качеством и надежностью.

[1] «Закон о преемственности и учете данных в медицинском страховании» (Health Insurance Portability and Accountability Act [HIPAA], Pub. L. 104–191, 110 Stat. 1936) в США (1996), «Закон о защите личных сведений и электронных документов» (Personal Information Protection and Electronic Documents Act [PIPEDA]) в Канаде (2000), «Общий регламент по защите данных» (General Data Protection Regulation [GDPR]) в ЕС (2016) и другие национальные законы о защите данных / конфиденциальности информации [в частности, Федеральный закон РФ № 152-ФЗ «О персональных данных» (2006). — *Примеч. пер.*] описывают обязанности, связанные с обработкой персональных идентификационных данных (например, ФИО, адресов, религиозной принадлежности, сексуальной ориентации и т. п.) и обеспечением неразглашения (ограничения доступа) подобной информации.

- **Риск злоупотребления данными.** Нецелевое использование данных может негативно сказываться на отдельных людях или организациях. Таким образом, предотвращать возможность злоупотреблений — этический долг специалистов, работающих с данными.
- **Экономическая ценность данных.** Поскольку данные имеют экономическую ценность, этические нормы владения данными должны определять, кому и каким образом эта ценность может быть доступна.

Организации обеспечивают защиту данных, руководствуясь прежде всего действующими правовыми актами и требованиями регулирующих органов. Тем не менее, поскольку данные касаются людей (клиентов, сотрудников, пациентов, поставщиков и т. д. и т. п.), профессионалы в области управления данными должны отдавать себе отчет в том, что имеются и этические (наряду с юридическими) причины обеспечивать защиту данных и не допускать злоупотреблений. Данные, на первый взгляд не отражающие никаких сведений о физических лицах, всё равно могут быть использованы для принятия решений, которые в конечном счете навредят людям.

Этическим императивом является не только защита, но и управление качеством данных. И те, кто принимает решения, и те, на ком эти решения сказываются, вправе рассчитывать на полноту и точность данных, служащих основанием для принятия того или иного решения. И с деловой, и с технической точки зрения профессионалы в области управления данными обязаны по этическим соображениям выполнять свою работу с минимально возможным риском искажения, неверного представления, нецелевого использования или ошибочной интерпретации данных. И эта обязанность охватывает весь жизненный цикл данных — от их создания до ликвидации.

К сожалению, во многих организациях этого не понимают и, как следствие, не принимают всех необходимых мер по исполнению своих этических обязанностей в сфере управления данными. Где-то по традиции занимают чисто техническую позицию и делают вид, будто им не до того, чтобы разбираться со смыслом обрабатываемых данных; где-то исходят из представления, что достаточно соблюдать букву закона — и никакого риска от данных, имеющихся в их распоряжении, не возникнет. Это весьма опасные заблуждения.

Среда, в которой обрабатываются и используются данные, стремительно эволюционирует и развивается. Сегодня организации находят такие применения данным, о возможности которых еще пару лет назад никто даже не догадывался. Законы вроде бы и предписывают некоторые этические принципы управления данными, но угнаться за всё новыми и новыми рисками, связанными со стремительным эволюционным развитием информационных технологий, законодатели пока не в состоянии. Организации должны отдавать себе отчет в том, что защита доверенных им данных — их этический долг, и развивать соответствующую корпоративную культуру, в которой ценится этичное обращение с информацией.

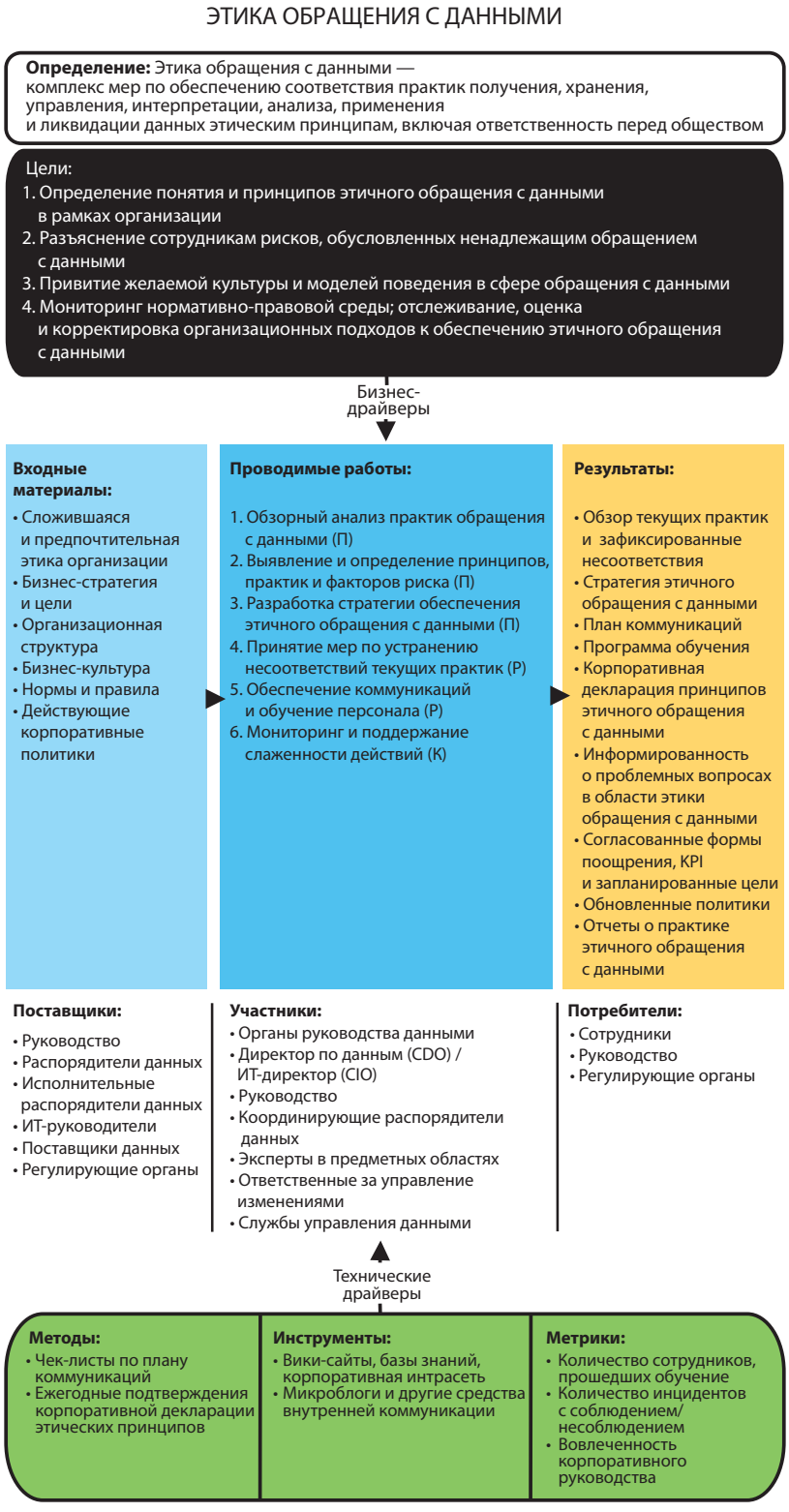


Рисунок 12.
Контекстная диаграмма: этика обращения с данными

2. БИЗНЕС-ДРАЙВЕРЫ

К этике в полной мере применим один из сформулированных Эдвардом Демингом[1] принципов управления качеством: «Всё должно делаться правильно даже при отсутствии надзора». В бизнесе этичный подход к данным всё чаще рассматривается в качестве конкурентного преимущества (Hasselbalch and Tranberg, 2016). Соблюдение норм этики при обращении с данными повышает доверие к организации, а также к ее данным и результатам деятельности. А это, в свою очередь, способствует улучшению отношений между организацией и заинтересованными лицами. Создание этичной культуры включает в себя обеспечение надлежащего руководства данными, в том числе учреждение структур, призванных контролировать и обеспечивать этичность как планируемых, так и достигнутых результатов обработки данных и не допускать, чтобы они подрывали доверие или могли интерпретироваться как посягательство на человеческое достоинство.

Обращение с данными происходит не в вакууме, и клиенты и другие заинтересованные лица ожидают этичного подхода и таких же этичных результатов от организаций и их процессов обработки данных. Снижение риска злоупотребления данными со стороны сотрудников, клиентов или партнеров служит ключевой причиной, по которой организации следует приступить к культивированию принципов этичного обращения с данными. Также организация несет моральную ответственность за обеспечение защиты данных от посягательств со стороны правонарушителей (защита от взлома и утечек; см. главу 7).

Различные модели владения данными по-разному влияют и на этику обращения с ними. К примеру, развитие технологий значительно расширило возможности организаций в плане обмена данными и их совместного использования. Но новые возможности обязывают организации еще этичнее относиться к принятию ответственных решений относительно дальнейшего распространения данных, права на которые принадлежат не им или не им одним.

Недавно появившиеся во многих организациях должности директора по управлению данными (Chief Data Officer, CDO), директора по рискам (Chief Risk Officer, CRO), директора по вопросам конфиденциальности информации (Chief Privacy Officer, CPO) и директора по аналитике (Chief Analytics Officer, CAO) нужны для того, чтобы всецело фокусироваться на контроле рисков за счет установления соответствующей практики обращения с данными. Но ответственность лежит не только на этих должностных лицах. Гарантией этичного обращения с данными может служить лишь понимание всеми сотрудниками организации рисков, обусловленных злоупотреблением данными, а также приверженность организации практике обращения с ними на основе принципов, обеспечивающих защиту индивидуумов, и нравственных обязательств, связанных с владением данными.

[1] Уильям Эдвардс («Эдвард») Деминг (*англ.* William Edwards «Edward» Deming, 1990–1993) — американский специалист по математической физике и классик научного управления качеством. — *Примеч. пер.*

3. ОСНОВНЫЕ ПОНЯТИЯ И КОНЦЕПЦИИ

3.1 Этические принципы, связанные с данными

Общепринятые постулаты биоэтики, призванные обеспечить сохранение человеческого достоинства, служат хорошей отправной точкой и для формулировки основополагающих принципов этики обращения с данными. К примеру, белмонтские этические принципы[1] вполне можно с небольшими доработками перенести с биомедицинских исследований на дисциплины управления информацией (US-HSS, 1979).

- **Уважение к личности.** Этот принцип отражает фундаментальное этическое требование уважительного отношения к людям, исключающего ущемление их человеческого достоинства и ограничение личных свобод и независимости. Кроме того, в случаях неизбежного «ограничения самостоятельности» защите достоинства и прав человека должно уделяться повышенное внимание.

Рассматривая данные в качестве актива, помним ли мы о том, что это данные о живых людях и их использование неизбежно сказывается на них и их интересах? Персональные данные, в частности, — это ведь принципиально иного рода «ресурс», нежели нефть или уголь. Неэтичное использование имеющихся у нас персональных данных может испортить отношения людей с окружающими, лишить работы, а то и вовсе места в обществе. Не ограничивают ли разрабатываемые нами информационные системы независимость, самостоятельность или свободу выбора? Учитываем ли мы, каким образом результаты обработки наших данных скажутся на людях с ограниченными умственными или физическими возможностями? Предусмотрели ли мы равные права по получению доступа и использованию данных для этих категорий граждан? Наконец, ведется ли обработка персональных данных на основе явного, информированного и имеющего юридическую силу согласия?

- **Обеспечение безусловной пользы.** Этот принцип включает два составных элемента: во-первых, не навредить; во-вторых, принести максимальную пользу и минимизировать потенциальный ущерб.

«Не навреди» — древнейший принцип врачебной этики, но ведь он явным образом применим и в контексте управления данными и информацией. Специалисты по работе с данными, руководствующиеся этическими принципами, должны выявлять всех заинтересованных лиц и проектировать процессы обработки данных таким образом, чтобы полученные результаты приносили

[1] Речь идет об этических принципах и руководствах, сформулированных в опубликованном в 1978 году «Белмонтском докладе» Национальной комиссии по защите прав людей — субъектов медико-биологических и поведенческих исследований (*Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, Report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research*). — *Примеч. пер.*

максимальную пользу, а риски причинения вреда были минимальными. На что изначально нацелен процесс обработки данных — на выигрыш за счет проигравших или на создание взаимовыгодных условий для сторон? Не слишком ли агрессивна процедура обработки данных и не имеется ли менее рискованных альтернатив удовлетворения потребностей бизнеса в информации? Достаточно ли прозрачно организовано обращение с данными и не кроются ли в сложившихся процессах источники потенциального вреда для людей?

- **Справедливость.** Этот принцип предписывает в равной мере справедливое и объективное отношение ко всем, то есть безусловное равноправие.

По поводу этого принципа могут возникнуть некоторые вопросы, в частности: равноправие распространяется на всех без исключения и в любых обстоятельствах или допускается дифференциация по группам людей и/или обстоятельствам при сохранении равенства в пределах каждой обособленной категории лиц/обстоятельств? Не приводит ли конечный результат на выходе процесса или алгоритма к непропорциональному ущербу или выгоде для какой-либо конкретной группы людей? Не используются ли для машинного обучения наборы данных, способные повлиять на формирование предвзятого отношения к определенным социокультурным группам?

В Менлоском докладе Министерства внутренней безопасности США белмонтские принципы адаптированы к исследованиям в области информационных и коммуникационных технологий. При этом добавлен четвертый этический принцип (US-DHS, 2012):

- Уважение закона и общественных интересов.

В 2015 году Европейский инспектор по защите данных (European data protection supervisor, EDPS) опубликовал экспертное заключение по вопросам цифровой этики, в котором особое внимание уделено «инженерным, мировоззренческим, правовым и моральным последствиям» новейших тенденций в области обработки данных и, в частности, больших данных. В документе содержится призыв целенаправленно сфокусироваться на обеспечении уважения к человеческому достоинству при обработке данных и сформулированы четыре принципа, гарантирующие этичное обращение с данными в информационных экосистемах (EDPS, 2015).

- Правила обработки данных и соблюдения прав на неприкосновенность частной жизни и защиту личных данных должны разрабатываться с ориентацией на будущее.
- Лица, ответственные за разработку правил обработки персональных данных, обязаны сохранять прозрачность своей деятельности.
- Проектирование и разработка продуктов и услуг, связанных с обработкой данных, должны учитывать требования конфиденциальности.
- Права и полномочия лиц, которых затрагивают процессы обработки данных, необходимо расширять.

Эти принципы в широкой трактовке вполне согласуются с белмонтскими, поскольку также нацелены прежде всего на защиту человеческого достоинства и свободы личности. EDPS причисляет право на неприкосновенность личной информации к фундаментальным правам человека. Тем самым перед разработчиками инновационных решений в сфере управления данными ставится непростая задача: рассматривать принципы защиты человеческого достоинства, неприкосновенности и свободы личности в качестве базовой платформы, на которой выстраивается устойчивая цифровая среда, а не в качестве досадных помех или технических препятствий, которые нужно как-то обойти. Кроме того, документ призывает их к прозрачности и постоянному информационному обмену со всеми заинтересованными сторонами.

В таком контексте руководство данными становится важнейшим инструментом обеспечения учета и соблюдения вышеназванных этических принципов в решениях, определяющих, кто именно и что именно имеет право делать с теми или иными данными и при каких условиях такая обработка допустима или необходима. Специалисты-практики должны в обязательном порядке учитывать этические последствия и риски, возникающие при обработке данных, для всех заинтересованных сторон и управлять ими так же, как и качеством данных.

3.2 Основополагающие принципы законодательства о конфиденциальности данных

Усилия по определению правильных и неправильных с точки зрения этики подходов к обращению с данными предпринимаются и в рамках публичной политики, и на законодательном уровне. Но законотворцы и регулирующие органы не в силах предусмотреть в своих законах и кодексах все варианты развития событий и стечения обстоятельств. К примеру, законы о неприкосновенности частной жизни и конфиденциальности информации в Евросоюзе, Канаде и США основаны на весьма различных подходах к кодификации этических принципов обращения с данными. Тем не менее их можно взять за основу политики организации.

Законы о неприкосновенности частной жизни и неразглашении конфиденциальной информации появились уже давно, причем сами эти концепции восходят к этическому императиву уважения и соблюдения прав человека. Еще в 1890 году американские правоведы Сэмюэл Уоррен и Луи Брэндайс[1] назвали неприкосновенность частной жизни и конфиденциальность информации о ней неотъемлемой частью прав человека, гарантированных Конституцией США. В 1973 году Федеральной торговой комиссией был утвержден первый вариант «Принципов честной информационной практики», а год спустя концепция конфиденциальности информации как фундаментального права была подтверждена в федеральном «Законе о неприкосновенности частной жизни»[2], прямо

[1] Сэмюэл Уоррен (*англ.* Samuel Dennis Warren II, 1852–1910) и Луи Брэндайс (*англ.* Louis Dembitz Brandeis, 1856–1941) — однокурсники по Гарвардской юридической школе (выпуск 1877 г.) и соучредители (1879) преуспевающей адвокатской конторы, здравствующей и поныне под вывеской *Nutter McClennen & Fish*, отличавшиеся прогрессивными для своего времени взглядами и опубликовавшие в 1890 г. эссе «Право на частную жизнь» («The Right to Privacy», 4, *Harvard L. R.* 193, Dec. 15, 1890), считающееся первым трудом с обоснованием права на неприкосновенность частной жизни. — *Примеч. пер.*

[2] Privacy Act of 1974 (Pub. L. 93–579, 88 Stat. 1896, enacted December 31, 1974, 5 U. S. C. § 552a). — *Примеч. пер.*

установившем, что «право на неприкосновенность частной жизни является основополагающим правом личности и охраняется Конституцией США».

Подписанная в 1950 году на волне всеобщей обеспокоенности массовыми нарушениями прав человека в годы Второй мировой войны «Европейская конвенция по правам человека»[1] включила в число фундаментальных прав и свобод человека как общее право на уважение и неприкосновенность частной (личной и семейной) жизни, жилища и корреспонденции (ст. 8), так и право на защиту репутации и неразглашение конфиденциальной информации (ст. 10), то есть на защиту персональных данных.

В 1980 году Организация экономического сотрудничества и развития (ОЭСР) установила основные принципы защиты данных в «Руководстве по честной обработке информации и трансграничных потоках персональных данных»[2], которые затем легли в основу законов ЕС о защите данных.

Установленные ОЭСР восемь базовых принципов обработки персональных данных предусматривают их применение на национальном уровне в целях обеспечения повсеместного уважения прав граждан на неприкосновенность их частной жизни и включают: принцип ограничения объема собираемых данных; принцип обеспечения качества данных; принцип конкретизации целей (перед сбором данных); принцип ограниченного использования (строго по целевому назначению и с согласия субъекта данных или по законному требованию властей); принцип обеспечения безопасности; принцип открытости (прозрачности); принцип личного участия (право доступа к процедурам и данным с целью убедиться в их достоверности и потребовать уничтожения недостоверных данных); принцип подотчетности организаций, контролирующих сбор и обработку данных, и их ответственности за обеспечение соблюдения всего комплекса этих принципов.

Принципы ОЭСР впоследствии сменились принципами, лежащими в основе «Общего регламента по защите данных» ЕС (GDPR, 2016); см. таблицу 1.

Таблица 1. Принципы GDPR

Принцип GDPR	Описание принципа
Честность, законность, прозрачность	Персональные данные должны обрабатываться согласно букве закона, честно и прозрачно в отношении субъектов данных
Ограничение целей использования	Персональные данные должны собираться лишь в строго и явным образом оговоренных законных целях и не могут обрабатываться или использоваться как-либо иначе, чем это диктуется их целевым назначением

[1] Имеющая полное название «Конвенция о защите прав человека и основных свобод» (*англ.* Convention for the Protection of Human Rights and Fundamental Freedoms) была подписана 04.11.1950 г., вступила в силу 03.11.1953 г., ратифицирована РФ 30.03.1998 г. (кроме протоколов 6, 12, 13 и 16). — *Примеч. пер.*

[2] «Руководство по честной обработке информации и трансграничным потокам персональных данных» (*OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, 1980) впоследствии легло в основу дополненной и переработанной «Концепции по защите неприкосновенности частной жизни OECD» (*The OECD Privacy Framework*, 2013). — *Примеч. пер.*

Принцип GDPR	Описание принципа
Минимизация объема данных	Персональные данные должны быть адекватными и релевантными, а объем собираемых данных не должен превышать минимума, необходимого для их целевого использования
Точность данных	Персональные данные должны быть точными и при необходимости своевременно обновляться. Должны предприниматься все разумные меры по оперативному выявлению и незамедлительному удалению или исправлению неточных данных
Ограничение сроков хранения	Персональные данные должны храниться в форме, допускающей идентификацию субъектов данных [физических лиц], не дольше, чем это необходимо для их обработки
Целостность и конфиденциальность	Процессы обработки персональных данных должны обеспечивать их надлежащую защиту от несанкционированного доступа, обработки и использования в незаконных целях, утечки, случайного уничтожения или повреждения. Для этого должны быть приняты все доступные технические и организационные меры
Ответственность и подотчетность	Должностные лица, ответственные за защиту данных, обязаны соблюдать все сформулированные принципы и отчитываться о принятых мерах по защите персональных данных с предоставлением подтверждающих документов

Эти принципы сбалансированы, в частности, и за счет в явном виде прописанных гарантий определенных прав физических лиц в отношении собранных о них персональных данных. Среди них — права доступа, уточнения и исправления неверных данных, переноса данных, а также право отказывать в обработке персональных данных, которые могут причинить вред, и требовать их уничтожения. В тех случаях, когда сбор и обработка персональных данных требуют явного согласия физического лица, такое согласие должно носить характер положительного, свободного и недвусмысленного волеизъявления после получения субъектом исчерпывающей и конкретной информации о целях, назначении, порядке обработки и гарантиях защиты данных. GDPR также требует эффективного руководства данными и документального подтверждения соблюдения принципов «проектируемой конфиденциальности» (privacy by design)[1].

В Канаде «Закон о защите личных сведений и электронных документов» (PIPEDA[2]) сочетает всеобъемлющий правовой режим обеспечения неприкосновенности частной жизни, включая защиту персональных данных, с отраслевым саморегулированием. Действие PIPEDA распространяется на все без исключения организации, которые практикуют сбор, использование и распространение личных сведений в процессе своей коммерческой деятельности. Правила, устанавливаемые PIPEDA[3] (табл. 2), носят характер юридически обязательных для исполнения всеми

[1] «Проектируемая конфиденциальность» представляет собой концепцию интеграции мер защиты персональных данных на этапе проектирования системы их обработки. При этом защита персональных данных рассматривается как одна из основных функций системы. — *Примеч. науч. ред.*

[2] *сокр.* от *англ.* Personal Information Protection and Electronic Documents Act.

[3] См.: http://bit.ly/2tNM53c

организациями, использующими персональную информацию о своих потребителях (кроме случаев, подпадающих под явным образом сформулированные исключения).

Таблица 2. Законодательно установленные в Канаде обязанности организаций по защите персональных данных

Принцип PIPEDA	Описание принципа
Ответственность и подотчетность	Организация отвечает за имеющиеся в ее распоряжении личные сведения и должна назначить лицо, ответственное за обеспечение их защиты в соответствии с этими принципами
Раскрытие целей	Организация должна в явном виде указывать цели сбора личной информации до начала ее фактического сбора или получения
Согласие	Организация должна удостовериться в понимании физическим лицом целей и порядка сбора, использования или раскрытия личных сведений и получить согласие с ними, за исключением случаев, когда это требование невыполнимо
Ограничение сбора, использования, раскрытия и хранения	Сбор личных сведений должен ограничиваться лишь теми данными, которые необходимы для использования в заявленных организацией целях, и вестись честными и законными средствами. Личные сведения не подлежат использованию или раскрытию в иных целях, кроме заявленных при сборе, за исключением случаев согласия самого лица или требования закона. Личные сведения могут храниться лишь до тех пор, пока они необходимы для использования в заявленных целях
Точность	Личные сведения до лжны быть точными, полными и по мере необходимости обновляемыми для использования по целевому назначению
Защитные механизмы	Личные сведения подлежат защите с использованием механизмов, соответствующих степени чувствительности информации
Открытость	Организация должна обеспечивать доступность информации о своих правилах и практиках управления личными сведениями для физических лиц
Личный доступ	По запросу физического лица ему должна предоставляться информация о наличии у организации личных сведений об этом лице, их использовании и раскрытии, а также доступ к этим сведениям. Физическому лицу должна быть гарантирована возможность оспаривать точность и полноту личных сведений о себе, а неточности и пробелы должны по мере возможности устраняться
Возможность оспаривания нарушений	У физического лица должна иметься возможность оспаривать нарушение любого из вышеизложенных принципов путем обращения к уполномоченному лицу или лицам в структуре организации, отвечающим за обеспечение соблюдения этих принципов

Кроме того, в Канаде существует должность Федерального уполномоченного по вопросам конфиденциальности, аппарат которого выполняет все функции по разбору жалоб граждан на организации. Однако этот уполномоченный — не более чем омбудсмен, и его решения носят исключительно рекомендательный характер и юридически не обязательны к исполнению даже его подчиненными (и, тем более, не создают правовых прецедентов).

В марте 2012 года Федеральная торговая комиссия США (FTC), отвечающая за защиту прав потребителей, выпустила доклад с детальными рекомендациями организациям относительно проектирования и реализации собственных программ защиты персональных данных, основанные на лучших практиках, описанных в докладе («проектируемая конфиденциальность») (FTC 2012). Доклад подтвердил приверженность Комиссии принципам честной информационной практики (табл. 3).

Таблица 3. Применяемые в США критерии программы защиты персональных данных

Принцип FTC	Описание принципа
Уведомление/ Осведомленность	Ответственные за сбор данных лица должны раскрывать сведения о своей практике сбора, обработки и использования персональных данных и сведений личного характера до начала их получения от потребителей
Выбор/Согласие	Потребителям должны предоставляться варианты выбора: согласны ли они на сбор персональных данных и сведений личного характера; каким образом эта информация будет собираться; может ли она использоваться в отличных от изначально заявленных целях, и в каких именно
Доступ/Участие	Потребители должны иметь возможность просматривать данные, собранные о них, и опротестовывать их неточность или неполноту
Целостность/ Безопасность	Ответственные за сбор данных лица должны принимать все разумные меры для защиты собранной о потребителях информации от искажений и несанкционированного доступа
Надзор/Санкции	Должен использоваться надежный механизм обеспечения соблюдения и наложения санкций на нарушителей принципов честной информационной практики

Эти рекомендации разработаны на основе Руководящих принципов честной обработки информации ОЭСР, включая особо выделенные требования минимизации данных (ограничений по сбору пределами разумных потребностей), ограничения срока их хранения, обеспечения точности и разумной защищенности данных о потребителях.

Другие важные аспекты честной информационной практики включают:

- обеспечение простоты выбора для потребителей с целью снижения нагрузки на них;
- обеспечение комплексного управления данными на протяжении всего их жизненного цикла;
- наличие опции «не отслеживать действия» (do not track);
- требование наличия явно и добровольно выраженного согласия на сбор и обработку данных;
- опасения относительно функциональных возможностей сбора данных, предоставляемых поставщиками крупномасштабных платформ; гарантию прозрачности и четких уведомлений и правил, связанных с обеспечением конфиденциальности;
- обеспечение доступа физических лиц к собранным о них данным;
- обучение потребителей практическим навыкам обеспечения конфиденциальности;
- «проектируемую конфиденциальность».

Глобальная тенденция такова, что повсеместно усиливается законодательное регулирование защиты персональных данных, а за основу модели берутся, в целом, стандарты ЕС[1]. Что касается движения данных через международные границы, то тут законы и ограничения сильно варьируются в зависимости от национальной юрисдикции. Даже внутри транснациональных организаций возникают сложности с глобальным информационным обменом, обусловленные правовыми ограничениями отдельных национальных законодательств. Следовательно, организациям важно учитывать в своих внутренних политиках и правилах, руководствах и инструкциях все применимые в законном порядке требования и ограничения, чтобы сотрудники могли безбоязненно оперировать имеющимися данными в пределах той степени юридического риска, который организация считает приемлемым.

3.3 Этические аспекты работы с данными в режиме онлайн

В наши дни в США десятками появляются инициативы и программы, призванные упорядочить этический кодекс поведения в режиме онлайн (Davis, 2012). Темы горячих дискуссий включают следующее.

- **Владение данными.** Как обеспечить защиту права распоряжаться собственными персональными данными в социальных сетях? Как уберечься от брокеров данных? Ведь присосавшиеся к информационным потокам агрегаторы персональных данных научились упаковывать их в столь глубоко запрятанные профили, о которых человек даже не подозревает.
- **Право на забвение.** Имеет ли человек право на удаление всякой информации о себе из сети интернет, в частности с целью исправления подпорченной репутации? Эта тема является частью более широких дискуссий о практиках сохранения данных на определенный срок.
- **Идентификация личности.** Вправе ли мы требовать однозначной и достоверной идентификации личности пользователей, блогеров, владельцев аккаунтов и т. п.? Или же они имеют право на сохранение анонимности?
- **Свобода слова.** Где именно пролегает граница между высказыванием собственного мнения и травлей, запугиванием, возбуждением розни по всяческим групповым признакам, «троллингом» или оскорблениями.

3.4 Риски, обусловленные неэтичными практиками обращения с данными

Большинство людей, работающих с данными, прекрасно понимают, что данные вполне можно использовать для формирования неверных представлений. Успевшая стать классической книга Дарелла Хаффа «*Как лгать при помощи статистики*» (1954) описывает широкий спектр приемов обмана общественности без явной подтасовки данных за счет несложных манипуляций с отображением имеющегося фактического материала. Используемые методы включают

[1] В частности, и Федеральный закон РФ «О персональных данных» № 152-ФЗ основан, в общих чертах, на тех же принципах защиты данных, унаследованных от ОЭСР, что и европейский регламент GDPR (см. гл. 2 закона «Принципы и условия обработки персональных данных»). — *Примеч. пер.*

тенденциозные выборки данных, манипуляции с масштабами шкал, отбрасывание не укладывающихся в модель точек данных и т. п. Подходы подобного рода практикуются в сфере работы с данными и в наши дни.

Одним из способов определения границ между допустимым и недопустимым в обращении с данными является выявление и изучение практик, которые, по мнению большинства людей, являются неэтичными. Этичное обращение с данными подразумевает решительное следование моральным принципам, к которым относится, в частности, стремление к достоверности. Подтверждение достоверности может включать оценку по критериям качества данных, таким как точность и актуальность. Сюда входит также оценка по критериям правдивости и прозрачности: недопустимо использовать данные для обмана или введения в заблуждение; и обязательно обеспечивать прозрачность источников данных и намерений, стоящих за тем или иным их использованием организацией. Далее описаны лишь некоторые сценарии неэтичных практик обращения с данными, идущих вразрез со сформулированными выше принципами.

3.4.1 Манипуляции с хронологией/временем

Не исключена возможность откровенной лжи посредством выборочного исключения и/или включения точек ввода/регистрации данных, используемых для формирования итоговых статистических отчетов. В частности, на фондовых рынках распространена практика публикации котировок акций «по состоянию на момент закрытия биржевых торгов», когда они, как правило, дорожают по сравнению с усредненными котировками за прошедший день. Для справки: такой подход квалифицируется как «временная фиксация рыночной конъюнктуры» и юридически поставлен вне закона.

Специалисты в области бизнес-аналитики (Business Intelligence, BI), вероятно, первыми замечают аномалии подобного рода. Сегодня они считаются крайне ценными игроками на главных площадках биржевых торгов во всем мире, поскольку помогают воссоздавать реальную картину конъюнктуры фондовых рынков, а заодно анализировать сводки и отчеты и сопоставлять их с действующими правилами и тревожными сигналами систем мониторинга биржевой активности. Сотрудники подразделений BI, руководствующиеся этическими принципами, должны передавать сведения о выявленных аномалиях функциональным подразделениям, отвечающим за руководство и управление данными.

3.4.2 Вводящие в заблуждение визуальные представления

Диаграммы и графики также могут использоваться для формирования обманчивых представлений о данных. К примеру, играя масштабом шкалы, можно представить тенденцию более выигрышной или неприглядной, чем она есть на самом деле. Выборочная отбраковка точек данных, сравнение по общему параметру несопоставимых по природе явлений, нарушение общепринятых правил графического представления (например, сумма долей, отображенных в секторной диаграмме, должна равняться строго 100% или единице, однако на практике «пироги» нарезают с использованием произвольных цифр) и ряд других приемов также используют для введения

людей в заблуждение, побуждая интерпретировать не данные как таковые, а картинки, имеющие к этим данным весьма косвенное отношение[1].

3.4.3 Нечеткие определения или некорректные сравнения

Новостные каналы сообщали, ссылаясь на данные Бюро переписи населения США за 2011 год, что в стране на 101,7 млн человек работающего населения приходится 108,6 млн получателей социальных пособий, то есть налицо чудовищная диспропорция по этому показателю[2]. Объяснение было найдено Media Matters[3]: цифра в 108,6 млн «получателей социальных пособий» взята из отчета Бюро переписи населения об... «участниках программ помощи нуждающимся», к которым отнесены «все без исключения лица, проживающие в домохозяйствах, где хотя бы одно лицо получало социальные льготы или пособия» в IV квартале 2011 года, то есть в число «получателей пособий» оказались массово включены лица, не получившие от правительства ни цента. С другой же стороны, к «работающему населению» отнесли лишь тех, кто имеет постоянную работу, то есть «трудоустроен на полную ставку», тогда как их иждивенцы, проживающие в одном домохозяйстве с работающими, к числу «получателей зарплат» отнесены не были[4].

А как было бы этично? Этично представлять информацию с контекстным разъяснением ее смыслового значения. В рассматриваемом примере этично было бы дать четкое, однозначное и корректное определение методики измерения числа «получателей пособий». Вынося же за рамки обязательный для понимания контекст, на поверхности презентации можно получить смысл, никоим образом не поддерживаемый исходными данными. Делается ли это с целью обмануть аудиторию или по невнимательности — вопрос второй, поскольку в любом случае налицо неэтичное использование данных.

Ну и, конечно же, самое последнее дело с точки зрения этики — злоупотребление статистикой и/или ее нецелевое использование.

«Статистическое сглаживание» показателей за отчетный период способно кардинально изменить восприятие чисел. Недавно появившийся термин «data mining snooping» (в буквальном переводе — «добыча данных с отслеживанием», однако в русскоязычных источниках чаще всего используется термин «слепое прочесывание данных») описывает новомодную тенденцию

[1] См.: How To Statistics (Website). *Misleading Graphs: Real Life Examples*. 24 January 2014, http://bit.ly/1jRLgRH; io9 (Website). *The Most Useless and Misleading Infographics on the Internet*, http://bit.ly/1YDgURl; http://bit.ly/2tNktve Google «misleading data visualization». Контрпримеры этичного визуального отображения данных см.: Tufte (2001).

[2] Общая численность населения США по последним данным Бюро оценивается в 327,2 млн чел. (2018), работающего населения — в 126,8 млн чел. (2016), уровень бедности — в 12,3% (2017), а вот показатель числа получателей социальных пособий в национальной демографической статистике отсутствует как таковой (http://bit.ly/2iMlP58). — *Примеч. пер.*

[3] Media Matters for America — основанный в 2004 г. и действующий на правах некоммерческой организации (IRC 501(c)(3)) «Центр прогрессивных исследований и комплексного мониторинга, анализа и исправления консервативной дезинформации, публикуемой СМИ США» (https://www.mediamatters.org/). — *Примеч. пер.*

[4] http://mm4a.org/2spKToU. Этот же пример заодно наглядно демонстрирует и один из приемов передергивания фактов на графиках: столбец с 108,6 млн получателей пособий визуально выглядит раз в пять выше столбца с 101,7 млн работающих из-за отсечки оси значений на уровне 100 млн чел.

в статистико-аналитических исследованиях больших массивов неупорядоченных данных. В рамках этого подхода на массив данных накладываются исчерпывающие корреляционные связи, то есть данные принудительно втискиваются в рамки некой статистической модели, после чего из массива вытягивается выборка, дающая формально «статистически значимые» результаты, которые в реальности являются чисто случайными и не выходят за пределы статистической ошибки в рамках совокупности исходных данных. Неспециалисты этим приемом вводятся в заблуждение с легкостью. Этот трюк сегодня наиболее распространен в финансах и медицине (Jensen, 2000; ma.utexas.edu, 2012)[1].

3.4.4 Предвзятость, систематические ошибки и искажения

Предвзятость подразумевает искажение перспективы вследствие тенденциозности восприятия. На личностном уровне предвзятость приводит к необоснованным субъективным суждениям или предрассудкам. В статистике — к отклонению полученной оценки от реального значения оцениваемой величины вследствие накопления систематических ошибок на стадии формирования выборки данных[2]. Систематические ошибки и искажения могут привноситься на разных фазах жизненного цикла данных: при создании (сборе и/или формировании) данных; при определении источников и категорий данных и формировании выборок для статистического анализа; при выборе методов анализа; наконец, при интерпретации и представлении результатов.

Этический принцип справедливости порождает безусловное обязательство сознавать возможность накопления систематических ошибок и искажений вследствие предвзятости на стадиях сбора, обработки, анализа и/или интерпретации данных. Особенно важно помнить об этом в случае статистической обработки больших наборов данных, результаты которой могут особо негативно сказаться на группах лиц, традиционно подвергавшихся дискриминации в силу исторически сложившегося предвзятого или несправедливого отношения. Использование данных без решения проблемы их искажения может способствовать дальнейшему усугублению предубеждений и приводить к снижению прозрачности процесса, придавая конечным результатам видимость беспристрастности и нейтральности, тогда как в реальности они таковыми являться не будут. Существует несколько типов искажений и систематических ошибок, обусловленных субъективизмом или предвзятостью.

- **Сбор данных для подтверждения предопределенного результата.** Аналитикам заказали собрать доказательную базу данных в подтверждение какого-либо заключения, требующегося заказчику, который к тому же может оказывать давление на исполнителей. В результате вместо объективного сбора и анализа данных происходит подгонка выборки исходных данных под заказанный результат.
- **Предвзятое использование собранных данных.** Данные собраны объективно и с минимальными систематическими погрешностями, тем не менее на аналитика оказывается серьезное

[1] См. также многочисленные статьи на эту тему У. Эдвардса Деминга: https://deming.org/deming/deming-articles

[2] http://bit.ly/2lOzJqU

давление с целью получения подтверждения предопределенного результата. В таких ситуациях дело может дойти и до прямой подтасовки данных на стадии анализа (например, часть мешающих получить «нужный» результат данных отправляется в корзину).

- **Выборочный поиск данных в подтверждение гипотезы.** Аналитику приходит в голову догадка, которой требуется найти подтверждение, что порождает поиск данных, соответствующих гипотезе, и игнорирование тех данных, которые ей противоречат.
- **Тенденциозная методология выборки.** Если статистическое исследование предусматривает формирование выборки исходных данных, возможность для привнесения искажений порой возникает еще на этапе выбора соответствующей методологии. Нереалистично ждать от лиц, собирающих данные, полной объективности, беспристрастности и непредвзятости при составлении выборки. Для минимизации возможных искажений следует использовать статистические инструменты формирования выборок и обеспечивать адекватные размеры самих выборок. Особенно важно помнить об однозначной тенденциозности наборов данных, подготовленных в целях обучения.
- **Влияние контекста и культуры.** Искажения нередко бывают обусловлены культурой или контекстом, что требует умения отрешиться от этих факторов для нейтрального восприятия изучаемой ситуации.

Спектр вопросов, касающихся искажений, весьма широк и в каждом конкретном случае зависит от различных факторов, таких как тип обрабатываемых данных, заинтересованные лица, порядок заполнения наборов данных, удовлетворяемые исследованием бизнес-нужды и ожидаемые результаты. Однако устранить абсолютно все искажения порой невозможно, а иногда — даже нежелательно. К примеру, в бизнесе дискриминация неплатежеспособных клиентов (с которыми бессмысленно и даже вредно продолжать иметь дело) закладывается бизнес-аналитиками во многие сценарии: эту категорию потребителей безжалостно вычищают из выборок или игнорируют при анализе. В подобных случаях долг аналитиков — документировать критерии определения изучаемой демографической выборки. В качестве противоположного примера стоит привести прогностические алгоритмы выявления «потенциальных преступников» или «криминогенных районов», широко используемые полицией. Такие методы статистического анализа рисков сопряжены со значительно более серьезным посягательством на этические принципы справедливости и равноправия, поэтому использовать их следует с большей осмотрительностью, обеспечивая, в частности, прозрачность алгоритмов прогнозирования и персональную ответственность за умышленное искажение данных, используемых для обучения этих алгоритмов[1].

[1] Примеры предвзятости при машинном обучении см.: Brennan (2015) и на сайтах Ford Foundation и ProPublica (http://bit.ly/2oYmNRu). Помимо систематических ошибок, обусловленных тенденциозностью выборок данных, использовавшихся для обучения, серьезную проблему представляет и непрозрачность предиктивных алгоритмов, используемых самообучаемыми машинами: по мере их усложнения всё труднее отслеживать логику и происхождение принимаемых ими решений. См.: Lewis and Monett (2017).

3.4.5 Преобразование и интеграция данных

Этические трудности при интеграции данных и их преобразовании обусловлены тем, что данные перемещаются из системы в систему, а при отсутствии должного контроля это сопряжено с риском неэтичного и даже противоправного обращения с данными. Этические риски такого рода связаны и с фундаментальными проблемами управления данными; в их числе:

- **Ограниченность знаний о первоисточнике и происхождении данных.** Если организация не знает, откуда взялись данные, и не располагает сведениями об их происхождении (lineage), то есть сведениями о том, как они изменялись при перемещении из системы в систему, у нее нет доказательств их соответствия действительности.
- **Некачественные данные.** У организаций должны иметься четкие и измеримые стандарты качества данных, и все используемые данные должны проверяться на предмет соответствия этим стандартам. Без подтверждения качества данных организация не вправе ручаться за их достоверность, а те, кто использует эти данные в практических целях, рискуют сами или подвергают риску других.
- **Ненадежные метаданные.** Потребители данных зависят от надежности и качества метаданных, включая однозначные и непротиворечивые определения отдельных элементов данных, задокументированные сведения о первоисточнике и происхождении данных (например, правила, использовавшиеся при интеграции данных). Без надежных метаданных данные могут быть неверно поняты или неправильно использованы. В случаях обмена данными между организациями, а тем более их трансграничного движения, метаданные должны обязательно включать признаки, указывающие на их источник, владельца и, при необходимости, особые требования к защите.
- **Отсутствие документированной истории исправления данных.** У организаций также должна иметься доступная для проверки информация от том, как именно и когда изменялись любые данные, имеющиеся в ее распоряжении. Даже если исправления вносятся с целью устранения неточностей, повышения качества или обеспечения лучшей защиты данных, они тем не менее могут оказаться несанкционированными. Поэтому любые исправления должны осуществляться строго в соответствии с формальным контролируемым процессом управления изменениями.

3.4.6 Обфускация / Редактирование данных

Под обфускацией (от *лат.* obfuscare — затенять, затемнять; и *англ.* obfuscate — делать неочевидным, запутанным, сбивать с толку) данных понимается практика преобразования исходной информации в анонимную или удаления из нее важных элементов (иногда используется также термин «редактирование данных» — data redaction). Однако для защиты данных одной лишь обфускации может оказаться недостаточно, если последующая обработка (анализ или сопоставление с другими наборами данных) может привести к их раскрытию. Этот риск не устраняется полностью ни одним из нижеперечисленных методов.

- **Агрегирование данных.** При агрегировании данных по некоторому набору измерений и одновременном удалении идентификационных данных полученный набор данных можно по-прежнему использовать в аналитических целях без опасения раскрытия персональной идентификационной информации. Распространенной практикой является агрегирование данных по географическому признаку (см. главы 7 и 14).
- **Маркировка данных.** Маркировка данных применяется для их классификации по степени важности с точки зрения защиты (секретная, конфиденциальная, персональная и т. п.) и для контроля их передачи различным группам пользователей — широкой публике, поставщикам или партнерам, — в том числе пользователям из определенных стран, а также другим группам, отобранным по самым разным признакам.
- **Маскировка данных.** Маскировка данных используется для разблокирования процессов, в которых участвуют операторы, вводящие данные. Процесс разблокируется только при соответствии введенных данных заданному формату и/или критериям. Операторы не видят, какие данные считаются приемлемыми, а какие нет, а просто вводят с клавиатуры предоставляемые ответы; если ответ правильный, осуществляется переход к дальнейшим действиям. Бизнес-процессы с маскировкой данных используются, в частности, в деятельности аутсорсинговых колл-центров, а также при работе с субподрядчиками, которым предоставляется ограниченный доступ к информации.

Практика использования экстремально больших массивов данных в аналитической деятельности, относящейся к науке о данных (data science), способствует резкому возрастанию уже не только теоретического, но и практического интереса к вопросам эффективности анонимизации. Обладая огромными массивами данных из разных источников и как следует углубившись в них, вполне возможно методами сравнительного анализа и сопоставления данных выявлять и идентифицировать конкретных людей, даже если исходные массивы прошли анонимизацию. В связи с этим первоочередной задачей при попадании информации в «озеро данных» (data lake) должен являться ее анализ на присутствие важных с точки зрения информационной безопасности данных и незамедлительное применение общепринятых методов их защиты. Но одного этого может оказаться недостаточно для полной гарантии невозможности утечки; именно потому ключевыми факторами для организации являются последовательное руководство и приверженность принципам этического обращения с данными (см. главу 14).

3.5 Формирование культуры этичного обращения с данными

Для того чтобы в организации сложилась культура этичного обращения с данными, необходимо понимание существующих практик и определение требований к ожидаемому поведению с последующей формализацией их в виде политик, правил и кодекса поведения, а также проведение обучения персонала и обеспечение надзора за соблюдением новых норм обращения с данными. Как и в случае с любыми другими инициативами, относящимися к руководству данными и изменению корпоративной культуры, тут требуется эффективное лидерство.

Требования этичного обращения с данными безусловно включают соблюдение норм действующего законодательства, но, помимо этого, они должны распространяться и на процедуры сбора, анализа и интерпретации данных, и на вопросы, связанные с повышением эффективности их использования как внутри, так и вне организации. В организационной культуре, где ценится этичное поведение, будут предусмотрены не только кодексы поведения, но и проработанные каналы коммуникации, и действенные рычаги принятия необходимых мер, которые помогают сотрудникам чувствовать, что их запросы к высшему руководству по поводу кажущихся им неэтичными практик или потенциальных этических рисков будут внимательно выслушаны, а необходимые меры по пресечению нарушений не будут носить карательного характера в адрес тех, кто сообщил о нарушениях. В целом же для привития культуры этичного обращения с данными не обойтись без формального процесса управления организационными изменениями (см. главу 17).

3.5.1 Обзорный анализ текущего состояния практик обращения с данными

Первый шаг к улучшению ситуации — понять текущее состояние. Обзорный анализ существующих практик обращения с данными необходим, чтобы определить, насколько прямо и точно они привязаны к этическим и юридическим нормам. Такой анализ также позволяет выявить, хорошо ли сотрудники понимают этические составляющие существующих практик и их вклад в выстраивание и поддержание доверительных отношений с клиентами, партнерами и другими заинтересованными сторонами. На основании результатов обзорного анализа можно сформулировать этические принципы, которые должны лечь в основу сбора, использования и надзора за данными на протяжении всего их жизненного цикла, включая деятельность по обеспечению коллективного использования данных.

3.5.2 Выявление и определение принципов, практик и факторов риска

Формальное закрепление принципов этичной практики обращения с данными призвано снизить риск их неправильного или злонамеренного использования во вред клиентам, сотрудникам, коммерческим партнерам и другим заинтересованным лицам либо самой организации. Стремясь к совершенствованию используемых практик, организация должна иметь представление как об общих принципах, таких как необходимость обеспечения неприкосновенности частной жизни и неразглашения персональных данных, так и о специфических отраслевых нормах — например, необходимости защиты финансовой информации или данных, связанных со здоровьем.

Подход организации к обеспечению этичной работы с данными должен соответствовать требованиям законодательства и нормативно-правового регулирования. В частности, организациям, осуществляющим свою деятельность в международных масштабах, нужно иметь общее представление об этических принципах, лежащих в основе законов каждой из стран, где они работают, а также о соглашениях между этими странами. В дополнение, у большинства организаций имеются и специфические риски, которые могут быть обусловлены, к примеру, технологическим ландшафтом, текучестью кадров, средствами сбора информации о потребителях или иными факторами.

Принципы должны согласовываться с рисками возникновения ущерба, которыми чревато несоблюдение этих принципов, и утвержденными практиками, помогающими этих рисков избежать. Наконец, необходимы механизмы контроля соблюдения установленных практик. Рассмотрим весь этот комплекс на следующем примере.

- **Руководящий принцип.** Люди имеют право на неразглашение информации о состоянии их здоровья. Следовательно, доступа к историям болезни не должен иметь никто, кроме лечащих врачей и иного персонала медучреждений, которым сведения о состоянии здоровья пациентов нужны для оказания медицинской помощи.
- **Риск.** Если открыть доступ к данным о здоровье пациентов слишком широкому кругу лиц, серьезно повышается вероятность утечки и/или разглашения этой конфиденциальной информации, то есть нарушения одного из неотъемлемых прав пациента.
- **Практика.** Доступ к историям болезни пациентов имеет только медперсонал, непосредственно занимающийся оказанием медицинской помощи.
- **Контроль.** Ежегодная проверка списков пользователей информационных систем, содержащих персональные данные о состоянии здоровья пациентов, с целью убедиться, что доступ к ним имеют лишь те, кому это положено.

3.5.3 Разработка стратегии обеспечения этичного обращения с данными и дорожной карты ее реализации

Завершив анализ текущего положения дел и разработку этических принципов, организация может приступать к формализации стратегии совершенствования обращения с данными. Она должна отражать и сами принципы, и ожидаемые нормы поведения при обращении с данными, посредством положения о ценностях и кодекса этического поведения. Компоненты подобной стратегии включают следующее.

- **Положение о ценностях.** Положение о ценностях описывает принципы, в которые верит организация. Примерами могут служить правда, честность или справедливость. Это создает основу для этичного обращения с данными и принятия решений.
- **Принципы этичного обращения с данными.** Принципы этичного обращения с данными описывают подход организации к решению проблем, связанных с данными. Например, как соблюдается право на неприкосновенность частной жизни и на неразглашение персональных данных. Принципы и правила поведения могут быть сведены в этический кодекс и дополнены политикой организации по обеспечению его соблюдения. Мероприятия по внедрению кодекса и политики должны быть включены в программу обучения и план коммуникаций.
- **Рамочная структура обеспечения соответствия.** Рамочная структура обеспечения соответствия включает этические факторы, побуждающие организацию к соблюдению законодательных и нормативных требований. Этичное поведение облегчает работу по обеспечению соответствия этим требованиям. Требования могут варьироваться в зависимости от страны и сектора экономики.

- **Оценки рисков.** Оценка рисков сводится к определению вероятности возникновения конкретных проблем и тяжести их последствий для организации. Результаты оценки следует использовать для приоритизации мер по снижению рисков, включая обеспечение соблюдения сотрудниками этических принципов.
- **Обучение и коммуникации.** Обучение должно включать обзор этического кодекса. Сотрудники должны дать расписку в том, что они его изучили и знают о последствиях неэтичного обращения с данными. Обучение должно быть непрерывным. Можно предусмотреть, например, ежегодное проведение занятий с подписанием подтверждения ранее взятых обязательств по соблюдению этического кодекса. Соответствующие коммуникации должны охватывать всех сотрудников.
- **Дорожная карта.** Дорожная карта должна включать график мероприятий, согласованных с руководством. График должен предусматривать обучение, мероприятия в соответствии с планом коммуникаций, мероприятия по идентификации и устранению несоответствий между текущими практиками и целевыми представлениями, а также по минимизации рисков и мониторингу. Необходимо разработать детальные формулировки, отражающие целевое состояние организации в части этического обращения с данными. Описать роли, обязанности и процессы, а также указать ссылки на экспертов, у которых можно получить дополнительную информацию. Кроме того, дорожная карта должна учитывать все применимые к организации законы и культурные факторы.
- **Аудит и мониторинг.** Этические идеи и положения этического кодекса могут быть закреплены с помощью соответствующего обучения. Целесообразно также осуществлять мониторинг деятельности по отдельным направлениям с целью подтверждения, что она ведется в соответствии с этическими принципами.

3.5.4 Принятие социально ответственной модели этических рисков

Специалисты в области BI, аналитики и науки о данных часто отвечают за следующие данные.

- Биографические и анкетные сведения: страна и/или место рождения, расовая, этническая и религиозная принадлежность и т. п.
- Род занятий и виды деятельности, включая политическую, социальную, а также подозрения на причастность к преступной деятельности.
- Информация об образе жизни: место жительства, уровень доходов, характер покупок, круг общения и/или переписки.
- Социальный статус, включая результаты анализа: например, число баллов или резюме предпочтений, по которым людей маркируют как «перспективных» или «бесперспективных» клиентов.

Подобными данными легко злоупотребить вопреки основополагающим принципам этики, предписывающим уважительное, благонамеренное и справедливое отношение к личности.

Деятельность в области BI, аналитики и науки о данных справедливо требует распространения сферы действия принципов этичности далеко за рамки организации, на которую работают

специалисты в этих областях; по сути, они несут ответственность за этичное обращение с данными перед обществом в целом. Этическую перспективу необходимо учитывать не только по причине возможности злоупотребления данными, но еще и потому, что организации несут социальную ответственность перед всеми людьми, данными о которых располагают, за предотвращение использования этих данных им во вред.

Например, организация может установить критерии отнесения людей к категории «плохих» клиентов с целью прекратить сотрудничество с этими потребителями. Однако если организация при этом является монополистом в сфере предоставления какой-либо жизненно важной услуги в конкретном регионе, то, отказав таким лицам в обслуживании, она тем самым причинит им вред.

Проекты, в рамках которых используются персональные данные, требуют соблюдения строгой дисциплины обращения с ними (см. рис. 13). В таких проектах должна быть предусмотрена ответственность за соблюдение этических норм в следующих областях:

Рисунок 13. Модель этических рисков для проектов выборочного обследования

- методология выборки изучаемых групп населения (стрелка 1);
- порядок сбора данных (стрелка 2);
- точное определение предмета анализа (стрелка 3);
- порядок доступа к результатам (стрелка 4).

В каждой из указанных областей следует тщательно изучать и минимизировать потенциальные этические риски, особое внимание уделяя возможным негативным последствиям для потребителей или граждан.

Модель рисков можно использовать, во-первых, для определения допустимости реализации проекта с учетом этических соображений, а во-вторых, для выработки методологии реализации проекта. Например, могут быть предусмотрены такие меры, как анонимизация данных, удаление персональных данных из файлов с данными, усиленная защита файлов с данными и/или строгий контроль доступа к ним, юридическое заключение о допустимости проекта в соответствии с действующими законами о защите персональных данных и неразглашении конфиденциальной информации о частных лицах. Отказ в обслуживании клиентов может не допускаться действующим законодательством, если организация является монопольным поставщиком на местном рынке (например, теплоэнергетических ресурсов, воды и т. п.).

Поскольку проекты в области анализа данных устроены весьма сложно, люди могут элементарно не видеть связанных с ними этических проблем. Организациям нужно активнее выявлять потенциальные риски самостоятельно, а также ценить и защищать активных граждан, бдительно выискивающих риски и бьющих тревогу по их поводу. Автоматизированного мониторинга для защиты от неэтичных действий и поведения недостаточно. Людям следует анализировать происходящее и почаще задумываться о том, не становятся ли они свидетелями предвзятости или дискриминации. Культурные и этические нормы поведения на рабочих местах оказывают значительное влияние на корпоративное поведение, поэтому следует тщательно изучить и использовать вышеописанную модель выявления и оценки этических рисков. DAMA International призывает профессионалов в области работы с данными занимать жесткую профессиональную позицию в отношении любых проявлений и рисков неэтичного обращения с данными и в обязательном порядке докладывать о неблагополучных ситуациях высшему руководству, которое может не вполне отдавать себе отчет о возможных пагубных последствиях использования тех или иных данных с неочевидными для них нарушениями этических норм.

3.6 Этика обращения с данными и руководство данными

Общая ответственность и надзор за соблюдением принципов и правил этичного обращения с данными возложены как на должностных лиц, отвечающих за руководство данными, так и на юридическую службу организации. Обе стороны обязаны совместно отслеживать последние изменения законодательства и минимизировать риск этически недопустимых событий за счет обеспечения полнейшего понимания работниками своих обязанностей по соблюдению норм этики обращения с данными. Со стороны руководства данными должны быть установлены стандарты и политики обращения с данными и предусмотрены механизмы надзора за их соблюдением на практике. Сотрудники вправе рассчитывать на честное обращение с их собственными персональными данными, защиту от преследований за сообщения о возможных нарушениях и невмешательство в их личную жизнь. В функции руководства данными безусловно должен входить

надзор за разработкой, проверкой и утверждением планов и решений, предлагаемых специалистами из сфер BI, аналитики и науки о данных.

Для того чтобы получить сертификат профессионала в области управления данными (Certified Data Management Professional, CDMP), выдаваемый DAMA International, специалист должен не только сдать квалификационные экзамены, но и присоединиться к формализованному этическому кодексу, включающему обязательство обращаться с данными этично и во благо всего общества, не ограничиваясь интересами организации-работодателя.

4. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Blann, Andrew. *Data Handling and Analysis*. Oxford University Press, 2015. Print. Fundamentals of Biomedical Science.

Council for Big Data, Ethics, and Society (website), http://bit.ly/2sYAGAq

Davis, Kord. *Ethics of Big Data: Balancing Risk and Innovation*. O'Reilly Media, 2012. Print.

European Data Protection Supervisor (EDPS). Opinion 4/2015 «Towards a new digital ethics: Data, dignity and technology», http://bit.ly/2sTFVlI

Federal Trade Commission, US (FTC). *Federal Trade Commission Report Protecting Consumer Privacy in an Era of Rapid Change*. March 2012, http://bit.ly/2rVgTxQ и http://bit.ly/1SHOpRB

GDPR REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

Hasselbalch, Gry and Pernille Tranberg. *Data Ethics: The New Competitive Advantage*. Publishare, 2016.

Huff, Darrell. *How to Lie with Statistics*. Norton, 1954. Print. [*Дарелл Хафф*. Как лгать при помощи статистики. — М.: Альпина Паблишер, 2015.]

Jensen, David. «Data Snooping, Dredging and Fishing: The Dark Side of Data Mining A SIGKDD99 Panel Report». *SIGKDD Explorations*. ACM SIGKDD, Vol. 1, Issue 2. January 2000.

Johnson, Deborah G. *Computer Ethics*. 4th ed. Pearson, 2009. Print.

Kaunert, C. and S. Leonard, eds. *European Security, Terrorism and Intelligence: Tackling New Security Challenges in Europe*. Palgrave Macmillan, 2013. Print. Palgrave Studies in European Union Politics.

Kim, Jae Kwan and Jun Shao. *Statistical Methods for Handling Incomplete Data*. Chapman and Hall/CRC, 2013. Chapman and Hall/CRC Texts in Statistical Science.

Lake, Peter. *A Guide to Handling Data Using Hadoop: An exploration of Hadoop, Hive, Pig, Sqoop and Flume*. Peter Lake, 2015.

Lewis, Colin and Dagmar Monett. *AI and Machine Learning Black Boxes: The Need for Transparency and Accountability*. KD Nuggets (website), April 2017, http://bit.ly/2q3jXLr

Lipschultz, Jeremy Harris. *Social Media Communication: Concepts, Practices, Data, Law and Ethics*. Routledge, 2014. Print.

Mayfield, M. I. *On Handling the Data*. CreateSpace Independent Publishing Platform, 2015. Print.

Mazurczyk, Wojciech et al. *Information Hiding in Communication Networks: Fundamentals, Mechanisms, and Applications*. Wiley-IEEE Press, 2016. Print. IEEE Press Series on Information and Communication Networks Security.

Naes, T. and E. Risvik eds. *Multivariate Analysis of Data in Sensory Science*. Volume 16. Elsevier Science, 1996. Print. Data Handling in Science and Technology (Book 16).

Olivieri, Alejandro C. et al., eds. *Fundamentals and Analytical Applications of Multi-way Calibration*. Volume 29. Elsevier, 2015. Print. Data Handling in Science and Technology (Book 29).

ProPublica (website). «Machine Bias: Algorithmic injustice and the formulas that increasingly influence our lives». May 2016, http://bit.ly/2oYmNRu

Provost, Foster and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, 2013. Print.

Quinn, Michael J. *Ethics for the Information Age*. 6th ed. Pearson, 2014. Print.

Richards, Lyn. *Handling Qualitative Data: A Practical Guide.* 3 Pap/Psc ed. SAGE Publications Ltd, 2014. Print.

Thomas, Liisa M. *Thomas on Data Breach: A Practical Guide to Handling Data Breach Notifications Worldwide*. LegalWorks, 2015. Print.

Tufte, Edward R. *The Visual Display of Quantitative Information*. 2nd ed. Graphics Pr., 2001. Print.

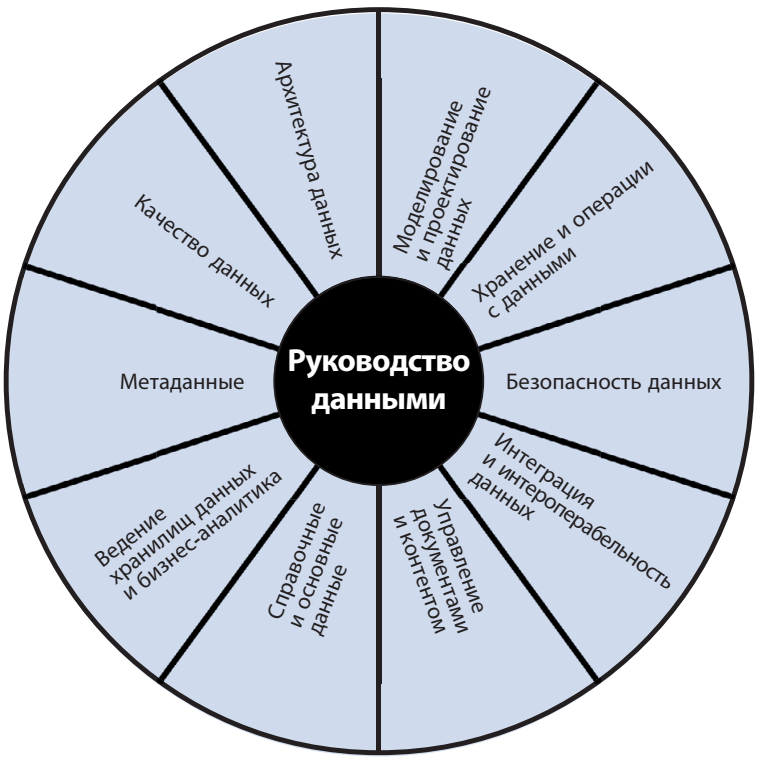
University of Texas at Austin, Department of Mathematics (website). *Common Misteaks Mistakes in Using Statistics*.

US Department of Health and Human Services. *The Belmont Report*. 1979, http://bit.ly/2tNjb3u (US-HSS, 1979).

US Department of Homeland Security. «Applying Principles to Information and Communication Technology Research: A Companion to the Department of Homeland Security Menlo Report». January 3, 2012, http://bit.ly/2rV2mSR (US-DHS, 2012).

Witten, Ian H., Eibe Frank and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. Morgan Kaufmann, 2011. Print. Morgan Kaufmann Series in Data Management Systems.

Руководство данными



DAMA-DMBOK2 Data Management Framework

1. ВВЕДЕНИЕ

Руководство данными (Data Governance, DG) определяется как деятельность по осуществлению руководящих и контрольных полномочий (планирование, мониторинг и обеспечение выполнения) в отношении управления информационными активами. Решения относительно данных принимаются в любой организации вне зависимости от того, введена ли в ней формально определенная функция руководства данными. В организациях, где действует формальная программа руководства данными, реализация полномочий по руководству и контролю носит более осознанный и целенаправленный характер (Seiner, 2014). Такие организации обладают большими возможностями по повышению ценности, извлекаемой из своих информационных активов.

Функция руководства данными выступает в качестве руководящей всеми остальными функциями управления данными. Целью руководства является обеспечение надлежащего управления данными в соответствии с политиками и лучшими практиками (Ladley, 2012). В то время как главным драйвером управления данными в целом является обеспечение извлечения ценности из информационных активов, функция руководства данными сосредоточена на том, каким образом принимаются решения, касающиеся данных, и как должны функционировать люди и процессы, имеющие к ним отношение. Общее содержание и ключевые направления программы руководства данными будут зависеть от потребностей конкретной организации, но большинство программ включает следующие компоненты.

- **Стратегия.** Определение, доведение до исполнителей и управление реализацией Стратегии работы с данными и Стратегии руководства данными.
- **Политика.** Определение и обеспечение соблюдения политик в отношении управления, доступа, использования, безопасности и качества данных и метаданных.
- **Стандарты и качество.** Определение и обеспечение соблюдения стандартов в области качества данных и архитектуры данных.
- **Надзор.** Практическое осуществление наблюдения, аудита и исправления выявленных недостатков в ключевых аспектах обеспечения качества, соблюдения политики и управления данными (часто такая деятельность называется *распоряжением — stewardship*).
- **Нормативно-правовое соответствие.** Обеспечение соблюдения организацией нормативно-правовых требований в отношении данных.
- **Управление проблемными вопросами.** Выявление, определение, эскалация и разрешение проблемных вопросов, связанных с безопасностью данных, доступом к данным, качеством данных, нормативно-правовым соответствием, владением данными, политикой, стандартами, терминологией и процедурами руководства данными.
- **Проекты по управлению данными.** Поддержка усилий, направленных на развитие практик управления данными.
- **Оценка информационных активов.** Введение стандартов и процессов, позволяющих согласованно определять ценность информационных активов для бизнеса.

Для достижения целей в указанных направлениях деятельности по руководству данными программа руководства должна предусматривать разработку политик и процедур, развитие практики распоряжения данными на различных уровнях внутри организации и инициацию мероприятий по управлению изменениями, направленных на активное разъяснение сотрудникам организации выгод от внедрения руководства данными и моделей деятельности, необходимых для успешного управления данными как активом.

Для большинства организаций формальное внедрение руководства данными требует поддержки со стороны сферы управления изменениями (см. главу 17) и одобрения кого-то из руководителей высшего уровня (C-level): например, директора по рискам (Chief Risk Officer, CRO),

Рисунок 14. Контекстная диаграмма: руководство и распоряжение данными

финансового директора (Chief Financial Officer, CFO) или директора по данным (Chief Data Officer, CDO).

Способность создавать и распространять данные и информацию в корне преобразила наши личные и экономические взаимодействия. Динамичность рынков и возросшее понимание высокой значимости данных в качестве дифференцирующего фактора в конкурентной борьбе заставляют организации перестраивать распределение полномочий и ответственности в сфере управления данными. Эта тенденция отчетливо прослеживается в финансовом, государственном секторах, в области электронной коммерции и розничной торговле. Организации прилагают всё больше усилий, чтобы стать управляемыми на основе данных (data-driven), используя проактивный подход и рассматривая определение требований к данным как составную часть разработки стратегии, программного планирования и внедрения технологий. Однако это зачастую влечет за собой серьезные культурные вызовы. Более того, поскольку недостаточно развитая организационная культура способна разрушить любую стратегию, усилия по внедрению руководства данными должны обязательно включать культурные изменения, а для этого, в свою очередь, требуется эффективное лидерство.

Чтобы организация могла получать реальную выгоду от данных как корпоративного актива, в ее организационную *культуру* должны быть привнесены практики оценки данных и управления данными. Даже с самой лучшей стратегией работы с данными планы по руководству и управлению ими не будут успешно реализованы, пока организация не признает необходимость проведения изменений и не начнет ими управлять. Для многих организаций именно изменение культуры становится главным вызовом. Один из основных принципов управления изменениями заключается в необходимости перемен на индивидуальном уровне (Hiatt and Creasey, 2012). Когда руководство и управление данными требуют серьезных изменений в поведении людей, без формального управления изменениями успеха не добиться.

1.1 Бизнес-драйверы

Наиболее распространенным драйвером внедрения руководства данными является необходимость соблюдения организацией нормативно-правовых требований, особенно в строго регулируемых отраслях, таких как финансово-банковский сектор и здравоохранение. Обеспечение соответствия постоянно развивающемуся законодательству требует внедрения строго регламентированных процессов руководства данными. Дополнительную движущую силу создает взрывное развитие расширенной аналитики (advanced analytics) и науки о данных (data science).

Наряду с требованиями нормативно-правового соответствия и аналитики, стимулирующим фактором внедрения руководства данными для многих организаций выступает выполнение собственной программы управления информацией. Программа может быть ориентирована на конкретные потребности бизнеса (например, управление основными данными), или на решение крупных проблем в области данных, или одновременно на то и другое. Типичный сценарий: компания нуждается в более качественных и детальных данных о клиентах, в связи с чем принимает решение о разработке системы управления основными данными (master data management, MDM)

о клиентах, а затем приходит к пониманию того, что успешное управление основными данными требует также и руководства данными.

Руководство данными — не самоцель. Оно должно быть согласовано непосредственно со стратегией развития организации. Чем более очевидна его помощь в решении проблем организации, тем более вероятно, что сотрудники изменят поведение и примут практики руководства данными. Драйверы внедрения руководства данными чаще всего включают различные аспекты снижения рисков или совершенствования процессов.

- Снижение рисков.
  - **Общее управление рисками.** Учет, анализ и минимизация финансовых и репутационных рисков, обусловленных данными, включая принятие мер по проблемным правовым (электронное раскрытие информации — E-Discovery) и нормативным вопросам.
  - **Безопасность данных.** Защита информационных активов посредством механизмов контроля доступа, возможности использования, целостности, согласованности, возможности проверки и безопасности данных.
  - **Конфиденциальность.** Контроль частных/конфиденциальных/персональных данных посредством политики и мониторинга нормативно-правового соответствия.

- Совершенствование процессов.
  - **Нормативно-правовое соответствие.** Способность эффективно и последовательно обеспечивать соответствие нормативно-правовым требованиям.
  - **Повышение качества данных.** Способность обеспечивать повышение эффективности бизнеса за счет повышения уровня надежности и достоверности данных.
  - **Управление метаданными.** Создание бизнес-глоссария, позволяющего определять и оперативно отыскивать необходимые данные в информационных системах организации; обеспечение ведения в организации широкого спектра других метаданных и предоставление к ним доступа.
  - **Эффективность управления проектами разработки.** Совершенствование управления жизненным циклом разработки систем (system development lifecycle, SDLC) с целью устранения проблем и реализации возможностей в области управления данными организации, включая управление устранением относящихся к данным технических недоработок посредством руководства жизненным циклом данных.
  - **Управление поставщиками.** Контроль исполнения контрактов, связанных с данными: например, на внедрение облачных хранилищ, приобретение данных, продажу данных как продукта, аутсорсинг обработки данных.

Важно ясно представлять отдельные бизнес-драйверы руководства данными в организации и согласовывать их с общей стратегией развития бизнеса. Слова «*организация руководства*

данными» (*DG organization*[1]) часто вызывают у руководителей высшего уровня отторжение, поскольку всё с ними связанное воспринимается как нечто сопряженное с дополнительными расходами безо всяких явных выгод. Поэтому важно с учетом сложившийся организационной культуры определить правильный язык для доведения этой концепции до руководства, а также выбрать операционную модель и роли в рамках программы ее реализации. Отметим, что по состоянию на время написания DMBOK2 термин *организация* (*organization*) при употреблении в смысле *организационной системы* активно замещается такими более общими терминами, как *операционная модель* (*operating model*) или *операционная рамочная структура* (*operating framework*)[2].

Хотя люди иногда утверждают, будто им трудно понять, что такое «руководство данными», понятие «*руководство*» как таковое является общепринятым и устоявшимся. Вместо того чтобы изобретать новые подходы, специалисты в области управления данными могут применить к руководству данными идеи и принципы руководства, уже применяемые в других областях. Достаточно часто проводятся аналогии между руководством данными и бухгалтерским учетом и аудитом. Аудиторы и контролеры устанавливают правила управления финансовыми активами. Специалисты в области руководства данными устанавливают правила управления активами информационными. Представители других направлений деятельности должны эти правила выполнять.

Руководство данными — не разовое мероприятие. Для его осуществления требуется постоянно действующая программа, нацеленная на обеспечение извлечения максимальной выгоды из имеющихся в ее распоряжении данных при одновременной минимизации рисков, связанных с этими данными. Команда руководства данными может быть как виртуальной, так и формально определенной организационной системой со специальными видами подотчетности. Для эффективной деятельности в рамках руководства данными требуется четкое понимание ролей и проводимых работ. Они должны быть встроены в операционную структуру с хорошо отлаженными и реализованными внутри организации функциями. Программа руководства данными должна учитывать характерные организационные и культурные аспекты, а также специфические для управления данными вызовы и возможности внутри организации (см. главы 1 и 16).

Руководство данными отделено от руководства ИТ (IT governance). В рамках руководства ИТ принимаются решения об инвестициях в информационные технологии, о портфеле ИТ-приложений и ИТ-проектов, иными словами — об аппаратном и программном обеспечении, а также об общей технической архитектуре. Руководство ИТ обеспечивает соответствие ИТ-стратегий и инвестиций целям и стратегиям организации. Стандартные подходы к руководству ИТ

[1] В данном издании применительно к понятиям «data governance organization» и «data management organization» слово «organization» переведено не как «организация», а как «организационная система» (здесь под организационной системой понимается совокупность организационной структуры и организационного механизма). Это сделано для того, чтобы избежать путаницы при использовании в одних и тех же подразделах слова «organization» как в смысле организационной системы, так и в смысле предприятия, учреждения, компании и т. п. — *Примеч. науч. ред.*

[2] В DMBOK2 термины «operating model» и «operating framework» используются как синонимы. — *Примеч. науч. ред.*

определяет методология COBIT[1], но лишь небольшая ее часть посвящена управлению данными и информацией. Некоторые из важных и широко обсуждаемых тем, например обеспечение соблюдения требований закона Сарбейнса — Оксли (США), в силу своей широты затрагивают сферы и корпоративного руководства, и руководства ИТ, и руководства данными. Но само по себе руководство данными, напротив, сфокусировано исключительно на управлении информационными активами и данными как активом.

1.2 Цели и принципы

Цель руководства данными (DG[2]) — создать в организации возможности для управления данными как активом. DG предоставляет принципы, политику, процессы, рамочную структуру, метрики и механизмы надзора для управления данными как активом и для руководства деятельностью по управлению данными на всех уровнях. Для достижения этой всеобъемлющей цели программа DG должна быть:

- **Устойчивой.** Программу DG нужно сделать «текущей» и «неуклонной». DG — не разовый проект с определенным сроком завершения: это непрерывный процесс, требующий приверженности всей организации. Программа DG требует внесения изменений в порядок управления данными и их использования. Для этого далеко не всегда нужно нагромождать новые организационные структуры. Но управление изменениями должно осуществляться таким образом, чтобы была обеспечена устойчивость полученных результатов после первоначального внедрения любого из компонентов DG. Устойчивое руководство данными, кроме того, зависит от руководителей высшего звена (leadership), оказываемой им поддержки (sponsorship) и четкого регулирования вопросов владения (ownership), которое подразумевает наделение конкретных сотрудников полномочиями на утверждение решений, касающихся тех или иных информационных ресурсов или областей данных.
- **Встроенной в процессы.** DG — не дополнительный процесс; работы, проводимые в рамках DG, должны быть включены в методики разработки программного обеспечения, процессы анализа данных, управления основными данными, управления рисками и другие процессы, предусматривающие работу с данными.
- **Измеримой.** Хорошо организованное DG приносит финансовую отдачу, но, чтобы ее продемонстрировать, необходимы понимание исходной точки и планирование измеримых улучшений.

[1] COBIT (*сокр.* от *англ.* Control Objectives for Information and Related Technology — Цели контроля для информационных и смежных технологий) — признанная международным сообществом методология управления ИТ. Разработана и поддерживается ISACA (ранее полностью — Information Systems Audit and Control Association) — международной ассоциацией профессионалов в области управления ИТ. — *Примеч. пер.*

[2] В данном издании в качестве сокращения термина «руководство данными» используется устоявшаяся аббревиатура DG (*сокр.* от *англ.* Data Governance). Аналогичные соглашения используются в отношении других областей знаний управления данными и относящихся к ним ключевых терминов (за исключением тех случаев, когда существует устоявшееся русское сокращение). — *Примеч. науч. ред.*

Реализация программы DG требует приверженности изменениям. Ниже сформулированы разрабатывавшиеся с начала нулевых годов принципы, способные заложить прочный фундамент для руководства данными[1].

- **Лидерство и стратегия.** Успешное руководство данными начинается с проявляемых руководителями дальновидности, лидерства и приверженности. Работы по управлению данными проводятся в соответствии со стратегией работы с данными, которая, в свою очередь, основывается на корпоративной стратегии развития бизнеса.
- **Определяющее влияние со стороны бизнеса[2].** Руководство данными является бизнес-программой, и в силу этого оно должно оказывать определяющее влияние на решения в области ИТ, связанные с данными, точно так же, как оно оказывает определяющее влияние на взаимодействие между бизнесом и данными.
- **Разделенная ответственность.** Во всех областях знаний по управлению данными руководство данными входит в сферу совместной ответственности распорядителей данных со стороны бизнеса и технических специалистов по управлению данными.
- **Многоуровневость.** Руководство данными осуществляется как на корпоративном уровне, так и на локальном, а часто и на промежуточных.
- **Базовая рамочная структура.** Поскольку все относящиеся к DG работы требуют координации между функциональными областями, программа DG должна установить операционную рамочную структуру, определяющую функциональные обязанности и взаимодействия.
- **Базовые принципы.** Руководящие принципы являются фундаментом для деятельности по руководству данными, а также (в особенности) для политики в области DG. Но часто организации вырабатывают политику, вовсе не руководствуясь какими-либо формальными принципами, а просто реагируя на отдельные проблемы. Хотя иногда принципы можно получить из политики путем реверс-инжиниринга, всё же стоит сперва четко сформулировать набор стержневых принципов и лучших практик, а затем на их основе вырабатывать политику. При этом ссылка на принципы может ослабить потенциальное сопротивление изменениям. Со временем в организации выявляются дополнительные руководящие принципы. Их следует публиковать во внутренней среде общего доступа наряду с другими артефактами руководства данными.

1.3 Основные понятия и концепции

Подобно тому как аудитор контролирует финансовые процессы, но не занимается непосредственно управлением финансами, руководство данными обеспечивает надлежащее управление ими, не вмешиваясь напрямую в сами процессы управления (см. рис. 15). Таким образом, введение функции руководства данными отражает *естественное разделение обязанностей между надзором и исполнением.*

[1] The Data Governance Institute, http://bit.ly/1ef0tnb

[2] В англоязычной литературе для обозначения такого влияния используется термин «business-driven» (управляемый бизнесом). — *Примеч. науч. ред.*

Рисунок 15. Руководство и управление данными

1.3.1 Датацентричная организация

Датацентричная (data-centric) организация дорожит данными как активом и управляет ими на всех фазах жизненного цикла, включая проектирование и текущие операции. Чтобы стать по-настоящему датацентричной, организация должна изменить способ трансформации стратегии в действие. Данные перестают считаться побочным продуктом процессов и приложений. Обеспечение высокого качества данных превращается в цель бизнес-процессов. Чем целенаправленнее организация стремится обосновывать принимаемые решения результатами аналитики, тем более важным приоритетом для нее становится эффективное управление данными.

Люди имеют склонность объединять понятия «данные» и «информационные технологии». Чтобы стать датацентричными, организациям нужно начать думать иначе и осознать, что управление данными отличается от управления ИТ. Это не столь просто, как может показаться. Существующая организационная культура с присущими ей внутренними политиками, неопределенностью с предоставлением полномочий владения, конкуренцией за бюджетные средства и унаследованными информационными системами может стать мощным препятствием на пути формирования четкого корпоративного представления о руководстве данными и управлении данными.

Хотя каждой организации необходимо развивать собственные принципы, те, кто стремится к извлечению максимальной ценности из данных, должны придерживаться следующих общих рекомендаций.

- Данными следует управлять как корпоративным активом.
- Организация должна поощрять повсеместное применение лучших практик управления данными.
- Корпоративная стратегия работы с данными должна быть строго согласована с общей стратегией развития бизнеса.
- Процессы управления данными должны непрерывно совершенствоваться.

1.3.2 Организационная система руководства данными

Слово «руководство» (governance) образовано от глагола «*руководить*» (*govern*), который в данном случае является ключевым. Смысл руководства данными проще всего понять на примере политического руководства. В отношении данных предусматриваются функции, подобные законодательным (определение политик, стандартов и корпоративной архитектуры данных), судебным (управление проблемными вопросами и эскалация) и исполнительным (защита и обслуживание, выполнение обязанностей по администрированию). Для лучшего управления рисками большинство организаций выбирают представительную форму руководства данными, обеспечивающую учет мнений всех заинтересованных сторон.

Каждая организация должна принять такую модель руководства данными, которая обеспечит поддержку ее бизнес-стратегии и при этом с наибольшей вероятностью будет иметь успех в ее культурном контексте. Организациям также нужно быть готовыми к дальнейшему развитию принятой модели с целью обеспечения соответствия новым вызовам. Возможные модели отличаются друг от друга по организационной структуре, степени формализации и подходам к принятию решений. Одни модели предусматривают централизованное управление, другие — распределенное.

Организационные системы руководства данными могут быть многоуровневыми, чтобы обеспечивать решение вопросов и проблем на разных уровнях управления — локальном, дивизиональном и корпоративном. Работа по руководству часто распределена между несколькими комитетами, за каждым из которых закреплен собственный круг задач и уровень надзорных полномочий.

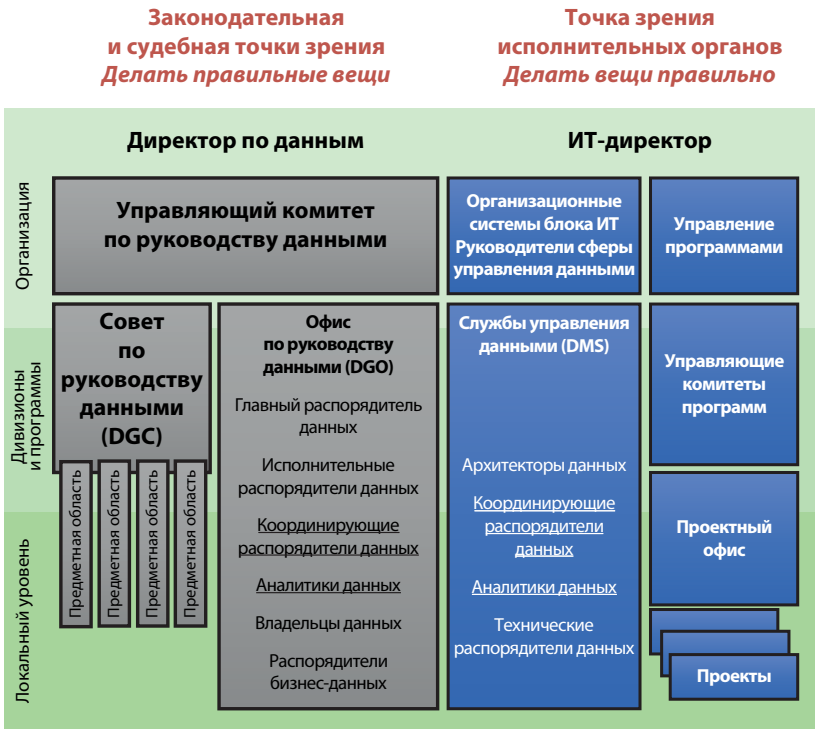


Рисунок 16. Основные элементы организационной системы руководства данными

На рисунке 16 представлена общая модель руководства данными, включающая распределение деятельности по различным уровням управления (вертикальная ось), а также разделение обязанностей по руководству данными внутри функциональных направлений организации и между сферами бизнеса и ИТ (горизонтальная ось). Таблица 4 описывает типичные комитеты и другие органы, которые могут быть созданы в рамках операционной рамочной структуры руководства данными. Следует обратить внимание на то, что это не описание организационной структуры. Схема показывает, каким образом взаимодействуют различные элементы организационной системы в процессе осуществления DG, и в то же время отражает вышеупомянутую тенденцию к уменьшению значения термина *организация* (*organization*) (при употреблении в смысле организационной системы) и к замещению его более общими терминами.

Таблица 4. Типичные комитеты и другие органы руководства данными

Орган руководства данными	Описание
Управляющий комитет по руководству данными	Высший орган руководства данными в организации, отвечающий за надзор, поддержку и финансирование DG. Представляет собой кросс-функциональную группу руководителей высшего звена. Обычно утверждает объемы финансирования деятельности по руководству данными и работ, поддерживаемых со стороны DG (в соответствии с рекомендациями DGC и директора по данным — CDO[1]). Этот комитет, в свою очередь, может быть подконтролен также относящемуся к высшему звену комитету по финансированию или другим комитетам, созданным в рамках отдельных инициатив
Совет по руководству данными (DGC[2])	Управляет инициативами в области DG (например, разработкой политик или метрик), разрешением проблемных вопросов и эскалацией. Состав формируется из руководителей и ответственных сотрудников в соответствии с используемой операционной моделью (см. рис. 17)
Офис по руководству данными (DGO)	Ведет текущую работу в части определений данных и стандартов по управлению данными корпоративного уровня во всех областях знаний DAMA-DMBOK. Формируется из сотрудников с координирующими ролями, которые можно обозначить как *распорядители данных* (*data stewards*) (или *хранители данных* — *data custodians*) и *владельцы данных* (*data owners*)
Команды по распоряжению данными	Заинтересованные группы сотрудников, фокусирующиеся на одной или нескольких предметных областях или проектах и сотрудничающие или осуществляющие взаимные консультации с проектными командами по вопросам определения данных или стандартов по управлению данными, относящимся к области интересов. Состоят из распорядителей данных со стороны бизнеса или ИТ, а также из аналитиков данных
Локальные комитеты (советы) по руководству данными	В крупных организациях могут формироваться комитеты или советы по руководству данными на уровне отдельных дивизионов или департаментов, работающие при содействии и под наблюдением корпоративного DGC. Небольшим организациям лучше избегать подобных сложностей

[1] В данном издании в качестве сокращения термина «директор по данным» используется устоявшаяся ангийская аббревиатура CDO (*сокр*. от *англ*. Chief Data Officer). Аналогичные соглашения используются в отношении названий других должностей (за исключением тех случаев, когда существует устоявшееся русское сокращение). — *Примеч. науч. ред.*

[2] В данном издании в качестве сокращений терминов «Совет по руководству данными» и «Офис по руководству данными» используются устоявшиеся аббревиатуры DGC (*сокр*. от *англ*. Data Governance Council) и DGO (*сокр*. от *англ*. Data Governance Office). Аналогичные соглашения используются в отношении названий других руководящих органов (за исключением тех случаев, когда существует устоявшееся русское сокращение). — *Примеч. науч. ред.*

1.3.3 Типы операционных моделей руководства данными

При централизованной модели одна организационная система руководства данными контролирует все работы по всем предметным областям. В реплицируемой модели одни и те же операционная модель и стандарты DG воспроизводятся в каждой бизнес-единице. Наконец, при федеративной модели одна организационная система руководства данными координирует деятельность нескольких бизнес-единиц с целью обеспечения согласованности определений и стандартов (см. рис. 17 и главу 16).

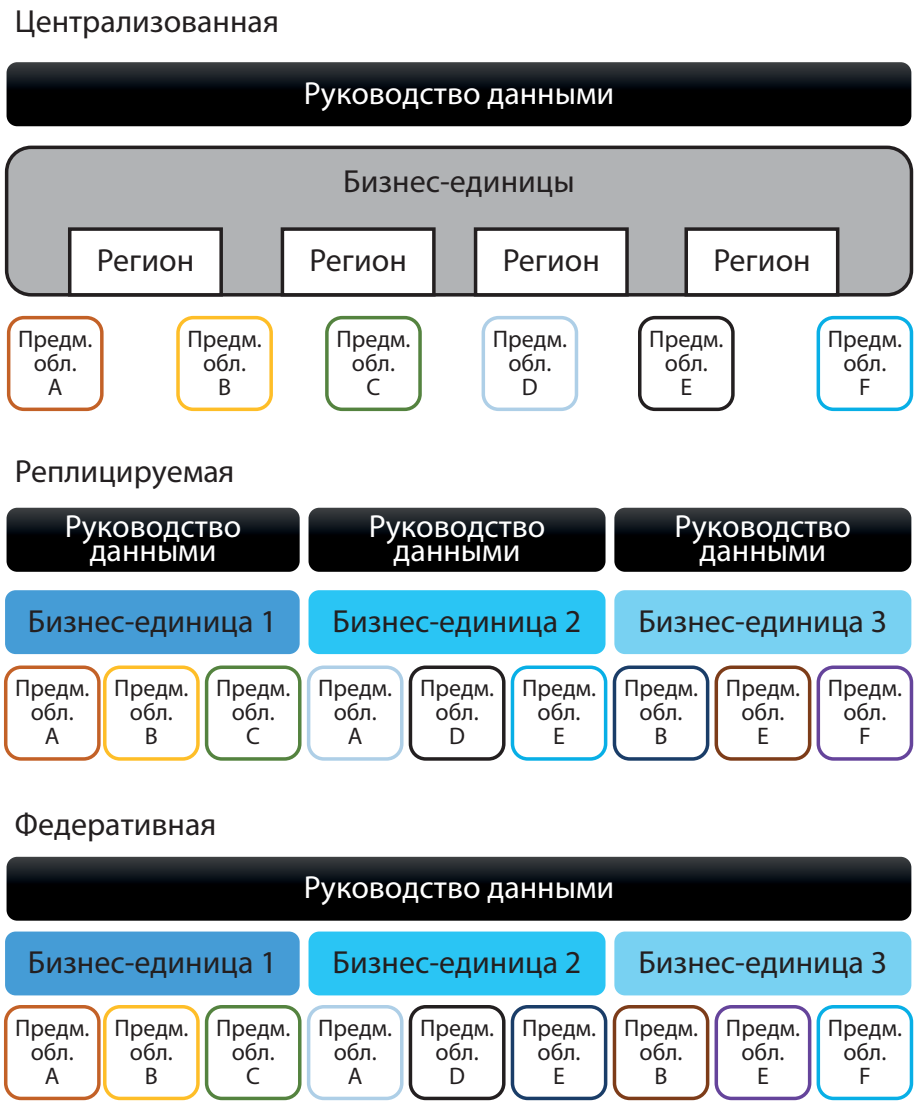


Рисунок 17. Примеры корпоративных операционных рамочных структур руководства данными[1]

[1] За основу взяты примеры из книги: Ladley (2012).

1.3.4 Распоряжение данными

Деятельность, связанная с несением ответственности (и подотчетностью) за данные и процессы, обеспечивающие эффективный контроль и использование информационных активов, чаще всего обозначается термином «*распоряжение данными*» (*data stewardship*). Распоряжение как деятельность может быть формализовано с помощью названий и описаний должностей или быть менее формализованной функцией, осуществляемой людьми, старающимися помочь организации извлечь ценность из своих данных.

Часто по отношению к тем, кто выполняет функции, подобные распорядительским, применяют такие термины-синонимы, как «*хранитель*» (*custodian*) или «*попечитель*» (*trustee*).

Приоритетные направления деятельности по распоряжению отличаются для разных организаций и зависят от стратегии организации, ее культуры, круга решаемых задач, уровня зрелости управления данными и степени формализации программы распоряжения данными. Однако в большинстве случаев работы по распоряжению данными сосредоточены на решении следующих (некоторых или всех) задач.

- **Создание и управление ключевыми метаданными.** Определение и управление бизнес-терминологией, допустимыми значениями и другими критически важными метаданными. Распорядители часто отвечают за корпоративный бизнес-глоссарий, который выполняет функции системы записи для бизнес-терминов, относящихся к данным.
- **Документирование правил и стандартов.** Определение/документирование бизнес-правил, стандартов данных, правил качества данных. Ожидания в отношении данных, используемые для определения данных высокого качества, часто формулируются в виде правил, которые скрыты в бизнес-процессах, создающих или потребляющих данные. Распорядители помогают выявить эти правила, подтвердить их единое понимание в рамках организации и убедиться в том, что они применяются соответствующим образом.
- **Управление проблемными вопросами в области качества данных.** Распорядители часто принимают активное участие в выявлении и разрешении проблемных вопросов, обусловленных недостаточным качеством данных, или содействуют их разрешению.
- **Осуществление текущей деятельности по руководству данными.** Распорядители отвечают за обеспечение постоянного соблюдения политик и поддержки инициатив в области руководства данными в рамках всех реализуемых проектов. Они также должны способствовать принятию решений по управлению данными, создающих условия для достижения организацией своих целей.

1.3.5 Типы распорядителей данных

Распорядитель (*steward*) — это лицо, чья работа заключается в управлении собственностью другого лица. Распорядители данных управляют информационными активами от имени других лиц и в интересах организации (McGilvray, 2008). Распорядители данных представляют интересы всех заинтересованных сторон и должны учитывать корпоративный взгляд на будущее для

обеспечения высокого качества и возможности эффективного использования корпоративных данных. Эффективные распорядители данных несут ответственность за деятельность по руководству данными и часть своего рабочего времени посвящают этой деятельности.

В зависимости от сложности организации и целей ее программы DG официально назначенные распорядители данных могут различаться по своей позиции в организации, направлению работы или по обоим указанным признакам. Например:

- **Главные распорядители данных** (Chief Data Stewards) могут возглавлять органы руководства данными вместо CDO или выступать в качестве CDO в виртуальной (основанной на комитетах) или распределенной организационной системе руководства данными. Они также могут быть исполнительными спонсорами (Executive Sponsors).
- **Исполнительные распорядители данных** (Executive Data Stewards) — это старшие руководители, входящие в состав Совета по руководству данными (DGC).
- **Распорядители корпоративных данных** (Enterprise Data Stewards) осуществляют надзор (oversight) за отдельными областями (domain) данных предприятия в процессе выполнения всех связанных с этими областями бизнес-функций.
- **Распорядители бизнес-данных** (Business Data Stewards) — это бизнес-специалисты, чаще всего признанные эксперты в той или иной предметной области, ответственные за соответствующее подмножество данных. Они работают с заинтересованными лицами (stakeholders) в части определения и контроля данных.
- **Владелец данных** (Data Owner) — это распорядитель бизнес-данных, который обладает подтвержденными полномочиями на утверждение решений, касающихся его области данных.
- **Технические распорядители данных** (Technical Data Stewards) — это ИТ-специалисты, работающие в одной из областей знаний управления данными. Среди них — специалисты по интеграции данных, администраторы баз данных, специалисты по бизнес-аналитике, аналитики качества данных или администраторы метаданных.
- **Координирующие распорядители данных** (Coordinating Data Stewards) возглавляют и представляют команды распорядителей бизнес-данных и технических распорядителей данных в обсуждениях как на уровне команд, так и с участием исполнительных распорядителей данных. Координирующие распорядители данных особенно важны в крупных организациях.

В первом издании DAMA-DMBOK утверждается что «лучшие распорядители данных часто находятся, а не создаются» (DAMA, 2009). Имеется в виду, что в большинстве организаций есть сотрудники, которые распоряжаются данными даже при отсутствии официальной программы руководства данными. Такие сотрудники уже вовлечены в работу по оказанию помощи организации в вопросах снижения связанных с данными рисков и получения наибольшей выгоды от данных. Формализация их ответственности в части распоряжения данными является официальным признанием полезности работы, которую они выполняют, и позволяет им вносить более существенный вклад и достигать большего успеха. И всё же, даже с учетом вышесказанного,

распорядители данных могут быть «созданы»; сотрудники могут пройти соответствующее обучение и стать специалистами в этой области. А те, кто уже занимается распоряжением данными, могут развивать свои навыки и знания и переходить на качественно новые уровни освоения этой деятельности (Plotkin, 2014).

1.3.6 Политики в области данных

Политики в области данных — это директивы, в которых принципы и цели управления данными представлены систематически в виде основных правил, регулирующих создание и приобретение, обеспечение целостности, безопасности и качества, а также использование данных и информации.

Политики в области данных являются всеобъемлющими. Они закладывают основу для создания стандартов данных, а также для формирования ожидаемого поведения, связанного с ключевыми аспектами управления данными и их использования. Политики в области данных сильно различаются в разных организациях. Эти документы описывают «что» (что делать и чего не делать) в части руководства данными, а стандарты и процедуры описывают «как». Политик должно быть относительно немного, и они должны быть изложены кратко и точно.

1.3.7 Оценка информационного актива

Оценка информационного актива — это процесс достижения понимания и расчета экономической ценности данных для организации. Поскольку данные, информация и даже бизнес-аналитика являются абстрактными понятиями, их трудно соотнести с экономическим эффектом. Ключом к определению ценности предметов, не имеющих замены (таких, как данные), является понимание того, как они используются и какие выгоды это приносит (Redman, 1996). В отличие от многих других активов (например, денег или оборудования), наборы данных не являются взаимозаменяемыми. Данные о клиентах одной организации отличаются от данных о клиентах другой по важным критериям; разница заключается не только в самих клиентах, но и связанных с ними данных (история покупок, предпочтения и т. д.). То, каким образом организация извлекает ценность из данных о клиентах (то есть что она узнает о своих клиентах из этих данных и как она применяет эти знания), может являться конкурентным преимуществом.

Большинство фаз жизненного цикла данных связаны с затратами (включая приобретение, хранение, администрирование и ликвидацию). Данные приносят ценность только тогда, когда они используются. При использовании данных также создаются затраты, связанные с управлением рисками. Поэтому ценность возникает, когда экономическая выгода от использования данных превышает затраты на их приобретение и хранение, а также на управление рисками, связанными с их использованием.

Некоторые другие способы оценки данных основаны на учете следующих аспектов.

- **Стоимость замены/восстановления.** Стоимость замены/восстановления данных, потерянных в результате аварии или нарушения целостности, включая используемые в организации данные транзакций, предметных областей, каталогов, документов и метрик.

- **Рыночная стоимость.** Стоимость данных как актива на момент слияния или поглощения.
- **Выявленные возможности.** Величина дохода, который может быть получен от реализации выявленных при работе с данными (в том числе с помощью бизнес-аналитики) возможностей, путем использования данных при заключении сделок или их продажи.
- **Продажа данных.** Некоторые организации оформляют данные в виде товарного продукта или продают ценные результаты, полученные в процессе анализа своих данных.
- **Стоимость рисков.** Оценка потенциальных затрат на выплату штрафов и устранение выявленных нарушений, а также судебных издержек, которые могут возникнуть в результате материализации юридических рисков, обусловленных следующими факторами.
  - Отсутствие данных, которые по закону должны быть в наличии.
  - Наличие данных, которых при определенных обстоятельствах быть не должно (например, незаконно собранные данные, обнаруженные в процессе раскрытия информации в соответствии с требованиями законодательства; данные, которые должны быть удалены, но не были удалены).
  - Неверные данные, наносящие ущерб клиентам, финансовому состоянию компании и ее репутации в дополнение к вышеуказанным расходам.
  - Ущерб, который могут причинить клиентам, компании и ее репутации неточные данные.
  - Снижение риска и цены риска компенсируется операционными затратами на улучшение и сертификацию данных.

Для получения представления о ценности информационного актива можно, к примеру, перевести «Общепринятые принципы бухгалтерского учета» (GAAP) на язык «Общепринятых информационных принципов»[1] (табл. 5).

Таблица 5. Принципы учета информационных активов

Принцип	Описание
Принцип подотчетности	Организация должна определить круг лиц, которые несут полную ответственность (и подотчетны) за данные и контент всех видов
Принцип управления данными как активом	Данные и контент всех видов являются активами и обладают всеми основными характеристиками других активов. Организация должна осуществлять управление ими, а также обеспечивать защиту и учет наравне с прочими материальными и финансовыми активами
Принцип аудита	Правильность данных и контента подлежит периодическому аудиту, проводимому независимой организацией
Принцип должной осмотрительности	Если стало известно о риске, то о нем следует доложить. Если риск возможен, то это следует подтвердить. Риски в области данных включают риски, обусловленные неудовлетворительными практиками управления данными

[1] Идея позаимствована из работы: Ladley (2010), p. 108–109, Generally Accepted Information Principles.

Принцип	Описание
Принцип непрерывности деятельности	Данные и контент жизненно важны для успешного и непрерывного осуществления бизнес-операций и управления (то есть они не рассматриваются в качестве временного средства для достижения результатов или побочного продукта бизнеса)
Принцип разумного уровня оценки	Данные как актив следует оценивать на уровне, который представляется наиболее разумным либо определяется самым простым из возможных способов
Принцип финансовой ответственности	Нарушения существующих юридических и этических норм при обращении с данными и контентом (злоупотребление или неудовлетворительное управление) влекут за собой финансовую ответственность
Принцип качества	Правильно передаваемый смысл, точность и надлежащая поддержка жизненного цикла данных и контента оказывают влияние на финансовое состояние организации
Принцип риска	Владение данными и контентом сопряжено с риском, который нужно учитывать, в том числе и формально, посредством выделения средств либо на его страхование, либо на покрытие издержек от потенциальных последствий его материализации, либо на принятие мер по его минимизации
Принцип ценности	Данные и контент обладают ценностью, зависящей от способов их использования для достижения целей организации, от присущих им свойств товара и/или от их вклада в оценку деловой репутации организации. Ценность информации для организации компенсируется (снижается) за счет затрат на ее обслуживание и перемещение

2. ПРОВОДИМЫЕ РАБОТЫ

2.1 Определение задач и функций руководства данными в организации

Усилия по осуществлению руководства данными должны способствовать реализации стратегии и целей бизнеса. Соответственно, бизнес-стратегия и цели организации лежат в основе стратегии работы с данными и предоставляют базовую информацию о том, каким образом должны проводиться работы по руководству и управлению данными.

Руководство данными позволяет реализовать разделение ответственности за принятие решений, касающихся данных. Работы по руководству данными пересекают организационные границы и границы информационных систем, что обеспечивает единое целостное представление о данных. Успешное руководство данными требует четкого понимания, по отношению к чему и к кому осуществляется руководство и кто именно руководит.

Руководство данными наиболее эффективно, когда оно носит характер усилий, предпринимаемых в масштабах организации, а не изолировано в рамках отдельной функциональной области. Определение границ руководства данными в организации обычно связано с определением предназначения организации (для этого лучше рассматривать организацию как *предприятие* (*enterprise*)[1]). В свою очередь, деятельность по руководству данными способствует осуществлению руководства организацией в соответствии с этим предназначением.

[1] Понятие «предприятие» (enterprise) более подробно обсуждается в главе 4. — *Примеч. науч. ред.*

2.2 Проведение оценки готовности

Значения оценочных показателей, которые описывают текущий уровень возможностей организации по управлению информацией, а также ее зрелости и эффективности в этой области, являются определяющими для планирования программы DG. Поскольку эти показатели могут быть использованы для измерения эффективности самой программы, они также представляют ценность для управления программой и поддержки ее выполнения.

Типичные предметы оценки:

- **Зрелость управления данными.** Следует понять, что именно организация делает с данными; оценить ее текущие возможности и способности по управлению данными. Особое внимание уделяется мнению бизнес-персонала в отношении того, насколько хорошо компания управляет данными и использует их себе на благо, а также объективным критериям, таким как используемые инструменты, уровни отчетности и т. п. (см. главу 15).
- **Способность к изменениям.** Поскольку DG требует изменений в поведении, важно измерить способность организации вносить в привычное поведение необходимые для реализации DG коррективы. Заодно это поможет выявить потенциальные места сопротивления. Часто DG требует введения в организации формального управления изменениями. В процессе управления изменениями должна быть проведена оценка способности организации к изменениям, предполагающая, в свою очередь, оценку существующей организационной структуры, восприятия корпоративной культуры и собственно самого процесса управления изменениями (Hiatt and Creasey, 2012) (см. главу 17).
- **Готовность к сотрудничеству.** Эта оценка характеризует способность организации к налаживанию сотрудничества при управлении и использовании данных. Поскольку такая деятельность, как распоряжение, по определению подразумевает кросс-функциональное взаимодействие, она по своей природе предполагает сотрудничество. Если в организации не умеют взаимодействовать, то такая корпоративная культура станет серьезной помехой для распоряжения данными. Никогда нельзя заранее предполагать, что в организации знают, как нужно сотрудничать. В сочетании с оценкой способности к изменениям оценка готовности к сотрудничеству дает достаточно глубокое понимание способности организации к внедрению DG с точки зрения корпоративной культуры.
- **Согласованность с бизнесом.** Иногда наряду с оценкой способности к изменениям проводится оценка согласованности с бизнесом, позволяющая определить, насколько хорошо организация подстраивает деятельность по использованию данных под стратегию развития бизнеса. Нередко приходится удивляться тому, что работа с данными ведется исключительно в ответ на конкретные запросы.

2.3 Выявление возможностей / угроз и согласование с бизнесом

Программа DG должна вносить вклад в успех организации за счет определения и реализации конкретных выгод (например, снижения размеров штрафов, налагаемых контрольно-надзорными

органами). Проведение работ по выявлению возможностей и угроз позволит определить и оценить эффективность действующих политик и руководств: какие риски они учитывают, а какие нет, какие модели поведения поддерживают и насколько хорошо реализованы на практике. Это также помогает определять новые возможности DG по увеличению полезности данных и контента. А работа по согласованию выявленных возможностей и угроз с бизнес-стратегией обеспечивает связь потенциальных выгод для организации с конкретными элементами программы DG.

Частью работы по выявлению возможностей и угроз является анализ качества данных (DQ[1]). Проведение оценки DQ позволит получить более глубокое представление об имеющихся проблемных вопросах и препятствиях, а также о негативном воздействии и рисках, связанных с низким качеством данных. Оценка DQ позволяет выявить бизнес-процессы, подверженные наибольшему риску при использовании в их реализации некачественных данных, а также определить финансовые и другие выгоды от создания программы обеспечения качества данных в рамках организации работ по руководству данными (см. главу 13).

Еще одним ключевым элементом деятельности, направленной на выявление возможностей и угроз, является проведение оценки сложившихся практик управления данными. При этом, например, можно определить наиболее активных пользователей и сформировать таким образом первоначальный список сотрудников, которые потенциально могли бы обеспечивать постоянную деятельность в области DG.

По итогам проведения работ по выявлению возможностей и угроз и согласованию их с бизнесом следует составить перечень требований к DG. Например, если нормативные риски грозят бизнесу финансовыми проблемами, необходимо перечислить мероприятия в рамках DG, которые обеспечат управление этими рисками. Включенные в перечень требования будут определять стратегию и тактику DG.

2.4 Создание точек взаимодействия внутри организации

Часть работ по обеспечению согласованности с бизнес-процессами связана с разработкой точек взаимодействия внутри организации, необходимых для осуществления руководства данными. Рисунок 18 иллюстрирует примеры точек взаимодействия, обеспечивающих согласованность и связность корпоративных подходов к руководству и управлению данными в областях, находящихся вне сферы полномочий директора по данным.

- **Закупки и контракты.** CDO ведет работу с подразделениями, обеспечивающими управление поставщиками/партнерами или закупками, с целью разработки и внедрения стандартного языка контрактов для его использования в договорах, связанных с управлением данными. Речь может идти, например, о контрактах на закупку данных как услуги (Data-as-a-Service, DaaS), предоставление облачных ресурсов и других аутсорсинговых услуг, проведение работ

[1] В данном издании для сокращения термина «качество данных» используется устоявшаяся аббревиатура DQ (*сокр.* от *англ.* Data Quality). — *Примеч. науч. ред.*

сторонними разработчиками, приобретение контента или лицензионных прав на его использование, а также приобретение и модернизацию датацентричных ИТ-инструментов.

- **Бюджет и финансирование.** Если CDO непосредственно не контролирует все бюджеты, связанные с закупками данных, его служба может стать главным звеном в части предотвращения дублирования усилий и обеспечения оптимизации закупаемых информационных активов.
- **Нормативно-правовое соответствие.** CDO должен четко сознавать, что он ведет работу в сложной нормативно-правовой среде, включающей нормы и правила локального, национального и международного уровня, а также понимать, какое влияние она оказывает на организацию и ее деятельность по управлению данными. С целью идентификации и отслеживания вновь появляющихся и потенциальных нормативно-правовых требований и факторов влияния должен осуществляться непрерывный мониторинг.
- **Жизненный цикл разработки систем (SDLC) / среда разработки.** Программа руководства данными определяет контрольные точки в рамках жизненных циклов систем или приложений, в которых могут быть разработаны корпоративные политики, процессы и стандарты.

Точки взаимодействия, посредством которых CDO оказывает влияние, позволяют поддерживать слаженность действий внутри организации по управлению данными и, как следствие, способствуют повышению оперативности и гибкости их использования. По сути, это ви́дение того, как будет восприниматься организацией руководство данными.

Рисунок 18. Точки взаимодействия CDO с организацией

2.5 Разработка стратегии руководства данными

Стратегия руководства данными определяет объем, содержание и общий подход к проведению работ по руководству. Стратегия DG должна быть комплексной и всеобъемлющей. При этом она должна быть четко связана со стратегией развития бизнеса, а также со стратегией управления данными и стратегией в области ИТ. Внедрение стратегии должно осуществляться итеративно, по мере разработки и утверждения отдельных частей. Содержание стратегии учитывает специфику каждой отдельно взятой организации, однако в любом случае результаты ее разработки должны включать следующее.

- **Общие положения.** Определяют бизнес-драйверы, ви́дение, миссию и принципы руководства данными, включая оценку готовности, внутреннюю оценку процессов, текущих проблем и критериев успеха.
- **Операционная рамочная структура и структура распределения ответственности.** Определяют основные структурные элементы организационного механизма и распределение обязанностей по проведению работ в области DG.
- **Дорожная карта внедрения.** Определяет сроки внедрения политик и руководящих документов, бизнес-глоссария, архитектуры, практики оценки информационных активов, стандартов и процедур, ожидаемых изменений в бизнес- и технологических процессах, документов, необходимых для поддержки работ по аудиту и обеспечения нормативно-правового соответствия.
- **Операционный план.** Описывает последовательность результатов, обеспечивающих устойчивое развитие работ по руководству данными (переход в целевое состояние).

2.6 Определение операционной рамочной структуры руководства данными

Разработка базовой концепции DG обычно особых трудностей не вызывает, а вот создание операционной модели, которую организация примет на практике, может вызывать затруднения. При построении операционной модели DG рекомендуется учитывать следующие аспекты.

- **Ценность данных для организации.** Если организация занимается продажей данных, значительное влияние DG на бизнес очевидно. Организации, использующие данные в качестве ключевого товара (например, Facebook или Amazon), нуждаются в операционной модели, отражающей роль данных в достижении успеха. В то же время в организациях, где данные используются только в качестве «смазочного материала» для осуществления операционной деятельности, программа DG будет реализована менее строго.
- **Бизнес-модель.** Бизнес-модель может быть централизованной или децентрализованной, локальной или международной и т. п., — все факторы подобного рода оказывают влияние на информационные потребности бизнеса и, как следствие, на определение операционной модели DG. В проекте целевой операционной рамочной структуры должны быть отражены специфические для данной организации связи с ИТ-стратегией, архитектурой данных и решениями по интеграции приложений (см. рис. 16).

- **Культурные факторы.** Такие, как дисциплинированность и адаптируемость к изменениям. Некоторые организации будут сопротивляться навязыванию руководства посредством внедрения политик и принципов. В таком случае стратегия руководства данными потребуется для разъяснения преимуществ операционной модели DG, которая при этом должна вписаться в сложившуюся организационную культуру (с учетом последующего проведения поэтапных изменений).

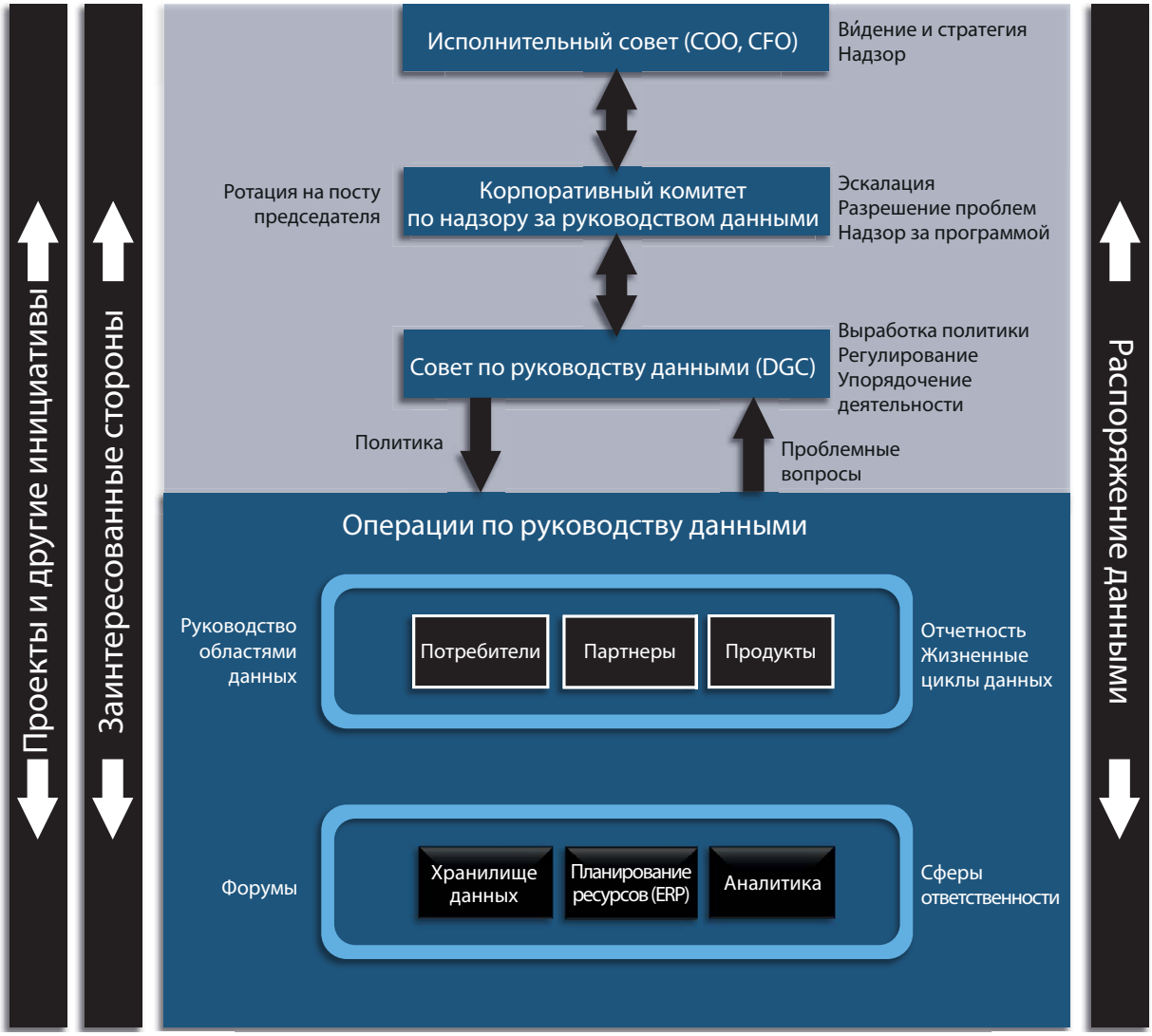


Рисунок 19. Пример операционной рамочной структуры DG

- **Влияние регламентации.** В организациях, деятельность которых сильно зарегламентирована, менталитет сотрудников существенно отличается от менталитета сотрудников в относительно свободных организациях. Соответственно, для организаций первого типа должна

отличаться и операционная модель DG. В частности, в ней могут быть предусмотрены связи с группой управления рисками, а также с юридической службой.

Часто в операционной модели выделяются несколько уровней. Такой подход подразумевает четкое определение сфер ответственности: кто отвечает за проведение работ по распоряжению данными, кто является владельцами данных и т. д. Операционная модель также определяет порядок взаимодействия между организационной системой руководства данными и ответственными за различные проекты или инициативы по управлению данными, порядок согласования и проведения мероприятий по управлению изменениями, необходимыми для реализации этой новой программы, а также модель разрешения проблемных вопросов в рамках руководства данными. Рисунок 19 описывает пример рамочной операционной структуры DG. Пример является иллюстративным. Артефакты подобного рода должны быть адаптированы к потребностям конкретной организации.

2.7 Выработка целей, принципов и политик

Выработка целей, принципов и политик на основе стратегии руководства данными обеспечивает перевод организации в желаемое будущее состояние.

Проекты целей, принципов и политик обычно подготавливаются либо профессионалами в области управления данными, либо сотрудниками, ответственными за разработку бизнес-политики, либо совместно теми и другими при содействии и под наблюдением со стороны представителей сферы руководства данными. Затем проекты рассматривают, дополняют и уточняют распорядители данных и менеджеры организации, после чего они передаются в Совет по руководству данными (или иной орган с аналогичными функциями), который проводит завершающее рассмотрение и корректировку документов, а также их утверждение.

Политики могут принимать самые разные формы, как видно из приведенного ниже примера.

- Офис по руководству данными (DGO) осуществляет сертификацию предназначенных для использования организацией данных.
- Кандидатуры владельцев данных подлежат утверждению DGO.
- Владельцы данных назначают распорядителей данных из своих функциональных областей. Распорядители данных отвечают за координацию текущей работы по руководству данными.
- Во всех возможных случаях для оценки уровня обслуживания большинства бизнес-потребностей вводятся стандартизованная отчетность и/или информационные панели / оценочные ведомости.
- Сертифицированные пользователи получают разрешение на доступ к сертифицированным данным в целях составления срочных/нестандартных отчетов.
- Все сертифицированные данные подлежат регулярной проверке с целью оценки их точности, полноты, согласованности, доступности, уникальности, соответствия предъявляемым требованиям, полезности.

В отношении политик в области данных в организации должны быть приняты меры по доведению их до сведения всех, кого они затрагивают, обеспечению их выполнения и осуществлению контроля выполнения. Действенность политик подлежит периодической оценке с целью их возможной корректировки. Совет по руководству данными может делегировать эти полномочия Управляющему комитету по распоряжению данными.

2.8 Поддержка проектов в области управления данными

Инициативы по расширению и совершенствованию возможностей в сфере управления данными приносят выгоды всей организации. Но они обычно требуют одновременной поддержки со стороны нескольких функциональных направлений или одобрения со стороны Совета по руководству данными. Для них сложно получить финансирование, поскольку в руководстве нередко воспринимают их как нечто отвлекающее от более важных дел, которыми нужно заниматься «здесь и сейчас». Ключ к успешному продвижению таких проектов — в четком обосновании их вклада в повышение эффективности работы и минимизацию рисков. Организациям, желающим извлекать максимум выгоды из имеющихся данных, нужно установить приоритеты в отношении создания и совершенствования возможностей по управлению данными, иначе желаемого результата не достичь.

DGC помогает составлять экономическое обоснование и курирует ход реализации проектов по управлению данными. Если в организации имеется отдельный Проектный офис (Project Management Office, PMO), DGC координирует свою работу с ним. Проекты по управлению данными могут включаться в общий портфель ИТ-проектов.

Также DGC может координировать свою работу по совершенствованию управления данными с крупными программами, реализуемыми в масштабах организации. Такой подход может быть применен, в частности, в рамках проектов, связанных с управлением основными данными, такими как проекты по внедрению систем планирования ресурсов предприятия (Enterprise Resource Planning, ERP), систем управления взаимоотношениями с клиентами (Customer Relationship Management, CRM), глобальных каталогов запасных частей.

Работа по реализации управления данными в рамках смежных проектов должна быть согласована с внутренним жизненным циклом разработки систем (SDLC), организационной системой управления предоставлением услуг, другими компонентами библиотеки инфраструктуры информационных технологий (Information Technology Infrastructure Library, ITIL) и процессами Проектного офиса[1]. Каждый проект, имеющий существенную информационную составляющую (а в наши дни иных практически не осталось), должен учитывать требования по управлению данными уже на *начальных* фазах цикла SDLC (фазах планирования и проектирования). Сюда входят требования к архитектуре, обеспечению нормативно-правового соответствия, идентификации и анализу в системах записи, а также требования по контролю качества и устранению дефектов данных. Также могут быть предусмотрены вспомогательные работы, включая проверочные стендовые испытания на предмет соответствия установленным требованиям.

[1] http://bit.ly/2spRr7e

2.9 Внедрение практики управления организационными изменениями

Управление организационными изменениями (Organizational Change Management, OCM) — стандартный комплекс механизмов проведения изменений в системах и процессах организации. Институт управления изменениями[1] постулирует, что OCM — это «нечто большее, нежели простой учет человеческого аспекта проектов». Его следует рассматривать в качестве подхода, используемого всей организацией в целом для обеспечения эффективного проведения изменений. Организации зачастую управляют чередой следующих друг за другом проектов, вместо того чтобы сосредоточиться на управлении эволюцией организации в целом (Anderson and Ackerson, 2012). Зрелая же в плане управления изменениями организация выстраивает отчетливое ви́дение своего будущего и активно направляет и контролирует изменения, обеспечивающие его достижение, включая идущие сверху вниз общеорганизационные проекты и координацию разработки и реализации малых проектов на местах. Инициативы по изменениям в такой организации гибко адаптируются на основе откликов, полученных в порядке обратной связи и тесного сотрудничества в масштабах всей организации (Change Management Institute, 2012) (см. главу 17).

Для многих организаций формализм и дисциплина, присущие DG, являют собой нечто новое и принципиально отличное от сложившейся практики. Соответственно, чтобы новый порядок прижился, требуется изменить привычки и характер взаимодействия людей. Поэтому необходима формальная программа OCM под эгидой авторитетного спонсора из числа высших руководителей организации — без нее не подвигнуть людей на изменение привычного поведения в направлении, необходимом для реализации устойчиво работающей программы DG. Заручившись поддержкой авторитетного лидера, организации следует создать команду, которая будет отвечать за следующие аспекты.

- **Планирование.** Планирование управления изменениями, включая анализ заинтересованных и затрагиваемых сторон, обеспечение поддержки сверху и определение подхода к коммуникациям, направленным на преодоление сопротивления изменениям.
- **Обучение сотрудников.** Разработка и выполнение учебных планов и программ переподготовки по предметам, необходимым для реализации программ DG.
- **Оказание влияния на разработку систем.** Взаимодействие с Проектным офисом с целью добавления в цикл SDLC шагов, относящихся к DG.
- **Внедрение политик.** Доведение до сотрудников политик в области данных и информации о приверженности организации деятельности по управлению данными.
- **Коммуникации.** Налаживание внутренних коммуникаций с целью повышения осведомленности сотрудников о роли и обязанностях распорядителей данных и других специалистов в области руководства данными, а также о целях и ожидаемых результатах проектов в области управления данными.

[1] Институт управления изменениями (*англ.* The Change Management Institute) — основанная в 2005 году «глобальная, независимая некоммерческая ассоциация профессионалов в области изменений». — *Примеч. пер.*

Коммуникации — жизненно важная составляющая процесса управления изменениями. Программа управления изменениями, обеспечивающая поддержку внедрения DG, должна сфокусировать коммуникации на следующих аспектах.

- **Содействие пониманию ценности данных как актива.** Разъяснение сотрудникам роли данных в реализации целей организации.
- **Мониторинг и обработка отзывов о работах в области DG.** Помимо распространения информации, планами коммуникаций должен быть предусмотрен и сбор отзывов, которые помогут усовершенствовать и саму программу DG, и процесс управления изменениями. Активно интересуясь мнением и полезными предложениями заинтересованных лиц, можно повысить приверженность сотрудников целям программы и одновременно выявить как успехи, так и недочеты, допущенные в ходе ее реализации и подлежащие устранению.
- **Внедрение обучения в области управления данными.** Обучение на всех уровнях организации повышает уровень знания и сознательного применения лучших практик и процессов.
- **Оценка эффектов от управления изменениями в пяти ключевых областях**[1]**:**
  - Осознание необходимости изменений.
  - Желание участвовать в осуществлении изменений, оказывать им поддержку.
  - Понимание того, что и как именно следует изменить в собственной работе.
  - Способность применять новые навыки и изменять свое поведение.
  - Закрепление и необратимость изменений.
- **Внедрение новых метрик и KPI**[2]**.** Система стимулирования работников должна быть пересмотрена с целью поощрения тех, кто придерживается лучших практик управления данными. Поскольку корпоративное руководство данными требует кросс-функционального сотрудничества, новые показатели и стимулы должны предусматривать поощрения за участие в совместных (охватывающих несколько подразделений) инициативах и проектах.

2.10 Внедрение практики управления проблемными вопросами

Процесс управления проблемными вопросами (issue management) включает процедуры выявления, оценки масштаба, приоритизации и разрешения спорных вопросов, имеющих отношение к сфере руководства данными. В их числе:

- **Распределение полномочий.** Проблемные вопросы, касающиеся прав и процедур принятия решений.
- **Эскалация проблем управления изменениями.** Проблемные вопросы, возникающие в процессе управления изменениями, которые не удается решить в рабочем порядке согласно действующим политикам и правилам.

[1] http://bit.ly/1qKvLyJ; см. также: Hiatt and Creasey (2012).

[2] В данном издании в качестве сокращения термина «ключевые показатели эффективности» используется устоявшаяся англоязычная аббревиатура KPI (*сокр.* от *англ.* Key Performance Indicators). — *Примеч. науч. ред.*

- **Нормативно-правовое соответствие.** Вопросы обеспечения соблюдения установленных нормативно-правовых требований.
- **Разрешение конфликтов и противоречий.** Приведение в соответствие между собой противоречивых политик, процедур, бизнес-правил, наименований, определений, стандартов, архитектур, интересов владельцев данных и сторон, заинтересованных в данных и информации.
- **Согласованность.** Проблемные вопросы, возникающие из-за несогласованности с политиками, стандартами, архитектурой и процедурами.
- **Управление договорами.** Вопросы заключения или пересмотра соглашений о совместном использовании данных, договоров купли-продажи данных, контрактов на использование облачных хранилищ.
- **Безопасность и идентичность данных.** Проблемы защиты персональных и конфиденциальных данных, включая расследование нарушений и посягательств.
- **Качество данных.** Выявление и решение проблем качества данных, включая вызванные чрезвычайными ситуациями или взломом защиты.

Многие проблемные вопросы могут быть решены на локальном уровне отдельными сотрудниками, входящими в команды по распоряжению данными. Однако проблемы, требующие обсуждения и/или эскалации, должны регистрироваться и в установленном порядке выноситься на рассмотрение команд по распоряжению данными или выше, на рассмотрение DGC (см. рис. 20). Для выявления тенденций, связанных с обнаруженными проблемами, можно использовать оценочную ведомость (scorecard) руководства данными, позволяющую уточнить детали: где внутри организации она возникла, в чем ее корневые причины и т. п. Проблемные вопросы, которые не могут быть разрешены DGC, выносятся на уровень высшего корпоративного руководства и/или управления.

Рисунок 20. Схема эскалации проблемных вопросов

Руководство данными требует наличия контрольных механизмов и процедур, обеспечивающих:

- выявление, сбор, протоколирование, отслеживание и обновление информации о текущем статусе проблемного вопроса;

- передачу проблемного вопроса исполнителю для принятия мер и контроль исполнения;
- документирование мнений и альтернативных предложений заинтересованных сторон относительно проблемных ситуаций и путей их разрешения;
- принятие, документирование и доведение до сведения сторон резолюций по проблемным вопросам;
- проведение объективных непредвзятых дискуссий и учет всех точек зрения;
- передачу проблемных вопросов на более высокий уровень рассмотрения.

Управление проблемными вопросами — важная функция. Успешное разрешение конфликтных ситуаций и устранение нарушений повышает авторитет команды DG, приносит прямую пользу потребителям данных и помогает снизить нагрузку на производственный персонал. Кроме того, способность устранять проблемы доказывает, что данными можно управлять, а их качество — неуклонно повышать. Но для успешного управления проблемными вопросами необходимы механизмы контроля, предоставляющие наглядную информацию о прилагаемых усилиях по разрешению вопросов и положительном эффекте от их разрешения.

2.11 Оценка требований по нормативно-правовому соответствию

Любая организация работает в нормативно-правовом поле и обязана соблюдать требования законов и нормативно-правовых актов, а также отраслевых норм и инструкций, в том числе касающихся управления данными и информацией. Поэтому составной частью функции руководства данными является мониторинг и обеспечение соблюдения нормативно-правового соответствия. Более того, зачастую именно ради этого первоначально и приступают к внедрению DG. В рамках руководства данными создаются контрольные механизмы, обеспечивающие мониторинг и документирование соблюдения требований, регулирующих порядок управления данными и их использования.

Некоторые регламентирующие документы глобального характера оказывают серьезное влияние на практику управления данными, в частности:

- **Стандарты бухгалтерского финансового учета.** Стандарты учета, устанавливаемые американскими регулирующими органами в этой сфере — Советом по стандартам учета в государственных учреждениях (Government Accounting Standards Board, GASB) и Советом по стандартам финансового учета (Financial Accounting Standards Board, FASB), также существенно влияют на то, как осуществляется управление информационными активами (в США).
- Документы Базельского комитета по банковскому надзору (BCBS) — **BCBS 239** («Принципы агрегирования рисков и отчетности по рискам») и «**Базель II**» (регламентирующее соглашение, определяющее требования к капитализации и другим аспектам деятельности банков) — содержат широкий ряд требований к банковской деятельности. С 2006 года финансовые учреждения, осуществляющие деятельность в странах ЕС, обязаны предоставлять стандартную информацию, подтверждающую ликвидность.

- **CPG 235.** Австралийское управление пруденциального регулирования (The Australian Prudential Regulation Authority, APRA) осуществляет надзор за банковской и страховой деятельностью. Оно публикует стандарты и руководства по выполнению требований этих стандартов. Среди них стандарт CPG 235, регламентирующий вопросы управления рисками, связанными с данными. Основное внимание в стандарте уделяется источникам риска и управлению данными на протяжении всего их жизненного цикла.
- **PCI-DSS** — стандарты защиты информации в индустрии платежных карт (The Payment Card Industry Data Security Standards).
- **Solvency II** — регламент ЕС для отрасли страхования, в целом воспроизводящий требования соглашения «Базель II».
- **Законы о конфиденциальности.** Должно обеспечиваться соблюдение всех местных, национальных и международных законов в этой сфере.

Органы руководства данными совместно с руководством бизнеса и технического блока проводят работу по оценке влияния действующих регулирующих документов на организацию. Они должны, например, ответить на следующие вопросы.

- Каким образом требования документа затрагивают организацию?
- Что понимается под соответствием требованиям? Какие правила и процедуры необходимы для обеспечения соответствия?
- Когда требуется соответствие? Как и когда отслеживается соблюдение соответствия?
- Если организация начнет применять отраслевые стандарты, достаточно ли этого для обеспечения соответствия требованиям?
- Как можно продемонстрировать соответствие?
- Какие риски и штрафные санкции влечет за собой несоответствие?
- Каким образом можно выявить несоответствие и сообщить о нем? Как осуществляется управление несоответствием и его устранение?

В рамках DG должен быть обеспечен контроль реагирования организации на нормативно-правовые требования или результаты проверок данных и действующих практик работы с данными (например, с помощью сертификации качества данных, включаемых в отчетность для регулирующего органа) (см. главу 6).

2.12 Внедрение руководства данными

Руководство данными не может быть внедрено моментально. Требуется тщательное планирование его внедрения — и не только из-за необходимости учесть все необходимые организационные изменения, но и просто из-за многогранности и сложности комплекса проводимых работ, которые должны быть скоординированы. Лучше всего создать дорожную карту внедрения с указанием сроков и взаимосвязей между работами по разным направлениям. Например, если в центре

внимания программы DG находятся вопросы обеспечения нормативно-правового соответствия, определяющее влияние на расстановку приоритетов должны оказывать требования регулирующих органов. При федеративной операционной модели DG (см. рис. 17) внедрение по различным направлениям бизнеса может проходить по разным графикам, которые зависят от уровней их вовлеченности и зрелости, а также от выделения средств.

Часть работ по внедрению закладывает фундамент руководства данными. От них зависят остальные внедренческие мероприятия. Эти работы выполняются в первую очередь и требуют постоянной поддержки и развития. На начальных стадиях приоритетными являются следующие направления деятельности.

- Определение процедур руководства данными, необходимых для достижения наиболее приоритетных целей.
- Выработка, согласование и ввод в действие бизнес-глоссария, а также документирование терминологии и стандартов.
- Согласование с корпоративной архитектурой и архитектурой данных для обеспечения лучшего понимания данных и систем.
- Финансовая оценка информационных ресурсов в целях оптимизации принятия решений и улучшения понимания роли данных в обеспечении успешной работы организации.

2.13 Поддержка стандартов и процедур

Стандарт определяется как «некоторая вещь, которая является очень хорошей и используется для вынесения суждений о качестве других вещей» или как «устанавливаемое полномочным органом правило измерения количества, веса, габаритов, ценности или качества»[1]. Стандарты помогают определять само понятие качества и судить о качестве вещей, выступая базой для сравнения. Они также предоставляют возможности для упрощения процессов. Утверждая стандарт, организация единожды принимает взвешенное решение и кодифицирует его в виде набора формальных утверждений (стандарта). Это избавляет от необходимости всякий раз заново обсуждать и принимать решение по тому же вопросу в рамках каждого нового проекта. Обеспечение соблюдения стандартов способствует получению согласующихся результатов от процессов, в которых эти стандарты применяются.

К сожалению, процесс разработки или принятия стандартов зачастую политизируется — и сформулированные выше цели забываются. Что касается стандартов в области данных руководства данными, то у большинства организаций просто нет достаточного опыта их разработки и внедрения. В некоторых случаях отсутствует понимание выгод от этой деятельности и, следовательно, ей не уделяют времени. В результате то, что называют «стандартами», очень широко варьируется не только от организации к организации, но и в пределах каждой из них; то же самое касается и ожиданий в отношении их соблюдения. Между тем стандарты DG должны быть обязательными для всех.

[1] http://bit.ly/2sTfugb

Стандарты данных могут принимать различные формы в зависимости от того, что они описывают: правила заполнения полей, правила, определяющие связи между полями, допустимые и недопустимые значения, форматы и т. д. Проект стандарта обычно создается профессионалами в области управления данными. Затем он должен быть рассмотрен, одобрен и принят DGC или специально созданной рабочей группой — например, Управляющим комитетом по стандартам данных. Уровень детализации стандартов при их документальном оформлении зависит отчасти от организационной культуры. Важно помнить, что документирование дает возможность зафиксировать детальную информацию или знания, которые без такой фиксации могут быть потеряны. Воссоздание или реверс-инжиниринг с целью получения доступа к этим знаниям являются очень затратными по сравнению с документированием в самом начале их выявления.

Стандарты данных должны оперативно доводиться до сведения заинтересованных лиц, контролироваться, а также периодически пересматриваться и обновляться. Но главное — предусмотреть средства обеспечения их соблюдения. Данные можно оценить путем сравнения со стандартами. Проверки работ по управлению данными могут проводиться DGC или Управляющим комитетом по стандартам данных либо по установленному графику, либо как часть процессов согласования в рамках цикла SDLC.

Процедуры управления данными — это документированные методы, технические приемы и шаги, которых необходимо придерживаться при выполнении определенных видов работ, обеспечивающих получение конкретных результатов, подкрепленных соответствующими артефактами. Процедуры — так же как политики и стандарты — сильно варьируются от организации к организации. И, как и в случае со стандартами данных, документы, описывающие процедуры, в явном виде фиксируют знания организации. Проекты документов, описывающих процедуры, обычно составляются профессионалами в области управления данными.

Примеры объектов стандартизации, которые можно выделить в рамках отдельных областей знаний по управлению данными, включают следующее.

- **Архитектура данных.** Корпоративные модели данных, стандарты на инструменты, соглашения об именовании систем.
- **Моделирование и проектирование данных.** Процедуры управления моделями данных, соглашения об именовании при моделировании, стандарты описаний, стандартные области и стандартные сокращения.
- **Хранение и операции с данными.** Стандарты на инструменты, стандарты в отношении восстановления баз данных и обеспечения непрерывности бизнеса, производительности баз данных, сохранения данных, сбора данных.
- **Безопасность данных.** Стандарты безопасности в отношении доступа к данным, процедуры мониторинга и аудита, стандарты по безопасности хранилищ данных, требования к подготовке персонала.
- **Интеграция и интероперабельность данных.** Стандартные методы и инструменты, используемые для обеспечения интеграции и интероперабельности данных.

- **Документы и контент.** Стандарты и процедуры в области управления контентом, включая использование корпоративных таксономий, поддержку раскрытия информации по запросам уполномоченных органов, сроки хранения документов и электронной почты, электронные подписи, порядок рассылки отчетов.
- **Справочные и основные данные.** Процедуры контроля в области управления справочными данными, системы записи данных, правила добавления записей и обеспечения обязательности их использования, стандарты по разрешению сущностей.
- **Ведение хранилищ данных и бизнес-аналитика.** Стандарты на инструменты, стандарты и процедуры обработки, стандарты форматирования отчетов и визуальных представлений, стандарты обработки больших данных.
- **Метаданные.** Стандартные бизнес- и технические метаданные, подлежащие сбору, а также процедуры интеграции и практики использования метаданных.
- **Качество данных.** Правила качества данных, стандартные методологии измерения показателей *качества данных*, стандарты и процедуры исправления данных.
- **Большие данные и наука о данных.** Идентификация источника, основания для сбора и порядок получения, система записи, распространение и обновление данных.

2.14 Разработка бизнес-глоссария

За содержание бизнес-глоссария обычно отвечают распорядители данных. Глоссарий необходим по той причине, что люди по-разному используют одни и те же слова. Особенно важно иметь четкие определения для данных, поскольку данные представляют не сами себя, а другие предметы (Chisholm, 2010). Кроме того, многие организации разрабатывают свой собственный внутренний словарь терминов, и глоссарий служит средством совместного использования этого словаря внутри организации. Разработка и документирование стандартных определений данных снижает неопределенность и улучшает коммуникации. Определения терминов должны быть четкими и ясными, строго и точно сформулированными, а также объяснять все исключения, синонимы или варианты использования. В состав сотрудников, участвующих в согласовании терминологии, должны входить представители ключевых групп пользователей. Архитектура данных часто может предоставить первоначальные варианты определений и разбивок терминов по типам и категориям из моделей предметных областей.

Основные задачи бизнес-глоссариев включают:

- обеспечение единого общего понимания основных понятий и терминов;
- снижение риска неправильного использования данных из-за неверного понимания связанных с бизнесом понятий и концепций;
- повышение терминологической согласованности между подразделениями, отвечающими за технологические ресурсы (с собственными соглашениями об именовании), и подразделениями, отвечающими за организацию и ведение бизнеса;

- максимизацию возможностей поиска и обеспечения доступа к документированным знаниям организации.

Бизнес-глоссарий — это не просто перечень терминов с определениями. Каждый термин должен быть связан с другими полезными метаданными: синонимами, метриками, данными о происхождении, бизнес-правилами, информацией о распорядителе данных, отвечающем за термин, и т. д.

2.15 Координация взаимодействия с архитектурными группами

Совет по руководству данными курирует разработку и утверждает артефакты архитектуры данных, такие как бизнес-ориентированная корпоративная модель данных. DGC также может создать Управляющий комитет по архитектуре данных предприятия (Enterprise Data Architecture Steering Committee) или Наблюдательный совет по архитектуре (Architecture Review Board, ARB) либо (при их наличии) взаимодействовать с ними для осуществления надзора за программой и ее итеративными проектами. Корпоративная модель данных должна разрабатываться и вестись общими усилиями архитекторов и распорядителей данных, работающих вместе в командах по предметным областям. В зависимости от особенностей организации эта работа может координироваться либо корпоративным архитектором данных, либо распорядителем данных. По мере возрастания бизнес-требований рабочие команды распорядителей данных должны вносить предложения по изменениям и разрабатывать расширения корпоративной модели данных.

Корпоративная модель данных должна быть рассмотрена, одобрена и формально принята Советом по руководству данными. Модель должна быть взаимоувязана со всеми ключевыми бизнес-стратегиями, процессами, организационными структурами и информационными системами. Стратегия работы с данными и архитектура данных являются центральными звеньями в обеспечении координации между сферами ответственности за то, чтобы «*делать вещи правильно*», и за то, чтобы «*делать правильные вещи*» в процессе управления данными.

2.16 Оказание содействия в финансовой оценке данных

Данные и информация являются активами, поскольку имеют ценность или могут ее создавать. В современной практике бухгалтерского учета данные принято относить к нематериальным активам наряду с программным обеспечением, документацией, экспертными знаниями, коммерческими тайнами и прочими объектами интеллектуальной собственности. Таким образом, денежная оценка данных представляет для организаций большую сложность. Поэтому DGC следует приложить усилия в этой области и установить соответствующие стандарты.

В некоторых организациях эта работа начинается с оценки размеров финансовых потерь бизнеса из-за недостаточной или недостоверной информации. Информационные пробелы (gaps) — то есть разница между необходимой и доступной информацией — возлагают на бизнес дополнительные обязательства. Затраты на ликвидацию или предотвращение возникновения таких

пробелов можно использовать в качестве оценки стоимости недостающих данных. Отталкиваясь от этого, организация может разработать модели оценки той информации, которая имеется в ее распоряжении.

Стоимостные оценки можно встраивать в дорожную карту стратегии работы с данными и применять их для экономического обоснования решений, призванных устранить корневые причины проблем с качеством данных, а также использовать в рамках иных инициатив по руководству данными.

2.17 Встраивание руководства данными в процессы

Одна из целей создания организационной системы руководства данными заключается во внедрении в многообразный спектр процессов организации моделей поведения, связанных с управлением данными как активом. Текущая деятельность по руководству данными требует планирования. Операционный план включает перечень шагов, обеспечивающих внедрение и непосредственное осуществление DG, с указанием соответствующих мероприятий, сроков и методов обеспечения устойчивого успеха.

Устойчивость подразумевает создание условий, гарантирующих наличие процессов и финансирования, необходимых для непрерывного функционирования организационной системы DG в масштабах организации. К главным условиям следует отнести *признание* организацией необходимости руководства данными, осуществление управления функцией DG, мониторинг и измерение результатов, а также устранение или преодоление типичных препятствий на пути к успешной реализации программы DG.

С целью углубления понимания организацией концепции руководства данными в целом и ее применения на местах, а также для обмена знаниями и опытом целесообразно создать сообщество по интересам в области руководства данными. Польза от сообщества особенно велика в первые годы реализации DG, но она, скорее всего, будет уменьшаться по мере повышения уровня зрелости этой деятельности.

3. ИНСТРУМЕНТЫ И МЕТОДЫ

Руководство данными в своей основе связано с организационным поведением. Это не та задача, которую можно решить с помощью технологий. Однако есть инструменты, поддерживающие процесс в целом. Например, руководство данными подразумевает постоянный обмен информацией. Программа DG должна в полной мере использовать возможности существующих коммуникационных каналов для регулярного доведения ключевых сообщений до сведения всех заинтересованных сторон, чтобы они всегда были в курсе последних изменений в политиках, стандартах и требованиях.

Кроме того, программа DG должна эффективно управлять собственной работой и собственными данными. Определенные инструменты способны помочь не только с решением

этих задач, но и с метриками, которые их поддерживают. Прежде чем выбирать инструмент для выполнения какой-либо конкретной функции, например решения по реализации бизнес-глоссария, организации следует определить общие цели и требования к DG с прицелом на возможность построения набора инструментов. В частности, некоторые решения по реализации глоссариев включают дополнительные компоненты для управления политиками и потоками работ. Если подобная функциональность необходима, следует уточнить требования к ней и провести тестирование на соответствие этим требованиям, прежде чем принимать решение об использовании инструмента. Иначе организация рискует обрасти множеством инструментов с дублирующими друг друга функциями, ни один из которых не будет по-настоящему соответствовать ее нуждам.

3.1 Присутствие в Сети / Веб-сайты

Программа руководства данными должна быть представлена в Сети. Основные документы могут публиковаться на собственном сайте программы или на корпоративном портале. Сайты могут хранить библиотеки документов, предоставлять доступ к поисковым функциям, обеспечивать управление простыми потоками работ. Сайт также способен помочь в создании бренда программы DG посредством использования логотипа и единообразного визуального представления материалов. На официальном сайте программы DG должны быть представлены:

- основные положения стратегии и программы руководства данными, включая формулировки ви́дения, миссии, целей и задач, описание преимуществ и дорожную карту внедрения;
- политики в области данных и стандарты данных;
- описания ролей и обязанностей распорядителей данных;
- новости, касающиеся программы;
- ссылки на тематические форумы сообщества по интересам в области руководства данными;
- ссылки на сообщения от руководства по вопросам, касающимся DG;
- отчеты о проведенных оценках качества данных;
- процедуры выявления и эскалации проблемных вопросов;
- ссылки на сервисы регистрации заявок и проблемных вопросов;
- документы, презентации и учебные программы, а также ссылки на соответствующие онлайн-ресурсы;
- контактная информация программы руководства данными.

3.2 Бизнес-глоссарий

Бизнес-глоссарий — ключевой инструмент DG. Он содержит согласованные определения бизнес-терминов и связывает их с данными. Сегодня доступно много инструментальных средств для создания глоссариев. Некоторые из них входят в состав более крупных продуктов (в частности, систем планирования ресурсов предприятия (ERP), решений по интеграции данных или управлению метаданными), а некоторые являются самостоятельными инструментами.

3.3 Инструменты для управления потоками работ

Организации покрупнее могут при желании воспользоваться каким-либо надежным инструментом для управления потоками работ, чтобы управлять процессами (например, процессом внедрения новых политик в области данных). Подобные инструменты позволяют привязывать процессы к документам и могут быть полезны при управлении политиками и разрешении проблемных вопросов.

3.4 Инструменты для управления документами

Инструменты для управления документами очень часто используются командами руководства данными как вспомогательное средство для управления политиками и процедурами.

3.5 Оценочная ведомость руководства данными

Подборка показателей для отслеживания работ по руководству данными и соответствия политикам может предоставляться Совету по руководству данными и Управляющему комитету по руководству данными в автоматическом режиме с помощью электронных оценочных ведомостей.

4. РЕКОМЕНДАЦИИ ПО ВНЕДРЕНИЮ

Как только определена программа руководства данными, а также разработан операционный план и подготовлена дорожная карта внедрения (на основе информации, собранной в рамках процесса оценки зрелости управления данными; см. главу 15), организация может начинать реализацию процессов и политик. Большинство стратегий развертывания носят постепенный ступенчатый характер, начиная с внедрения DG либо в какой-либо из крупных функциональных областей, такой как управление основными данными (MDM), либо в отдельном регионе или подразделении. Случаи внедрения DG сразу в масштабах всей организации встречаются редко.

4.1 Организация и культура

Как отмечалось в разделе 2.9, формализм и дисциплина, свойственные руководству данными, для многих организаций являются новыми и необычными. Руководство данными добавляет ценность посредством внесения изменений в поведение. Вполне вероятны сопротивление изменениям, нежелание переучиваться и усваивать новые методы принятия решений и управления проектами.

Эффективные и долгосрочные программы руководства данными требуют культурного сдвига в организационном мышлении и поведении по отношению к данным. Кроме того, требуется наличие постоянно действующей программы управления изменениями, поддерживающей новое мышление, образ действий, политики и процессы для достижения нового желаемого поведения. Вне зависимости от того, насколько четкой или экзотичной является стратегия руководства данными, игнорирование культурных факторов приведет к снижению ее шансов на успех. Нацеленность на управление изменениями должна являться неотъемлемой частью стратегии внедрения.

Целью организационных изменений является устойчивость. Устойчивость — это характеристика процесса, позволяющая оценить его способность продолжать добавлять ценность. Для обеспечения устойчивости программы руководства данными требуется планирование изменений (см. главу 17).

4.2 Согласование действий и коммуникации

Программы руководства данными реализуются поэтапно в контексте более широких стратегий развития бизнеса и управления данными. Поэтому для достижения успеха необходим учет расширенного спектра целей и задач. Команде DG нужно проявлять гибкость и оперативно корректировать подходы к управлению данными по мере изменения условий. Ниже перечислены инструменты, которые требуются для управления изменениями и своевременного оповещения о них.

- **Стратегическая карта развития бизнеса / DG.** Карта увязывает работы по DG с потребностями бизнеса. Периодическая оценка и доведение до всеобщего сведения вклада DG в развитие бизнеса жизненно важны для обеспечения поддержки программы на постоянной основе.
- **Дорожная карта DG.** Дорожная карта DG не должна быть жесткой. Она должна быть гибкой и предусматривать возможность адаптации к текущим изменениям внешних условий или приоритетов бизнеса.
- **Регулярно обновляемое экономическое обоснование DG.** Экономическое обоснование должно регулярно обновляться с целью учета изменения приоритетов и финансово-экономических реалий внутри организации.
- **Метрики DG.** По мере повышения уровня зрелости DG растет количество и меняется состав используемых метрик.

5. МЕТРИКИ

Для противодействия сопротивлению или проблемам, обусловленным продолжительной кривой обучения (learning curve), программа DG должна предполагать оценку прогресса и успеха с помощью метрик, которые демонстрируют, каким образом участники программы добавляют бизнес-ценность и достигают поставленных целей.

Для управления требуемыми изменениями в поведении важно отслеживать измеримые показатели прогресса внедрения руководства данными и соблюдения требований DG, а также увеличения ценности, которую руководство данными приносит организации. Метрики, подтверждающие ценность DG и наличие у организации всех необходимых ресурсов для продолжения эффективного руководства данными по завершении развертывания программы DG, также необходимо фиксировать, чтобы гарантировать ее долгосрочную устойчивость. Примеры измеримых показателей:

- Ценность
  - Вклад в достижение бизнес-целей
  - Снижение риска
  - Повышение эффективности операций
- Эффективность
  - Достижение целей и решение задач
  - Масштаб использования соответствующих инструментов распорядителями данных
  - Эффективность коммуникаций
  - Эффективность обучения/тренингов.
  - Скорость адаптации к изменениям
- Устойчивость
  - Реализация политик и процедур (работают ли они в соответствии с ожиданиями)
  - Соблюдение стандартов и процедур (следуют ли сотрудники правилам и инструкциям, а также изменяется ли их поведение в нужную сторону)

6. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Adelman, Sid, Larissa Moss and Majid Abai. *Data Strategy*. Addison-Wesley Professional, 2005. Print.

Anderson, Dean and Anderson, Linda Ackerson. *Beyond Change Management*. Pfeiffer, 2012.

Avramov, Lucien and Maurizio Portolani. *The Policy Driven Data Center with ACI: Architecture, Concepts, and Methodology*. Cisco Press, 2014. Print. Networking Technology.

Axelos Global Best Practice (ITIL website), http://bit.ly/1H6SwxC

Brzezinski, Robert. *HIPAA Privacy and Security Compliance — Simplified: Practical Guide for Healthcare Providers and Practice Managers*. CreateSpace Independent Publishing Platform, 2014. Print.

Calder, Alan. *IT Governance: Implementing Frameworks and Standards for the Corporate Governance of IT*. IT Governance Publishing, 2009. Print.

Change Management Institute and Carbon Group. *Organizational Change Maturity Model*, 2012, http://bit.ly/1Q62tR1

Change Management Institute (website), http://bit.ly/1Q62tR1

Chisholm, Malcolm and Roblyn-Lee, Diane. *Definitions in Data Management: A Guide to Fundamental Semantic Metadata*. Design Media, 2008. Print.

Cokins, Gary et al. *CIO Best Practices: Enabling Strategic Value with Information Technology*, 2nd ed. Wiley, 2010. Print.

De Haes, Steven and Wim Van Grembergen. *Enterprise Governance of Information Technology: Achieving Alignment and Value, Featuring COBIT 5*. 2nd ed. Springer, 2015. Print. Management for Professionals.

DiStefano, Robert S. *Asset Data Integrity Is Serious Business*. Industrial Press, Inc., 2010. Print.

Doan, AnHai, Alon Halevy and Zachary Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012. Print.

Fisher, Tony. *The Data Asset: How Smart Companies Govern Their Data for Business Success*. Wiley, 2009. Print.

Giordano, Anthony David. *Performing Information Governance: A Step-by-step Guide to Making Information Governance Work*. IBM Press, 2014. Print. IBM Press.

Hiatt, Jeff and Creasey, Timothy. *Change Management: The People Side of Change*. Prosci, 2012.

Huwe, Ruth A. *Metrics 2.0: Creating Scorecards for High-Performance Work Teams and Organizations*. Praeger, 2010. Print.

Ladley, John. *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program*. Morgan Kaufmann, 2012. Print. The Morgan Kaufmann Series on Business Intelligence.

Ladley, John. *Making Enterprise Information Management (EIM) Work for Business: A Guide to Understanding Information as an Asset*. Morgan Kaufmann, 2010. Print.

Marz, Nathan and James Warren. *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications, 2015. Print.

McGilvray, Danette. *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. Morgan Kaufmann, 2008. Print.

Osborne, Jason W. *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. SAGE Publications, Inc, 2013. Print.

Plotkin, David. *Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance*. Morgan Kaufmann, 2013. Print.

PROSCI (website), http://bit.ly/2tt1bf9

Razavi, Behzad. *Principles of Data Conversion System Design*. Wiley-IEEE Press, 1994. Print.

Redman, Thomas C. *Data Driven: Profiting from Your Most Important Business Asset*. Harvard Business Review Press, 2008. Print.

Reinke, Guido. *The Regulatory Compliance Matrix: Regulation of Financial Services, Information and Communication Technology, and Generally Related Matters*. GOLD RUSH Publishing, 2015. Print. Regulatory Compliance.

Seiner, Robert S. *Non-Invasive Data Governance*. Technics Publications, LLC, 2014. Print.

Selig, Gad. *Implementing IT Governance: A Practical Guide to Global Best Practices in IT Management*. Van Haren Publishing, 2008. Print. Best Practice.

Smallwood, Robert F. *Information Governance: Concepts, Strategies, and Best Practices*. Wiley, 2014. Print. Wiley CIO.

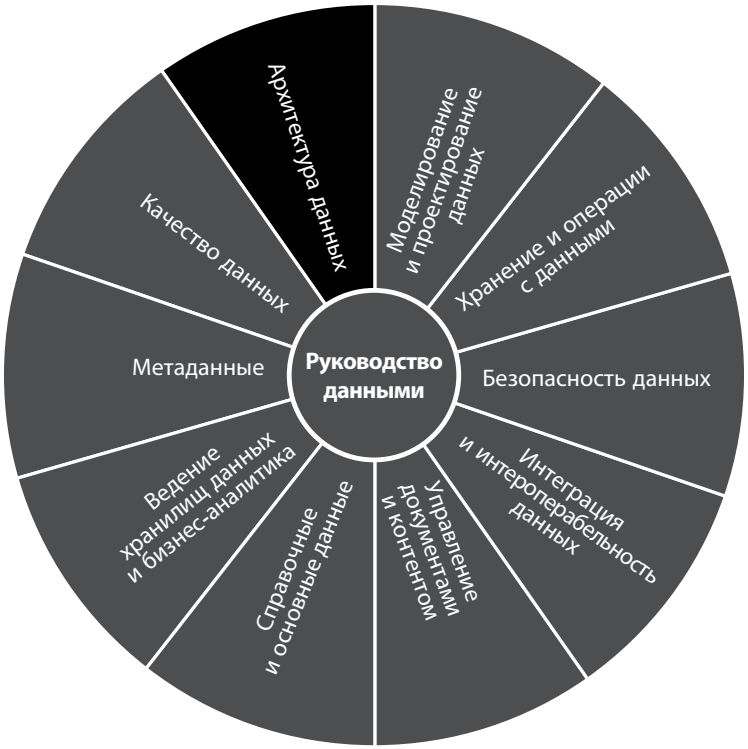
Soares, Sunil. *Selling Information Governance to the Business: Best Practices by Industry and Job Function*. MC Press, 2011. Print.

Tarantino, Anthony. *The Governance, Risk, and Compliance Handbook: Technology, Finance, Environmental, and International Guidance and Best Practices*. Wiley, 2008. Print.

The Data Governance Institute (website), http://bit.ly/1ef0tnb

The KPI Institute and Aurel Brudan, ed. *The Governance, Compliance and Risk KPI Dictionary: 130+ Key Performance Indicator Definitions*. CreateSpace Independent Publishing Platform, 2015. Print.

Архитектура данных



DAMA-DMBOK2 Data Management Framework

1. ВВЕДЕНИЕ

Архитектурой принято называть как искусство и науку о проектировании и строительстве чего-либо (в особенности жилых зданий), так и результаты этой деятельности — то есть сами строения. В более общем смысле архитектурой называют упорядоченную компоновку составляющих элементов, предполагающую оптимизацию функциональности, производительности, технологичности, стоимости и эстетичности создаваемой конструкции или системы.

Термин «архитектура» также применяется для описания отдельных аспектов проектирования информационных систем. Стандарт ISO/IEC/IEEE 42010:2011 «Systems and software engineering. Architecture description» («Системная и программная инженерия. Описание архитектуры») определяет

архитектуру как «основные понятия или свойства системы в окружающей среде, воплощенной в ее элементах, отношениях и конкретных принципах ее проекта и развития»[1]. Однако на практике под термином «архитектура» в зависимости от контекста может подразумеваться описание текущего состояния систем, компонентов совокупностей систем, проектирование систем (как дисциплина и практика), планируемый проект системы или совокупности систем (будущее состояние или предполагаемая архитектура), артефакты, характеризующие систему (архитектурная документация), или команда, выполняющая работу по проектированию (архитекторы или архитектурная команда).

Архитектурная практика осуществляется на различных уровнях внутри организации (организация в целом, направление деятельности, проект и т. д.) и применительно к различным аспектам (инфраструктура, приложения или данные). Людям, не связанным с архитектурой и не разбирающимся во всех различиях между названными уровнями и аспектами, легко в них запутаться, а потому бывает сложно понять, чем именно занимаются архитекторы. Одна из причин, по которой архитектурные рамочные структуры (*architecture framework*)[2] являются крайне полезными, как раз и заключается в том, что они помогают неспециалистам получить представление о связях и соотношениях между этими компонентами.

Дисциплина «*Архитектура предприятия*» (*Enterprise Architecture*)[3] охватывает архитектуры нескольких предметных областей (доменов — domains), включая бизнес, данные, приложения и технологии. Отлаженные практики управления архитектурой предприятия помогают организации четко понимать текущее состояние своих информационных систем, проводить изменения, направленные на переход в желаемое будущее состояние, обеспечивать соблюдение нормативно-правовых требований, повышать эффективность и производительность своей работы. Эффективное управление данными и системами хранения и использования данных — одна из общих целей всех разделов науки об архитектуре предприятия.

В настоящей главе архитектура данных рассматривается с трех позиций.

- **Выходные результаты архитектуры данных**, такие как модели, определения и описания потоков данных на различных уровнях, то есть всё то, что принято называть артефактами архитектуры данных.
- **Работы, проводимые в области архитектуры данных**, такие как формирование, развертывание и внедрение целевых решений в области архитектуры данных.

[1] См.: ГОСТ Р 57100-2016/ISO/IEC/IEEE 42010:2011. — *Примеч. пер.*

[2] В ГОСТ Р 57100-2016 понятие *architecture framework* переведено как «структура архитектуры» и определено как «условности, принципы и практики для описания архитектур, установленные в пределах заданной области применения и/или объединения заинтересованных сторон». — *Примеч. пер.*

[3] В данном случае под «предприятием» (*enterprise*) понимается одна или несколько организаций, разделяющих определенную миссию, цели и задачи для получения выхода (результата) в виде продукции или услуги. См., например: ГОСТ Р ИСО 15704-2008 «Промышленные автоматизированные системы. Требования к стандартным архитектурам и методологиям предприятия». Таким образом, в качестве предприятия может рассматриваться как целая корпорация, так и ее подразделение; как государственное учреждение, так и коммерческая фирма или, например, несколько фирм с общими владельцами. — *Примеч. науч. ред.*

- **Организационное поведение в рамках работ в области архитектуры данных**; в частности, сюда относятся формы сотрудничества, образы мышления, навыки, распределенные по различным ролям, имеющим отношение к архитектуре данных.

Эти три компонента являются важнейшими составляющими архитектуры данных.

Архитектура данных образует фундамент управления данными. Поскольку большинство организаций располагают объемами данных, которые не могут быть осмыслены отдельными сотрудниками, возникает насущная потребность в их представлении на разных уровнях абстракции таким образом, чтобы они были понимаемы и руководство могло принимать относительно этих данных соответствующие решения.

К артефактам архитектуры данных относятся спецификации, используемые для описания текущего состояния, определения требований к данным, порядка интеграции данных и контроля информационных активов в соответствии с действующей стратегией работы с данными. Архитектура данных организации описывается с помощью целостного комплекса проектных документов различной степени абстракции, включая стандарты, определяющие порядок сбора, хранения, упорядочения, использования и удаления данных. Она также делится на описания всех хранилищ данных и описания всех маршрутов их перемещения по информационным системам организации.

Наиболее детализированным архитектурным проектным документом в области данных является оформленная надлежащим образом корпоративная модель данных, включающая наименования элементов данных, подробные определения данных и метаданных, концептуальные и логические сущности и связи между ними, а также бизнес-правила. Наряду с другими документами в состав проектной документации входят физические модели данных, но только в качестве продуктов области моделирования и проектирования, а не области архитектуры данных.

Архитектура данных наиболее полезна в тех случаях, когда она в полном объеме обеспечивает потребности на корпоративном уровне. Единая корпоративная архитектура данных позволяет последовательно и согласованно осуществлять стандартизацию и интеграцию данных в масштабах всей организации.

Создаваемые архитекторами артефакты образуют чрезвычайно важные метаданные. В идеале хранение и управление всеми архитектурными артефактами должно осуществляться в корпоративном репозитории архитектурных артефактов.

Мы находимся в середине третьей волны цифровизации конечных потребителей. Первая волна перевела в электронную форму все банковские и финансовые транзакции; вторая — предоставление всевозможных услуг; третья же заключается в стремительном распространении интернета вещей и телематики. В рамках третьей волны на цифровые технологии переходят традиционные отрасли промышленности, такие как производство автомобилей, медицинских инструментов и промышленного оборудования.

Это происходит практически во всех отраслях. Для новых автомобилей Volvo предоставляется круглосуточный сервис, не только обеспечивающий решение технических проблем, но и подсказывающий местонахождение ближайших ресторанов и магазинов. Любая техника — от мостовых

кранов и штабелеукладчиков до анестезиологических аппаратов — непрерывно собирает и передает операционные данные, обеспечивающие сервисы поддержки ее работоспособности. Производители и поставщики всё больше предлагают не продажу оборудования, а его повременную аренду или плату за использование по мере надобности. Многие подобные компании, однако, не имеют достаточного опыта работы по таким схемам, поскольку раньше сбытом их продукции занимались предприятия розничной торговли, а послепродажным обслуживанием — специализированные сервисные центры.

Организации, которые смотрят в будущее, должны привлекать к разработке новых рыночных предложений профессионалов в области управления данными (например, корпоративных архитекторов данных или стратегических распорядителей данных), поскольку в наши дни такие предложения предполагают применение аппаратного и программного обеспечения, а также сервисов, обеспечивающих сбор и/или доступ к данным.

1.1 Бизнес-драйверы

Цель архитектуры данных — служить мостом между бизнес-стратегией и ее технологической реализацией. Будучи частью архитектуры предприятия, архитектура данных должна:

- стратегически подготавливать организации к быстрому развитию продуктов, услуг и данных с целью полного использования бизнес-возможностей, которые открываются вместе с появлением новых технологий;
- переводить бизнес-потребности на язык требований к данным и системам, с тем чтобы бизнес-процессы не испытывали дефицита в необходимой информации;
- обеспечивать управление сложным процессом предоставления данных и информации в масштабах предприятия;
- способствовать повышению согласованности между бизнес- и ИТ-процессами;
- служить средством гибкого проведения изменений и преобразований.

Именно эти бизнес-драйверы определяют всю меру ценности архитектуры данных.

Архитекторы данных создают и поддерживают знания организации о данных и системах, через которые эти данные распространяются. Такие знания позволяют организации управлять данными как активом и повышать получаемую от них пользу за счет выявления возможностей по их применению, а также снижения издержек и рисков.

1.2 Результаты и практики разработки архитектуры данных

Главными результатами разработки архитектуры данных являются:

- требования по хранению и обработке данных;
- проектные решения по структурам и планы, обеспечивающие выполнение текущих и долгосрочных требований организации в отношении данных.

Рисунок 21. Контекстная диаграмма: архитектура данных

Архитектуру стремятся сделать такой, чтобы она приносила организации ценность. Ценность же достигается за счет оптимизации требуемых ресурсов, операционной и проектной эффективности, а также расширения возможностей организации по использованию данных. Чтобы этого добиться, требуются качественное проектирование и планирование, равно как и способность обеспечить эффективную реализацию проектов и планов.

Для достижения этих целей архитекторы данных определяют и поддерживают спецификации, которые:

- определяют текущее состояние данных в организации;
- предоставляют стандартный бизнес-словарь для данных и компонентов;
- обеспечивают согласованность архитектуры данных с корпоративной стратегией и бизнес-архитектурой;
- отражают стратегические требования к данным;
- очерчивают высокоуровневые интегрированные проектные решения, призванные обеспечить выполнение этих требований;
- обеспечивают интеграцию разрабатываемых решений с дорожной картой реализации общей корпоративной архитектуры организации.

Практика разработки архитектуры данных в целом включает:

- использование артефактов архитектуры данных (основных рабочих описаний — master blueprints) для определения требований к данным, проведения интеграции данных, контроля информационных активов и согласования инвестиций в данные с бизнес-стратегией;
- сотрудничество с различными заинтересованными лицами, принимающими участие в развитии бизнеса или разработке информационно-технологических систем, получение от них знаний и оказание влияния;
- использование архитектуры данных для определения и закрепления семантики организации с помощью создания общего бизнес-словаря.

1.3 Основные понятия и концепции

1.3.1 Предметные области архитектуры предприятия

Архитектура данных выстраивается в контексте других предметных областей (доменов) архитектуры, включая бизнес, приложения и технологическую инфраструктуру. Таблица 6 содержит их сравнительное описание. Архитекторы, занимающиеся различными доменами, должны совместно определять направления и требования к разработке, согласовывая их между собой, поскольку каждый домен влияет и накладывает ограничения на другие (см. также рис. 22).

Таблица 6. Архитектурные домены

Домен	Бизнес-архитектура предприятия	Архитектура данных предприятия	Архитектура приложений предприятия	Технологическая архитектура предприятия
Назначение	Выявление путей создания предприятием ценности для потребителей и других заинтересованных лиц	Описание того, как данные должны быть организованы и как должно осуществляться управление данными	Описание структуры и функциональности используемых предприятием приложений	Описание технологической инфраструктуры, необходимой для обеспечения функционирования систем и предоставления с их помощью ценности
Элементы	Бизнес-модели, процессы, возможности, сервисы, события, стратегии, словарь	Модели данных, определения данных, спецификации отображения данных (data mapping), потоки данных, интерфейсы прикладного программирования (application programming interface, API), для работы со структурированными данными	Информационные системы, поддерживающие бизнес; пакеты прикладного ПО; базы данных	Технические платформы, сети, средства обеспечения безопасности и интеграции данных
Зависимости	Устанавливает требования для других доменов	Обеспечивает управление данными, создаваемыми и требуемыми согласно бизнес-архитектуре	Обеспечивает обработку данных в соответствии с бизнес-требованиями	Предоставление ресурсов для размещения решений, определенных архитектурой приложений
Роли	Бизнес-архитекторы и аналитики, распорядители бизнес-данных	Архитекторы данных и специалисты по разработке моделей данных, распорядители данных	Архитекторы приложений	Архитекторы инфраструктуры

1.3.2 Архитектурные рамочные структуры

Архитектурная рамочная структура является базовой структурой, используемой для разработки широкого спектра родственных архитектур. Такие структуры служат способом осмысления и понимания архитектуры. Они представляют собой своего рода общую «архитектуру архитектуры».

Компьютерное сообщество при Институте инженеров по электротехнике и электронике (IEEE Computer Society), разрабатывающее и сопровождающее уже упомянутый стандарт архитектуры ISO/IEC/IEEE 42010:2011, ведет детальную таблицу сравнения применяющихся архитектурных

рамочных структур[1]. Наиболее распространенные структуры и методики включают архитектуру данных в качестве одного из архитектурных доменов.

1.3.2.1 МОДЕЛЬ ЗАХМАНА

Самая известная рамочная структура архитектуры предприятия, «модель Захмана», была разработана Джоном Захманом[2] в 1980-х годах (см. рис. 22). С тех пор она продолжала развиваться. Захман обратил внимание на тот факт, что к построению (созданию) зданий, самолетов, предприятий, цепочек создания стоимости, проектов или систем имеют отношение различные группы людей, у каждой из которых есть собственная точка зрения (перспектива) на архитектуру создаваемого объекта. Эту концепцию он применил к требованиям для различных типов и уровней архитектуры предприятия.

Модель Захмана представляет собой онтологию — матрицу 6×6, охватывающую полный набор моделей, требуемых для описания предприятия, и связей между ними. Архитектурная рамочная структура Захмана не описывает, как именно создавать входящие в нее модели. Она просто показывает, что эти модели должны быть созданы.

	Что	Как	Где	Кто	Когда	Зачем	
Руководство	Идентификация объектов	Идентификация процессов	Идентификация местоположений	Идентификация обязанностей	Идентификация сроков	Идентификация мотивов	Бизнес-контекст
Бизнес-менеджмент	Определение объектов	Определение процессов	Определение местоположений	Определение обязанностей	Определение сроков	Определение мотивов	Бизнес-концепции
Архитекторы	Представление объектов	Представление процессов	Представление местоположений	Представление обязанностей	Представление сроков	Представление мотивов	Бизнес-логика
Инженеры	Спецификация объектов	Спецификация процессов	Спецификация местоположений	Спецификация обязанностей	Спецификация сроков	Спецификация мотивов	Физический уровень
Технические специалисты	Конфигурация объектов	Конфигурация процессов	Конфигурация местоположений	Конфигурация обязанностей	Конфигурация сроков	Конфигурация мотивов	Сборка
Предприятие	Реализация объектов	Реализация процессов	Реализация местоположений	Реализация обязанностей	Реализация сроков	Реализация мотивов	Реализация
	Наборы объектов	Потоки процессов	Расположение	Распределение обязанностей	Временные циклы	Мотивы	

Рисунок 22. Упрощенное представление модели Захмана

Столбцы матрицы отражают *обсуждаемые вопросы* (что, как, где, кто, когда и зачем), а строки — *преобразования в процессе материализации* (выявление — identification, определение — definition,

[1] http://www.iso-architecture.org/ieee-1471/afs/frameworks-table.html

[2] Джон Захман (*англ.* John A. Zachman, p. 1934) — американский специалист по информационным технологиям для бизнеса, разработавший описываемую модель в рамках концепции планирования бизнес-систем IBM, одним из авторов которой является. — *Примеч. пер.*

представление — representation, спецификация — specification, конфигурация — configuration, реализация — instantiation). В ячейках на пересечении строк и столбцов отражены уникальные типы артефактов архитектуры данных.

Обсуждаемые аспекты представляют собой основные вопросы, которые могут быть заданы относительно какой-либо сущности. Применительно к архитектуре предприятия столбцы матрицы можно интерпретировать следующим образом.

- **Что** (столбец объектов): сущности (объекты), используемые для построения архитектуры.
- **Как** (столбец процессов): проводимые работы.
- **Где** (столбец местоположений): местоположения бизнес-структур и технологических структур.
- **Кто** (столбец обязанностей): роли и организационные системы.
- **Когда** (столбец привязки по времени): сроки, интервалы, события, циклы, расписания.
- **Зачем** (столбец мотивации): цели, стратегии и средства.

Процесс материализации состоит из шагов, необходимых для перевода абстрактной идеи в конкретный образец (реализация). Эти шаги отражены в строках матрицы, обозначенных с помощью названий соответствующих ролей: планировщик, владелец, проектировщик, разработчик, внедренец, пользователь. Каждой из перечисленных ролей соответствует отличная от других перспектива процесса в целом, а также собственный круг решаемых проблем. Эта специфика и показана в строках. Например, каждая перспектива выражает различное отношение к столбцу «**Что**» (предметы или данные).

- **Перспектива руководства** (бизнес-контекст). Перечни составляющих бизнеса, определяющие содержание моделей идентификации.
- **Перспектива бизнес-менеджмента** (бизнес-концепции). Уточнение бизнес-менеджерами (как владельцами) связей между бизнес-концепциями и отражение их в моделях определения.
- **Перспектива архитектора** (бизнес-логика). Логические системные модели, детализирующие системные требования, и проектные решения без учета ограничений, отраженные архитекторами (как проектировщиками) в моделях представления.
- **Перспектива инженера** (физический уровень). Физические модели, оптимизирующие проектные решения с целью их реализации для конкретных применений с учетом ограничений по используемым технологиям, человеческим ресурсам, стоимости и срокам. Определяются инженерами (как разработчиками) в моделях спецификации.
- **Перспектива технического специалиста** (сборка). Чисто технический, без учета контекста взгляд на то, каким образом отдельные компоненты должны быть собраны и функционировать. Отражается техническими специалистами (как внедренцами) в конфигурационных моделях.
- **Перспектива пользователя** (реализация). Реальные функционирующие объекты, с которыми работают сотрудники (как пользователи). Эта перспектива моделей не предусматривает.

Как уже отмечалось, каждой ячейке, определяемой в результате пересечения строки и столбца, в модели Захмана соответствует уникальный тип разрабатываемого артефакта. Каждый такой артефакт описывает, каким образом соответствующая перспектива отвечает на основные вопросы, обсуждаемые в процессе создания архитектуры.

1.3.3 Корпоративная архитектура данных

Архитектура данных предприятия (*enterprise data architecture*), или *корпоративная архитектура данных*[1], определяет стандартные термины и проектные решения, применяемые в отношении важных для организации элементов. Проект корпоративной архитектуры данных включает описание бизнес-данных как таковых, а также описание порядка их сбора, хранения, интеграции, перемещения и распространения.

По мере поступления данных в организацию через каналы связи или интерфейсы, обеспечивается их защита и интеграция; они сохраняются, регистрируются, каталогизируются, распространяются, включаются в отчеты, анализируются и предоставляются заинтересованным лицам. Попутно данные могут подвергаться верификации, улучшению, связыванию, сертификации, агрегированию, анонимизации и использованию в целях аналитики вплоть до момента их архивации или удаления. Следовательно, описания корпоративной архитектуры данных должны включать как корпоративные модели данных (с указанием структуры и спецификаций данных), так и описания потоков данных.

- **Корпоративная модель данных (Enterprise Data Model, EDM).** EDM представляет собой целостную, не зависящую от технических средств реализации концептуальную или логическую модель данных, отражающую единый согласованный взгляд на данные в масштабах всей организации. Термин *корпоративная модель данных* обычно используется для обозначения высокоуровневой упрощенной модели данных, но уровень абстракции может быть различным в зависимости от целей ее представления. EDM включает данные о ключевых сущностях предприятия (на уровне бизнес-концепций — business concepts) и связях между ними, критически важные руководящие бизнес-правила и некоторые ключевые атрибуты. EDM закладывает на будущее основу для всех проектов в области данных или связанных с данными. Модели данных уровня отдельных проектов должны создаваться на основе EDM. EDM подлежит обязательной проверке всеми заинтересованными сторонами для обеспечения согласованного мнения о том, что в модели зафиксировано правильное представления об организации.
- **Описание потоков данных.** Содержит требования и основное рабочее описание (master blueprint) организации хранения и обработки данных в целом по всем базам данных, приложениям, платформам и сетям. Потоки данных отражают их перемещение с целью использования

[1] В данном издании применительно к терминам, начинающимся со слова «enterprise» в качестве определения (например, «enterprise architecture» или «enterprise data architecture»); это слово чаще всего переводится как «корпоративный» — «корпоративная архитектура», «корпоративная архитектура данных» и т. п. (широко распространенный вариант перевода в подобных случаях). Такой перевод облегчает восприятие текста в тех местах, где указанные термины применяются совместно с термином «organization» («организация»). — *Примеч. науч. ред.*

в бизнес-процессах, на отдельных рабочих местах, сотрудниками с определенными бизнес-ролями, а также отдельными техническими компонентами.

Эти два типа спецификаций должны быть между собой хорошо согласованы. Как уже отмечалось, и модель, и потоки данных должны быть отражены в трех состояниях — текущем, целевом (архитектурная перспектива) и переходном (проектная перспектива).

1.3.3.1 КОРПОРАТИВНАЯ МОДЕЛЬ ДАННЫХ

В одних организациях EDM создается как самостоятельный артефакт; в других под ней подразумевается совокупность моделей данных, отражающих различные перспективы с различными уровнями детализации, но в комплексе дающих полное и непротиворечивое представление об общепринятом в организации понимании описываемых сущностей, атрибутов и связей в масштабах предприятия. EDM включает как универсальные для всего предприятия модели (концептуальную и логическую модели корпоративного уровня), так и модели данных, используемые конкретными приложениями и/или проектами, а также определения, спецификации, отображения (mappings) данных и бизнес-правила.

Отправной точкой для разработки корпоративной модели данных может стать принятие организацией стандартной модели, применяемой в отрасли. Как правило, такие модели содержат полезные руководства и рекомендации. Однако, даже если организация начинает с покупки модели данных, выработка всех необходимых моделей в масштабах предприятия требует значительных вложений средств. Помимо прочего работа включает определение и документирование словаря организации, бизнес-правил и бизнес-знаний. Сопровождение и расширение EDM требует постоянной готовности тратить время и усилия.

Организация, осознавшая необходимость в EDM, должна определить, сколько времени и усилий она готова посвятить ее построению и ведению. EDM могут создаваться с различными уровнями детализации, а потому нужно изначально определиться с имеющимися ресурсами и исходя из этого спланировать объем работ по подготовке первоначального содержания модели. Со временем можно по мере надобности расширять объемы и прорабатывать дополнительные детали собираемых данных, необходимых для оптимальной работы предприятия, что, как правило, и делается. Самые успешные EDM выстраиваются поэтапно, итерационно и послойно. Рисунок 23 показывает, как связаны модели различных типов и как концептуальные модели могут быть привязаны к физическим моделям данных приложений. На рисунке отражены следующие уровни представления.

- Концептуальная общая модель данных, предоставляющая обзор всех предметных областей организации.
- Представления сущностей и связей по каждой предметной области.
- Детализированные, с частично описанными атрибутами логические представления тех же предметных областей.
- Логические и физические модели на уровне отдельных приложений или проектов.

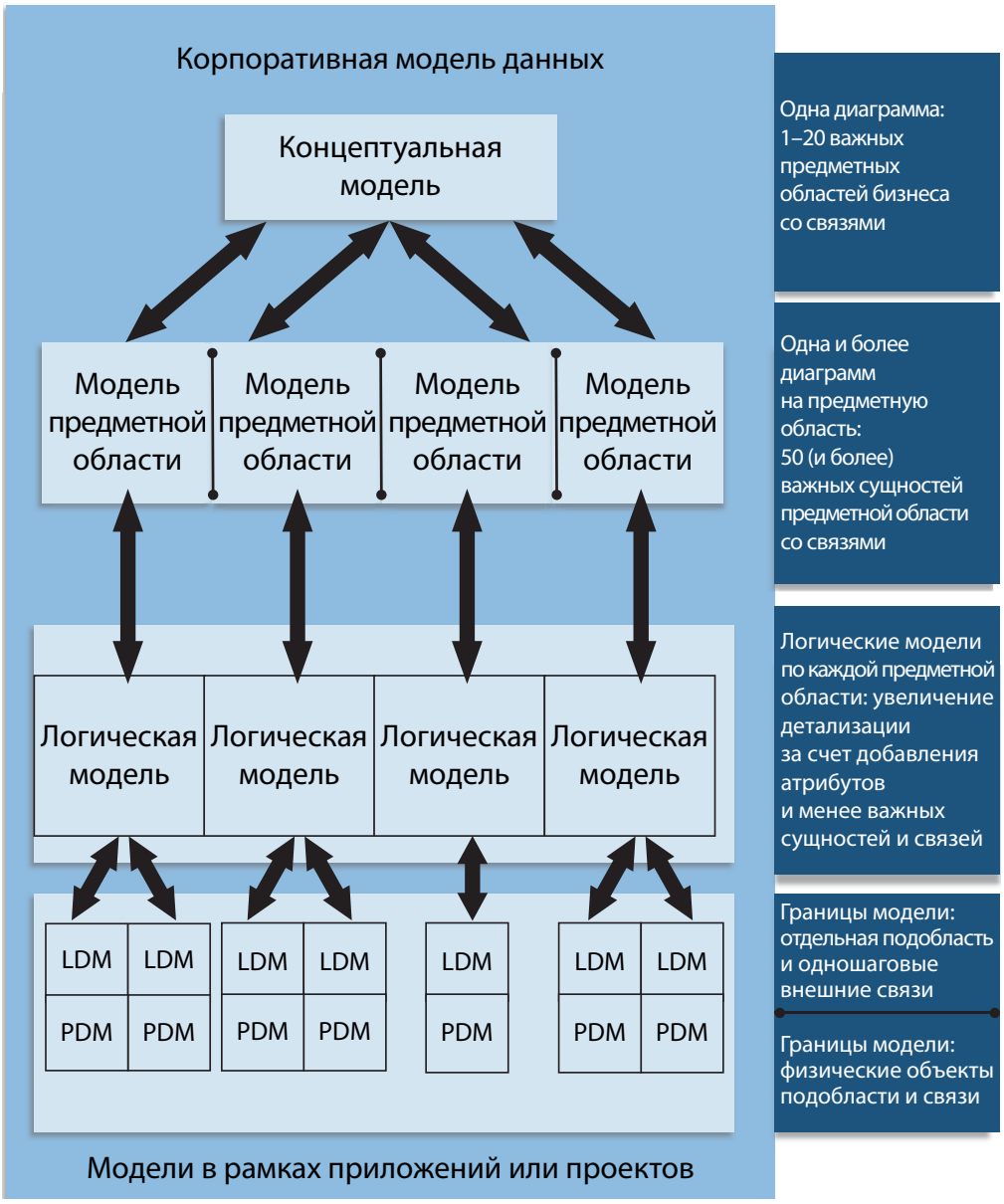


Рисунок 23. Корпоративная модель данных

Все уровни в совокупности составляют корпоративную модель данных. Структура связей позволяет проследить сущность с верхнего уровня до нижнего и между моделями на одном уровне.

- **Вертикальные связи.** Модели каждого уровня отображаются на модели других уровней. Например, таблица (или файл) с данными о мобильном устройстве MobileDevice в физической модели на уровне проекта может быть связана с сущностью MobileDevice логической модели проекта, сущностью MobileDevice предметной области Product (Продукт) корпоративной

логической модели, концептуальной сущностью Product модели предметной области Product и, наконец, с сущностью Product корпоративной концептуальной модели.

- **Горизонтальные связи.** Сущности и связи могут появляться во многих моделях одного уровня; сущности в логических моделях одной тематики могут быть связаны с сущностями другой тематики, помеченными или описанными как внешние по отношению к предметной области на графическом изображении модели. Сущность Product Part (Комплектующий элемент продукта) может фигурировать как в моделях, описывающих предметную область Product, так и в моделях, относящихся к другим областям: например, Sales (Продажи), Inventory (Запасы), Marketing (Маркетинг), — за счет включения ее в эти модели с помощью внешних ссылок.

Разработка EDM на всех уровнях ведется с использованием стандартных методов моделирования данных (см. главу 5).

На рисунке 24 представлен упрощенный пример диаграмм, описывающих три предметные области. Каждая диаграмма содержит концептуальную модель данных с набором сущностей. Связи могут пересекать границы между предметными областями; каждая сущность в корпоративной модели данных должна принадлежать строго одной предметной области, но может быть связана с сущностями из любой другой предметной области.

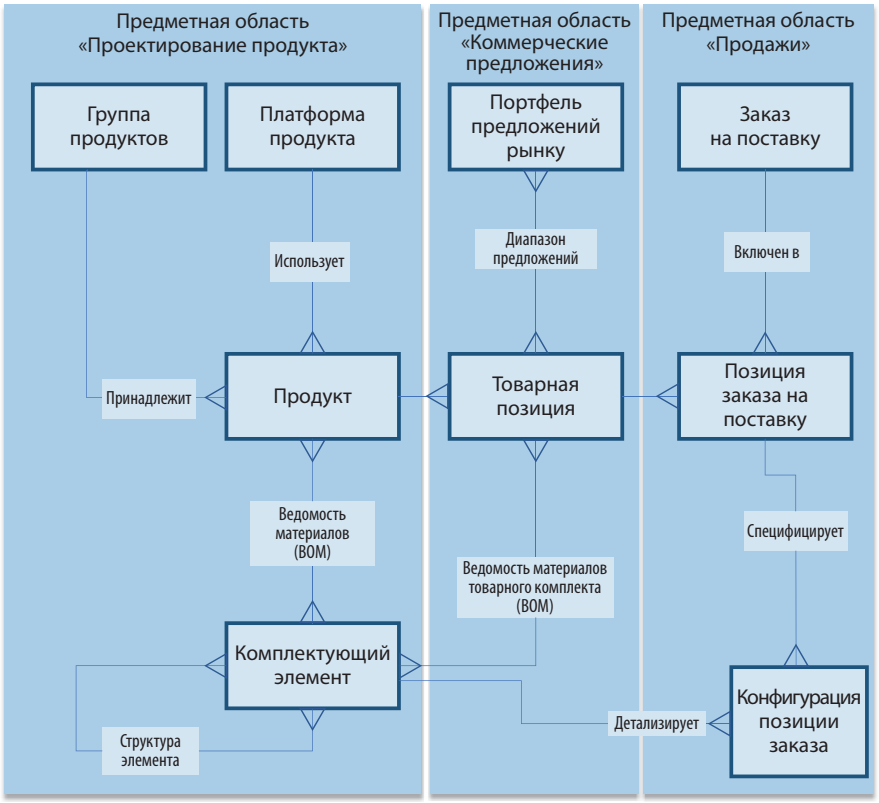


Рисунок 24. Примеры диаграмм, отражающих модели предметных областей

Таким образом, корпоративная концептуальная модель данных создается путем объединения моделей предметных областей. При этом EDM может выстраиваться как сверху вниз, так и снизу вверх. Первый подход подразумевает сначала определение предметных областей, а лишь затем заполнение их моделями. При втором подходе структура предметных областей определяется существующими моделями данных. На практике обычно рекомендуется использовать разумное сочетание обоих подходов: сначала обобщаются в предметные области имеющиеся и используемые предприятием модели (подход «сверху вниз»), а последующее дополнение EDM новыми моделями производится посредством делегирования функции моделирования предметных областей проектным командам отдельных проектов (подход «снизу вверх»).

Дискриминаторы предметной области (то есть принципы, определяющие ее уникальную структуру) должны быть одними и теми же в рамках всей EDM. Часто используются, в частности, следующие принципы: использование правил нормализации данных; отделение предметных областей от портфелей проектов систем (то есть финансирования); формирование предметных областей исходя из структуры руководства данными и распределения полномочий владения (организационный принцип); использование процессов верхнего уровня (основанных на цепочках создания стоимости); разделение по бизнес-возможностям (на основе архитектуры предприятия). Структура предметной области обычно наиболее эффективна с точки зрения построения архитектуры данных, если она сформирована с использованием правил нормализации. В процессе нормализации устанавливаются основные сущности, определяющие и составляющие каждую предметную область.

1.3.3.2 ОПИСАНИЕ ПОТОКОВ ДАННЫХ

Потоки данных являются одним из способов документального оформления происхождения (lineage) данных. Они фиксируют маршруты прохождения данных через бизнес-процессы и системы. Описанный от начала до конца поток данных показывает, где данные возникают, где хранятся и используются, а также все преобразования данных в процессе их движения как внутри, так и между различными процессами и системами. Анализ происхождения помогает объяснить состояние данных в каждой точке потока.

Потоки данных отображают и документируют взаимосвязи между данными и:

- приложениями, используемыми в рамках бизнес-процесса;
- хранилищами или базами данных в среде функционирования;
- сегментами сети (полезно для описания мер безопасности);
- бизнес-ролями — показывая, какие роли отвечают за создание, чтение, обновление, удаление данных (CRUD[1] — Create, Read, Update, Delete);
- местами, в которых происходят изменения данных.

[1] CRUD — часто используемый акроним, обозначающий четыре базовые операции, выполняемые при работе с данными: создание (create), чтение (read), обновление (update), удаление (delete). — *Примеч. науч. ред.*

Потоки данных могут документироваться с разной степенью детализации — до уровня предметной области, сущности или даже атрибута. Системы могут быть представлены сегментами сети, платформами, наборами часто используемых приложений или отдельными серверами. Для схематического представления потоков данных могут использоваться двумерные матрицы (рис. 25) или диаграммы потоков данных (рис. 26).

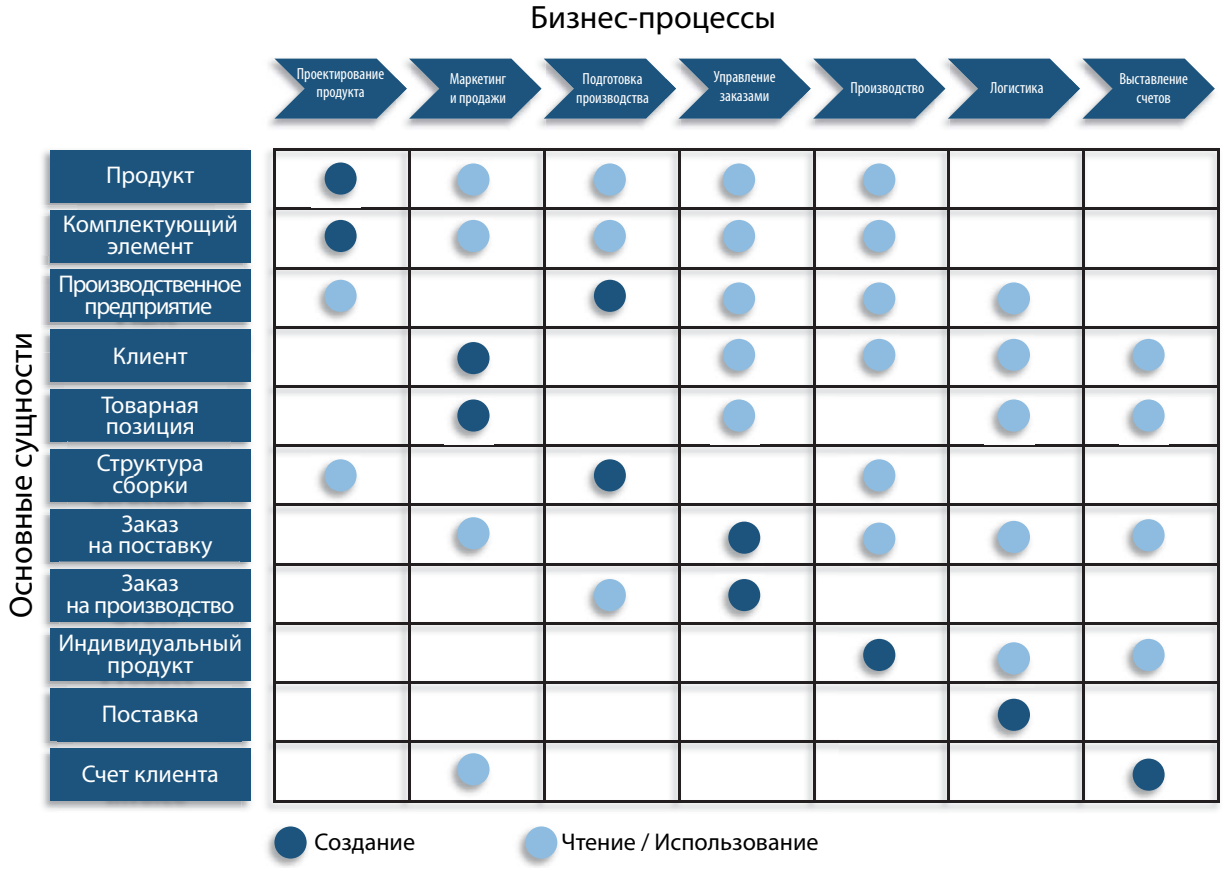


Рисунок 25. Поток данных, представленный в виде матрицы

Матричное представление наглядно показывает, в каких процессах создаются и используются данные каждой категории. Преимущество отображения потребностей в данных в виде матрицы заключается в следующем: оно учитывает, что потоки данных совсем не обязательно движутся только в одном направлении; обмен данными между процессами происходит в соответствии с типом связи «многие ко многим», иногда по весьма сложным схемам, при которых любые данные могут возникать то там, то здесь. К тому же матричный формат удобен для определения ответственных за получение различных данных и обусловленных данными зависимостей процессов друг от друга, что, в свою очередь, помогает лучше документировать каждый процесс. Те, кто предпочитает описывать деятельность организации в терминах бизнес-возможностей, могут

отображать потоки данных в аналогичном формате, просто заменив «процессы» на «возможности» в столбцах матрицы и в названии оси. Построение таких матриц — давно сложившаяся практика моделирования работы предприятия. Разработаны они были еще IBM в рамках методологии планирования бизнес-систем (Business Systems Planning, BSP), а популяризованы в 1980-е годы Джеймсом Мартином[1] в его методологии планирования информационных систем (Information Systems Planning, ISP).

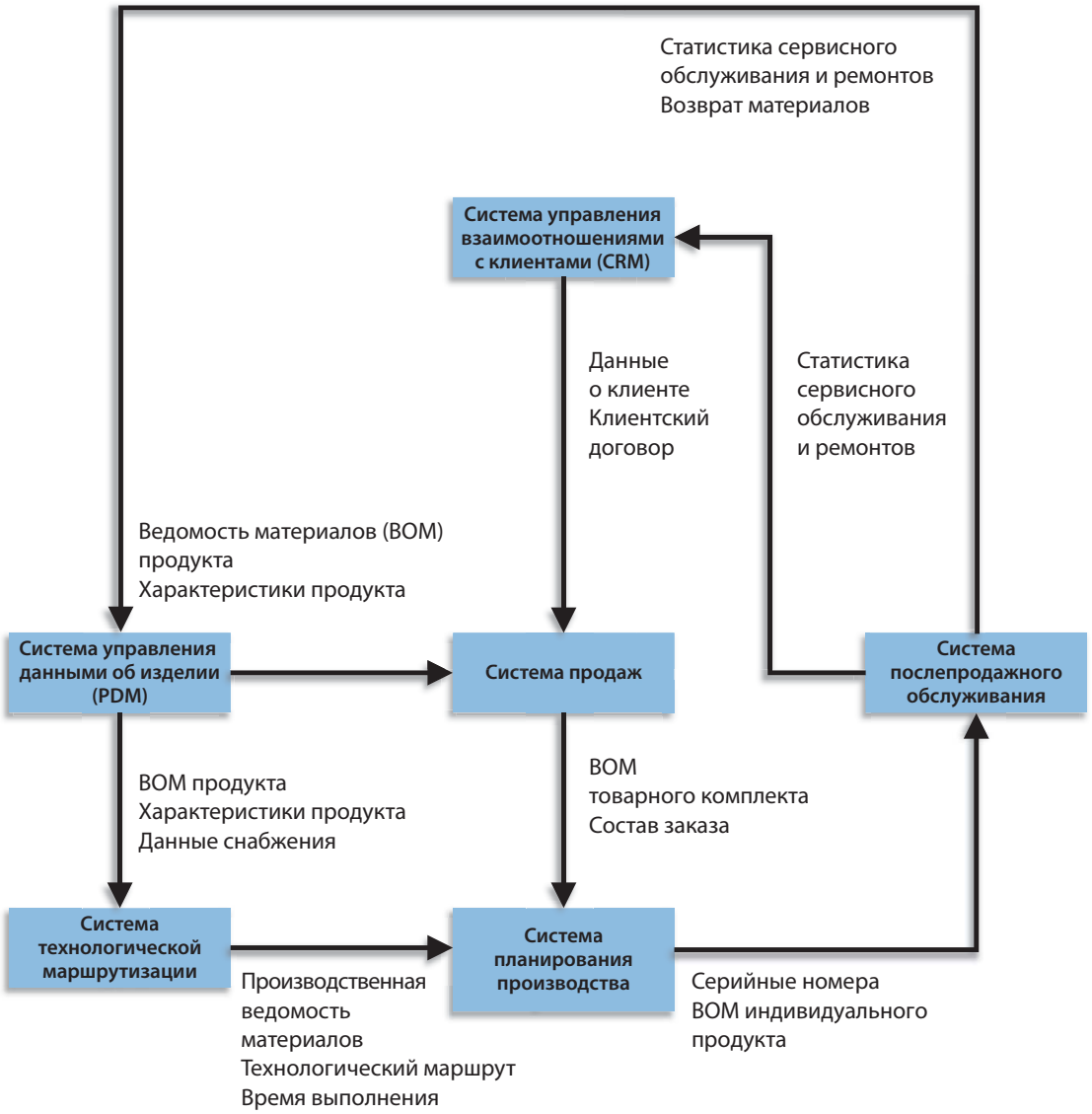


Рисунок 26. Пример диаграммы потоков данных

[1] Джеймс Мартин (*англ.* James Martin, 1933–2013) — английский специалист по информационным и компьютерным системам, основоположник концепции автоматизации разработки программного обеспечения (CASE), автор более сотни книг, в 1959–1980 годах работавший в IBM. — *Примеч. пер.*

Рисунок 26 представляет пример традиционной высокоуровневой диаграммы потоков данных между системами с краткими описаниями видов перемещающихся данных. Подобные диаграммы могут быть построены в различных форматах и описывать движение данных на различных уровнях детализации.

2. ПРОВОДИМЫЕ РАБОТЫ

Создание архитектуры данных (и архитектуры предприятия) сопряжено с необходимостью учета сложного комплекса вопросов, обусловленных следующими двумя основными точками зрения на архитектурные решения.

- **Ориентированная на качество.** Основное внимание направлено на совершенствование деятельности в рамках бизнес-цикла и цикла разработки в области ИТ. Без должного управления архитектурой качество архитектурных решений ухудшается. Системы со временем будут чрезмерно усложняться и утрачивать гибкость, а это создает дополнительные риски для организации. Неконтролируемое распространение и копирование данных, запутанные взаимосвязи делают организации менее эффективными и снижают доверие к данным.
- **Ориентированная на инновации.** Основное внимание направлено на трансформацию бизнеса и ИТ в свете новых ожиданий и перспектив. Продвигать инновации за счет внедрения прорывных технологий и методов использования данных — еще одна задача современного корпоративного архитектора.

К двум этим драйверам нужно подходить дифференцированно. Подход, ориентированный на качество, укладывается в традиционное представление о работе по проектированию архитектуры, предусматривающее ее поэтапное совершенствование. Архитектурные задачи распределяются по проектам, в которых принимают участие архитекторы (или соответствующая работа выполняется с помощью делегирования). В любом случае архитектор, как правило, не теряет из виду цельность архитектуры и руководствуется долгосрочными установками, связанными с руководством данными, стандартизацией и структурированной разработкой. При подходе же, ориентированном на инновации, может рассматриваться краткосрочная перспектива, а также предполагается использование еще не апробированных схем ведения бизнеса и передовых технологий. Такая направленность часто требует, чтобы архитекторы вступали в контакт с теми людьми внутри организации, которые обычно не входят в круг общения профессионалов в сфере ИТ (например, с представителями подразделений, отвечающих за разработку новых продуктов или бизнес-моделей).

2.1 Внедрение практики разработки и сопровождения архитектуры данных

В идеале архитектура данных должна быть неотъемлемой частью архитектуры предприятия. Но если в организации не внедрена функция поддержки корпоративной архитектуры, она тем

не менее может создать команду по поддержке архитектуры данных. В таком случае она должна определить и принять концептуальную рамочную структуру, которая поможет четко сформулировать цели и драйверы поддержки архитектуры данных. Эти драйверы в дальнейшем будут влиять на подход, содержание и приоритеты, отражаемые в дорожной карте.

Рамочную структуру следует выбирать согласно роду деятельности (то есть в госсекторе должна использоваться модель, предназначенная для государственных организаций; в бизнесе — для коммерческих, и т. п.). Представления и таксономия, определяемые рамочной структурой, должны быть полезными и понятными с точки зрения различных заинтересованных сторон. В рамках инициатив по созданию архитектуры данных это вдвойне важно, поскольку именно на данном этапе определяется системная и бизнес-терминология. Архитектура данных находится в тесной и неразрывной связи с архитектурой бизнеса.

Практика разработки и сопровождения корпоративной архитектуры данных обычно включает проведение работ (последовательное или параллельное) по следующим направлениям.

- **Стратегия.** Выбор рамочных моделей, формулировка подходов, разработка дорожных карт.
- **Признание и культура.** Информирование и мотивирование к изменениям поведения.
- **Организация.** Организация деятельности в области архитектуры данных посредством распределения обязанностей и внедрения механизма подотчетности.
- **Рабочие методы.** Определение лучших практик и проведение работ в области архитектуры данных в рамках проектов организации, связанных с разработкой, с учетом координации с работами в области корпоративной архитектуры.
- **Результаты.** Предоставление артефактов архитектуры данных в соответствии с дорожной картой.

Кроме того, корпоративная архитектура данных влияет на содержание и границы проектов, а также на разрабатываемые системы.

- **Определение проектных требований в области данных.** Архитектура данных накладывает определенные требования в части корпоративных данных по каждому отдельному проекту.
- **Проверка проектных решений в области данных.** Проведение анализа проектных решений позволяет убедиться в том, что концептуальные, логические и физические модели данных проекта согласуются с архитектурой и обеспечат поддержку долгосрочной стратегии развития организации.
- **Определение и учет факторов, оказывающих влияние на происхождение данных.** Гарантирует, что бизнес-правила в приложениях, задействованных по всему потоку данных, являются согласующимися и прослеживаемыми.
- **Контроль репликации данных.** Репликация является распространенным способом повышения производительности приложения и облегчения доступа к данным, но может привести к их несогласованности. Руководство архитектурой данных дает уверенность в наличии

достаточного контроля репликации (методов и механизмов), поддерживающего требуемый уровень согласованности. (При этом следует заметить, что согласованность данных требуется не для всех приложений.)

- **Обеспечение соблюдения стандартов архитектуры данных.** Разработка и обеспечение соблюдения стандартов в отношении жизненного цикла архитектуры данных. Стандарты могут быть сформулированы в виде принципов, процедур и руководств, а также представлены в формате рабочих документов, описывающих ожидания по обеспечению соответствия.
- **Стимулирование использования новейших технологий работы с данными и решений по проведению обновлений.** Архитектура данных совместно с архитектурой предприятия обеспечивают предоставление возможностей по регулярному обновлению применяемых технологий работы с данными. Для этого предусматриваются механизмы управления версиями, пакетами обновлений и политиками, которые использует каждое приложение, а также дорожная карта внедрения новых технологий.

2.1.1 Оценка существующих спецификаций архитектуры данных

В любой организации имеется представленная в том или ином виде документация на существующие системы. Необходимо идентифицировать эти документы и оценить их на предмет точности, полноты и уровня детализации, а при необходимости доработать и привести в соответствие с текущим состоянием.

2.1.2 Разработка дорожной карты

Если бы организация создавалась с чистого листа (вне всякой привязки к существующим процессам), оптимальная архитектура должна была бы основываться исключительно на данных, требуемых для эффективной работы подразделений; приоритеты определялись бы исключительно бизнес-стратегией, а решения принимались без учета предыдущей деятельности. В реальных организациях подобные случаи практически не встречаются. Даже в идеальной ситуации всевозможные зависимости от данных, которые нужно учитывать и контролировать, развиваются очень стремительно. Дорожная карта служит хорошим средством управления такими зависимостями и принятия решений, нацеленных на перспективу. Она помогает организации находить возможности для компромиссов и выстраивать прагматичный план, сбалансированно учитывающий потребности и возможности бизнеса, внешние требования и имеющиеся ресурсы.

Дорожная карта реализации корпоративной архитектуры данных разрабатывается на среднесрочную перспективу от трех до пяти лет. Вместе с бизнес-требованиями, результатами анализа фактических условий и техническими оценками дорожная карта описывает, каким образом целевая архитектура превратится в реальность. Дорожная карта развития корпоративной архитектуры данных должна быть интегрирована с общей дорожной картой реализации корпоративной архитектуры, которая включает высокоуровневые вехи, потребности в ресурсах, финансовые оценки, разбитые по направлениям развития бизнес-возможностей. Дорожная карта должна строиться с учетом оценки зрелости управления данными (см. главу 15.)

Большинство бизнес-возможностей требуют данных в качестве входных ресурсов; какие-то из них также производят данные, от которых зависят остальные бизнес-возможности. Корпоративная архитектура и корпоративная архитектура данных могут быть взаимоувязаны путем преобразования этого потока в цепочку зависимостей между бизнес-возможностями.

Управляемая на основе бизнес-данных (business-data-driven) дорожная карта начинается с мероприятий, относящихся к наиболее независимым бизнес-возможностям (не зависящим от результатов другой деятельности), и заканчивается самыми зависимыми направлениями работ. Последовательность проработки бизнес-возможностей будет соответствовать общему порядку возникновения бизнес-данных. Рисунок 27 содержит пример такой цепочки зависимостей с наименьшей зависимостью в верхней части. Управление продуктом и Управление клиентами не требуют каких-либо данных и, таким образом, производят основные данные. Самые зависимые направления отражены в нижней части. Здесь Управление счетами клиентов зависит от Управления клиентами и Управления заказами на продажу, которое, в свою очередь, также зависит от двух направлений.

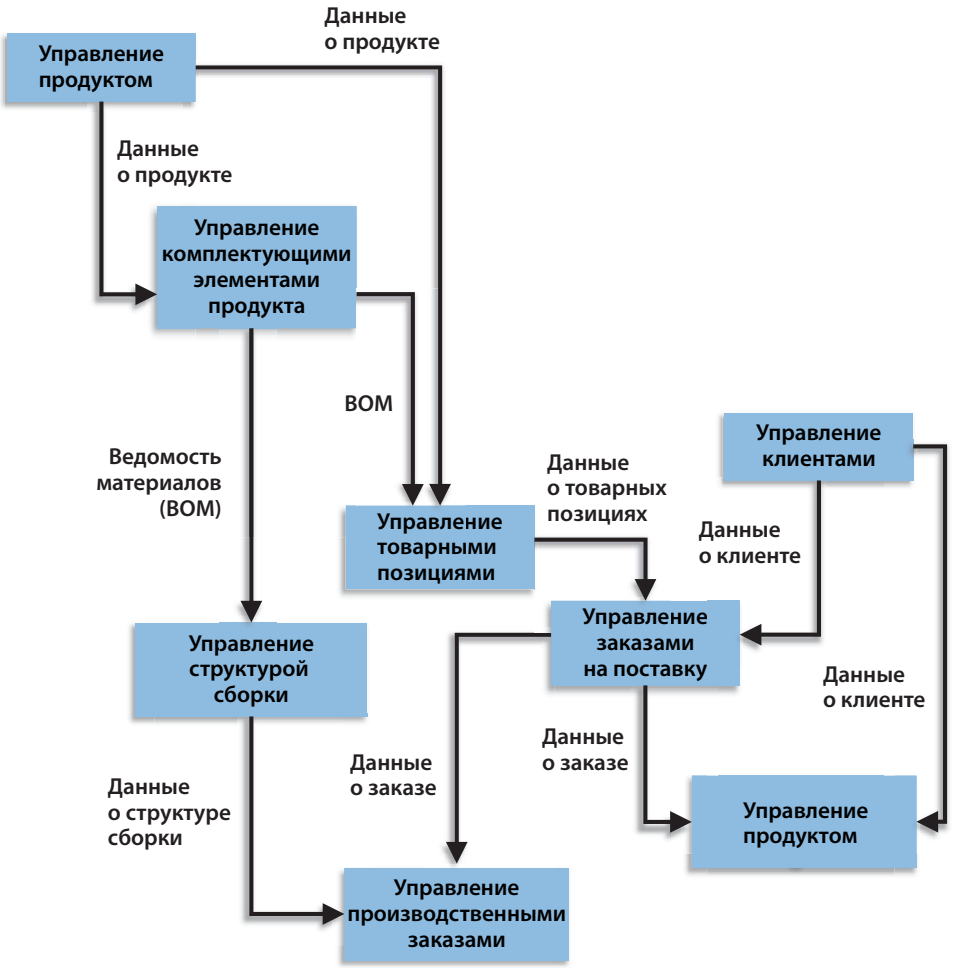


Рисунок 27. Зависимости бизнес-возможностей в отношении данных

Таким образом, в идеале следует начинать разработку дорожной карты с возможностей по Управлению продуктом и Управлению клиентами, а затем пошагово осуществлять разрешение каждой зависимости, спускаясь по цепочке сверху вниз.

2.1.3 Управление корпоративными требованиями в рамках проектов

Архитектура не должна замыкаться в границах, существовавших на момент ее разработки. Модели данных и другие спецификации, описывающие архитектуру данных, должны быть достаточно гибкими для того, чтобы адаптироваться к будущим требованиям. Модель данных уровня архитектуры должна обеспечивать единое глобальное представление об организации вместе с ясными определениями, понятными всем сотрудникам.

Проекты разработки (связанные с разработкой и внедрением) реализуют решения по сбору, хранению и распространению данных, которые основаны на бизнес-требованиях и стандартах, установленных корпоративной архитектурой данных. Этот процесс по самой своей природе носит пошаговый характер.

На уровне отдельного проекта процесс определения требований посредством моделирования данных начинается с анализа потребностей бизнеса. Часто эти потребности касаются исключительно специфических целей проекта и не имеют отношения к организации в целом. Процесс определения требований, кроме того, должен включать выработку определений терминов, а также другие работы, поддерживающие использование данных.

Важно, чтобы архитекторы данных были способны воспринимать требования с учетом общей архитектуры. По завершении работы над проектной спецификацией архитекторы данных должны определить:

- соответствуют ли корпоративные сущности, указанные в спецификации, согласованным стандартам;
- какие сущности из спецификации следует включить в общую корпоративную архитектуру данных;
- нет ли необходимости обобщить или доработать представленные в спецификации сущности и определения (с прицелом на будущие потребности, исходя из наблюдаемых тенденций);
- нет ли необходимости в новых архитектурных решениях по предоставлению данных, или разработчиков следует нацелить на использование уже имеющихся схем.

Организации часто откладывают решение насущных вопросов, касающихся архитектуры данных, до момента, пока в проектах не возникает необходимость разработки хранилищ данных и их интеграции. Однако предпочтительнее всё же проводить рассмотрение таких вопросов, начиная уже с ранней стадии планирования и затем на протяжении всего цикла жизни проекта.

Работы в рамках проектов, затрагивающих корпоративную архитектуру данных, включают следующее.

- **Определение содержания и границ проекта.** Необходимо убедиться, что содержание, границы и взаимосвязи соответствуют корпоративной модели данных. Кроме того, следует понять потенциальный вклад проекта в общую корпоративную архитектуру данных, в частности ответив на вопросы, что именно будет смоделировано и спроектировано и какие из имеющихся компонентов должны (или могут) быть повторно использованы в проекте. В тех областях, которые предстоит разработать, проект должен предусматривать определение новых зависимостей от заинтересованных сторон, которые не вошли в его границы (например, зависимостей в рамках нисходящих процессов). Информационные артефакты, которые в проекте будут определены как подлежащие совместному или повторному использованию, должны быть включены в корпоративную логическую модель данных и размещены в назначенных для них репозиториях.
- **Углубление понимания бизнес-требований.** Сбор связанных с данными требований (например, в отношении сущностей, источников, доступности, качества, уязвимых мест), а также оценка экономической целесообразности и реальной ценности выполнения этих требований для бизнеса. Также сюда относится оценка выгод для бизнеса от выполнения этих требований.
- **Проектирование.** Составление детализированных целевых спецификаций, включая описание бизнес-правил, применяемых на протяжении жизненного цикла данных. Утверждение конечных результатов, а также, при необходимости, определение и передача в заинтересованные подразделения требований по дополнению и расширению имеющихся стандартизованных моделей. Корпоративная логическая модель данных и корпоративный архитектурный репозиторий — подходящие места для поиска архитекторами данных проекта пригодных к повторному использованию структурных компонентов, к которым можно обеспечить общий доступ в масштабах организации. Кроме того, необходимо регулярно отслеживать и применять на практике технические стандарты в области данных.
- **Реализация.**
  - **При покупке** готовых коммерческих приложений (Commercial Off The Shelf, COTS) следует проводить их реверс-инжиниринг с целью сравнения с имеющейся структурой данных. Нужно выявлять и документировать пробелы (gaps) и расхождения в структурах, определениях и правилах. В идеале поставщики должны предоставлять модели данных к продаваемым продуктам; на практике многие этого не делают, поскольку рассматривают модели как свою собственность. Если это возможно, желательно купить такую модель с детальными определениями.
  - **При повторном использовании данных** необходимо сопоставить модели данных приложения с общеприменимыми в организации структурами данных, а также существующими и вновь внедряемыми процессами, чтобы понять, каким образом должны быть реализованы операции создания, чтения, обновления и удаления данных (CRUD). Следует добиваться использования систем записи или других надежных источников данных. Также необходимо выявлять и документировать пробелы и несоответствия.

- **При окончательной сборке** необходимо организовывать хранение данных в соответствии с их структурой. Проводить интеграцию нужно в строгом соответствии со стандартизованными или разработанными спецификациями (см. главу 8).

Роль корпоративных архитекторов данных в конкретных проектах, а также процесс выстраивания в проекте работы по проектированию архитектуры зависят от методологии разработки.

- **Каскадные (waterfall) методы.** Изучение требований и конструирование систем поэтапно и последовательно, в рамках общего процесса корпоративного проектирования. Такая методология предусматривает наличие контрольных точек принятия решения о переходе к следующему этапу (tollgates), обеспечивающих управление ходом изменений. Включить работы по проектированию архитектуры данных в подобные модели особого труда не составляет. Главное при таком подходе — постоянно помнить о том, что проектирование ведется в масштабах всего предприятия.
- **Методы приращений (incremental).** Изучение потребностей и конструирование решений в виде последовательности небольших шагов (мини-каскадов — mini-waterfalls). В рамках такого подхода создаются прототипы на основе достаточно расплывчатых общих требований. Ключевой является фаза инициации; лучше уже на ранних шагах разработать полную и подробную концепцию архитектуры данных.
- **Гибкие, итеративные методы (подход Agile).** Проведение изучения, проектирования, тестирования и выпуска отдельных пакетов обновлений короткими итерациями (называемых «спринтами» — sprints). Подход Agile хорош тем, что, если потребуется отказаться от результатов отдельно взятого неудачного микропроекта, потери не будут слишком серьезными. Agile-методы (Scrum, быстрая разработка — Rapid Development и унифицированный процесс — Unified Process) способствуют продвижению объектно-ориентрованного моделирования, которое придает особое значение дизайну пользовательского интерфейса и программного обеспечения, а также поведению систем. Эти методы следует дополнять спецификациями моделей данных и операций по их сбору, хранению и распространению. Опыт применения методологии DevOps[1] (недавно появившаяся и сразу же завоевавшая популярность разновидность Agile-подхода) свидетельствует о повышении эффективности решений в области проектирования данных и программного обеспечения при условии тесного взаимодействия между программистами и архитекторами данных, а также соблюдения теми и другими стандартов и руководств.

2.2 Интеграция с корпоративной архитектурой

Работы по созданию спецификаций корпоративной архитектуры данных, начиная от уровня предметной области до более детальных, а также во взаимосвязи с другими архитектурными

[1] DevOps (от *англ.* development и operations (разработка и операции); по-русски обычно произносится как «девóпс») — набор практик, нацеленных на активное взаимодействие специалистов по разработке со специалистами по информационно-технологическому обслуживанию и на взаимную интеграцию их рабочих процессов друг в друга. — *Примеч. науч. ред.*

доменами, как правило, выполняются в рамках плановых проектов, предусмотренных бюджетом организации. Плановые проекты обычно и оказывают определяющее влияние на архитектурные приоритеты. Несмотря на это, к вопросам архитектуры данных масштаба организации следует подходить проактивно. В действительности архитектура данных также может влиять на содержание и границы проектов. Следовательно, лучше интегрировать рассмотрение проблем корпоративной архитектуры данных с деятельностью по управлению портфелем проектов. Это обеспечит успешное выполнение мероприятий дорожной карты и внесет вклад в повышение качества получаемых результатов.

По аналогичным соображениям работы в области корпоративной архитектуры данных должны учитываться при планировании разработки и интеграции приложений. Целевой ландшафт приложений и дорожную карту продвижения к этому ландшафту обязательно следует рассматривать с точки зрения архитектуры данных.

3. ИНСТРУМЕНТЫ

3.1 Инструменты моделирования данных

Инструменты моделирования данных и репозитории моделей необходимы для управления корпоративной моделью данных на всех уровнях. Большинство инструментов моделирования данных включают функции прослеживания происхождения данных и связей между ними, что позволяет архитекторам управлять комплексами взаимосвязанных моделей, созданных с различными целями и на разных уровнях абстракции (см. главу 5).

3.2 Программное обеспечение для управления ИТ-активами

Программное обеспечение для управления активами используется в целях учета систем, описания их составных частей и отслеживания связей между системами. Среди прочего такие средства позволяют организации обеспечивать соблюдение требований лицензионных соглашений в отношении программного обеспечения и собирать данные, связанные с активами, которые могут быть использованы для минимизации затрат и оптимизации требуемых ИТ-ресурсов. Поскольку с помощью подобных инструментов составляется инвентарная опись ИТ-активов, параллельно они накапливают ценные метаданные о системах и содержащихся в них данных. Эти метаданные очень полезны при проектировании потоков данных или исследовании их текущего состояния.

3.3 Приложения для графического проектирования

Приложения для графического проектирования используются для построения архитектурных диаграмм моделей данных, потоков данных, цепочек создания стоимости данных и прочих архитектурных артефактов.

4. МЕТОДЫ

4.1 Проекции на фазы жизненного цикла

Архитектурные решения могут создаваться на будущее, а также могут быть уже внедренными и действующими или планируемыми к выводу из эксплуатации. Следует четко документировать, к каким фазам и аспектам жизненного цикла относятся конкретные продукты архитектурного проектирования. Например:

- **текущий период** — поддерживаемые и активно используемые в текущий момент продукты;
- **период развертывания** — продукты, которые будут введены в действие и готовы к использованию в ближайшие один-два года;
- **стратегический период** — продукты, которые будут введены в действие и готовы к использованию не ранее чем через два года;
- **отменяемые —** продукты, от использования которых организация решила отказаться или планирует отказаться в течение года;
- **предпочтительные** — продукты, преимущественно используемые большинством приложений;
- **ограниченного использования** — продукты, используемые лишь отдельными приложениями;
- **перспективные** — продукты, эксплуатируемые в экспериментальном режиме и исследуемые на предмет оценки целесообразности их внедрения в будущем;
- **на рассмотрении** — продукты, прошедшие оценку, с указанием результатов оценки, по которым пока что не принято решение о присвоении им одного из вышеперечисленных статусов.

Подробнее об управлении технологиями работы с данными см. главу 6.

4.2 Четкость и ясность графических представлений

Описания моделей и диаграммы должны отображать информацию в соответствии с принятым набором соглашений о визуальном представлении. Соглашения должны последовательно применяться во избежание неправильного толкования представленной информации или ее искажения в документе. Свойства графического архитектурного документа, обеспечивающие сведение к минимуму отвлекающих моментов и отражение максимума полезной информации, включают следующее.

- **Четкая и последовательная легенда.** Легенда должна определять все объекты и линии с расшифровкой их значения. На всех диаграммах легенда обязана располагаться в одном и том же месте.
- **Соответствие между всеми объектами диаграммы и легендой.** При использовании шаблонов с готовыми легендами на диаграмме могут отсутствовать объекты некоторых типов, представленных в легенде, но она не должна содержать объектов, типы которых в легенде не описаны.

- **Четкое и последовательное отображение направления линий.** Все потоки должны всегда начинаться с одной стороны или из одного угла (обычно слева) и продвигаться к противоположной стороне или углу, насколько это возможно. В случае петель и циклов линии со стрелками, ведущие в обратном по отношению к общему потоку направлении, следует пускать в обход основной части схемы, чтобы они четко выделялись.
- **Единообразное отображение пересечений линий.** Линии могут пересекаться, но эти пересечения графически должны четко отличаться от точек, в которых линии соединяются. При проведении одной линии над другой необходимо использовать специальные обозначения пересечения. Нельзя допускать слияния расположенных рядом линий. Следует сводить к минимуму количество линий, которые пересекаются.
- **Единообразное отображение объектов каждого типа.** Любые различия в размерах, цветах, толщине линий и т. п. должны о чем-то сигнализировать, иначе они будут лишь отвлекающим фактором.
- **Выравнивание и симметрия.** Диаграммы с объектами, упорядоченными по вертикали и горизонтали в четкие столбцы и ряды, читаются лучше диаграмм с хаотично разбросанными объектами. Далеко не всегда можно обеспечить строгое выравнивание всех объектов, но, если удастся выровнять хотя бы половину (по вертикали и/или горизонтали), диаграмма будет восприниматься гораздо лучше.

5. РЕКОМЕНДАЦИИ ПО ВНЕДРЕНИЮ

Как отмечалось во вводной части главы, важнейшие составляющие процесса создания архитектуры данных — разработка артефактов, проведение работ и формирование поведения. Следовательно, внедрение корпоративной архитектуры данных предполагает:

- организацию команд и форумов по корпоративной архитектуре данных;
- создание исходных версий артефактов архитектуры данных, таких как корпоративная модель данных, описание потоков данных в масштабах предприятия и дорожная карта внедрения;
- формирование и внедрение архитектурного подхода в области данных в практику выполнения проектов, связанных с разработкой;
- повышение уровня информированности организации и развитие общего понимания ценности усилий по созданию архитектуры данных.

Внедрение архитектуры данных требует осуществления деятельности не менее чем по двум из вышеперечисленных направлений, поскольку успех достигается только при совместном (или хотя бы параллельном) проведении работ. Приступать к внедрению можно сначала в какой-то одной части организации или в отдельной предметной области — например, по отношению к данным

о продуктах или клиентах. По мере развития навыков и повышения уровня зрелости сферу внедрения можно расширять.

Модели и другие артефакты архитектуры данных обычно первоначально создаются в рамках отдельных проектов разработки и лишь затем стандартизируются и переходят под управление архитекторов данных. Следовательно, значительная часть работы по созданию архитектуры выполняется в ходе первых ориентированных на архитектурную деятельность проектов, то есть еще до появления артефактов, предназначенных для повторного использования. В связи с этим бывает полезно включать в бюджет таких проектов отдельную статью затрат на проектирование архитектуры данных.

Корпоративный архитектор данных работает в тесном сотрудничестве с бизнес- и техническими архитекторами для достижения общей цели — повышения эффективности и гибкости функционирования организации. Бизнес-драйверы совершенствования общей корпоративной архитектуры также в значительной мере влияют и на стратегию внедрения корпоративной архитектуры данных.

Для ввода в действие корпоративной архитектуры данных в условиях организационной культуры, ориентированной на поиск новых решений с применением прорывных экспериментальных технологий, требуется гибкий (agile) подход к внедрению. Он может предусматривать, например, общую верхнеуровневую концептуальную модель предметной области, гибко и оперативно дополняемую по мере необходимости детализациями на локальных уровнях в рамках «спринтов» Agile. При таком гибком подходе корпоративная архитектура данных развивается пошагово, за счет небольших приращений. Однако этот путь требует гарантии обязательного участия архитекторов данных во всех инициативах в области разработки с самого их начала, поскольку в условиях инновационной культуры они развиваются очень быстро.

Наличие эффективных драйверов внедрения корпоративной архитектуры может способствовать началу проведения на корпоративном уровне некоторых пилотных работ по созданию архитектуры данных в интересах плановых проектов разработки. Обычно корпоративный архитектор данных начинает проектирование с областей основных данных (master data), наиболее нуждающихся в улучшении и совершенствовании, а после того, как они определены и согласованы, расширяет модель посредством включения в нее данных, ориентированных на бизнес-события (транзакционные данные). Это традиционный подход к внедрению, при котором корпоративные архитекторы данных выпускают рабочие описания и шаблоны, а они, в свою очередь, используются применительно ко всему системному ландшафту; при этом должен быть обеспечен надлежащий контроль соответствия посредством различных механизмов руководства.

5.1 Оценка готовности / Оценка рисков

Проекты по инициированию архитектурной деятельности сопряжены со значительно бóльшими рисками, чем любые другие проекты, особенно во время первой такой попытки внутри организации. Самые серьезные риски бывают обусловлены следующими факторами.

- **Недостаточная поддержка со стороны руководства.** Любая реорганизация предприятия или учреждения в период планового выполнения проекта способна повлиять на архитектурный процесс. Например, у нового руководства возникают вопросы в отношении проведения архитектурной деятельности, и оно может не дать согласия на продолжение участниками проекта их работы по созданию архитектуры данных. Лишь заручившись надежной поддержкой руководства, в целом можно рассчитывать на то, что архитектурный процесс переживет реорганизацию. Следовательно, необходимо постараться привлечь к процессу разработки архитектуры данных как минимум нескольких руководителей высшего или хотя бы старшего звена, понимающих преимущества архитектуры данных.
- **Отсутствие документально подтвержденных достижений.** Для успеха начинания важно иметь твердого сторонника и поручителя (спонсора — sponsor) из числа старших по рангу, который к тому же не сомневается в способности команды выполнять функцию по созданию и сопровождению архитектуры данных. Полезно бывает заручиться поддержкой и привлечь в качестве консультанта на критически важных этапах проектирования кого-то из главных архитекторов.
- **Настороженность или обеспокоенность спонсора проекта.** Если спонсор требует, чтобы всякий обмен информацией и сообщениями (равно как и все согласования) проходил строго через него/нее, это может указывать на различные неблагоприятные факторы: он или не вполне понимает свою роль, или боится за свое место, или преследует иные интересы, нежели те, что диктуются задачами проектирования архитектуры данных, или не уверен в способности архитекторов данных справиться со своими задачами. Причины для беспокойства у спонсора могут быть самые разные, но он обязан предоставить руководителю проекта и архитектору данных возможность выполнения ведущих ролей в реализации проекта. Необходимо всячески добиваться независимого положения специалистов, а также повышения уверенности спонсора.
- **Контрпродуктивные решения руководства.** Бывает так, что руководство вроде бы и осознаёт всю степень ценности хорошо организованной архитектуры данных, но не понимает, как обеспечить ее реализацию. Из-за этого могут внезапно приниматься решения, противоречащие усилиям архитекторов данных. Это свидетельствует не о потере поддержки со стороны руководства, а всего лишь о том, что архитектору данных нужно чаще, четче и яснее доносить свою позицию до его сведения.
- **Культурный шок.** Следует принимать в расчет то, как изменится рабочая культура тех сотрудников, которых затронет архитектурная деятельность. Можно попытаться представить, насколько просто или сложно им будет изменить свое поведение внутри организации.
- **Неопытный руководитель проекта.** Необходимо удостовериться, что руководить проектом поручено лицу с достаточным опытом в области корпоративной архитектуры данных, особенно если значительная часть проекта связана непосредственно с данными. Если выяснится, что опыта недостаточно, убедите спонсора проекта либо заменить руководителя, либо провести его обучение (Edvinsson, 2013).

- **Одностороннее представление.** Случается так, что владелец (owner) (или владельцы) одного из бизнес-приложений (например, ERP-системы) проявляют тенденцию навязывать свои взгляды на общую корпоративную архитектуру данных в ущерб более сбалансированному и всестороннему представлению.

5.2 Организационные и культурные изменения

Скорость, с которой организация осваивает архитектурную практику, зависит от того, насколько адаптивна ее культура. Природа проектной работы подразумевает взаимодействие архитекторов с разработчиками и другими креативно мыслящими специалистами повсюду в организации. Зачастую такие люди привыкли работать каждый по-своему. Соответственно, они могут как принять, так и отторгнуть изменения, необходимые для формального внедрения архитектурных принципов и инструментов.

Сфокусированные на результате, стратегически ориентированные организации находятся в лучшем положении с точки зрения восприятия архитектурных практик. Таким организациям чаще всего свойственны целеустремленность, знание проблем клиентов и партнеров, способность к расстановке приоритетов исходя из общих задач.

Способность организации к освоению практики разработки и сопровождения архитектуры данных зависит от ряда факторов, среди которых:

- восприимчивость организационной культуры к архитектурному подходу (выработка культуры, ориентированной на архитектурные решения);
- признание организацией данных как бизнес-актива, а не только как объекта заботы сферы ИТ;
- способность организации отрешиться от локального взгляда на данные и представлять их комплексно, на корпоративном уровне;
- способность организации интегрировать результаты архитектурной деятельности в методологию реализации проектов;
- степень принятия формального руководства данными;
- способность составить целостное представление об организации, а не фокусироваться исключительно на реализации проектов и ИТ-решениях (Edvinsson, 2013).

6. РУКОВОДСТВО АРХИТЕКТУРОЙ ДАННЫХ

Работы в области архитектуры данных направлены на непосредственную поддержку деятельности по согласованию и контролю данных. Архитекторы данных часто выполняют роль связующего звена между бизнесом и деятельностью по руководству данными. Поэтому организационные системы, осуществляющие деятельность в области архитектуры данных и руководства данными, должны действовать согласованно. В идеале за каждой предметной областью и даже за каждой сущностью предметной области должны быть закреплены ответственные архитектор данных

и распорядитель данных. Кроме того, следует согласовывать функции надзора за организацией бизнеса и надзора за процессами. Предметные области бизнес-событий должны быть согласованы с функцией руководства бизнес-процессами, поскольку любая сущность, относящаяся к событию, обычно соответствует тому или иному бизнес-процессу. Работы в области руководства архитектурой данных включают следующее.

- **Надзор за проектами.** Сюда относится обеспечение того, чтобы в проектах проводились требуемые работы в области архитектуры данных, использовались и развивались имеющиеся в активе организации архитектурные решения, а также чтобы реализация проектов осуществлялась в соответствии с установленными архитектурными стандартами.
- **Управление архитектурными решениями, жизненным циклом и инструментами.** Архитектурные решения подлежат четкому определению, оценке и сопровождению. Корпоративная архитектура данных служит в качестве «плана зонирования» (zoning plan) для проведения интеграции в долгосрочной перспективе. Будущая архитектура оказывает влияние на задачи проекта и учитывается при определении приоритетности проектов в портфеле проектов организации.
- **Определение стандартов.** Определение правил, руководств и спецификаций, устанавливающих порядок использования данных.
- **Создание артефактов, относящихся к данным.** Артефакты, обеспечивающие соблюдение директив руководства.

6.1 Метрики

Метрики для оценки эффективности корпоративной архитектуры данных отражают архитектурные цели и характеризуют соответствие архитектурным требованиям, тренды внедрения и ценность, приносимую архитектурой данных бизнесу. Значения метрик архитектуры данных обычно фиксируются ежегодно в рамках мониторинга общей удовлетворенности клиентов проектами.

- **Уровень соблюдения архитектурных стандартов.** Отражает, насколько строго проекты выполняют требования принятых архитектур данных и придерживаются практики учета в своих процессах корпоративной архитектуры. Метрики, отражающие случаи отказа от соблюдения архитектурных требований, могут быть также полезны в качестве средства обеспечения понимания узких мест и препятствий на пути принятия архитектурной практики.
- **Тренды внедрения.** Позволяют отслеживать, насколько корпоративная архитектура повлияла на улучшение способности организации реализовывать проекты как минимум по двум направлениям.
  - **Оценки количества используемых / повторно используемых / замененных / отмененных архитектурных артефактов.** Соотношение этих показателей позволяет судить об интенсивности внедрения и доле новых архитектурных разработок.

- **Оценки эффективности выполнения проектов.** Оцениваются сроки и затраты на реализацию проектов в целях выявления резервов для совершенствования за счет применения повторно используемых артефактов и руководящих артефактов.

- **Оценки ценности для бизнеса** позволяют отслеживать прогресс в достижении ожидаемых бизнес-эффектов и выгод.
  - **Повышение гибкости бизнеса.** Показатели, отражающие выгоды, полученные за счет усовершенствований жизненного цикла данных, или, в качестве альтернативы, стоимость потерь, обусловленных задержками.
  - **Качество бизнес-решений.** Показатели степени соответствия результатов внедренных бизнес-решений изначально запланированным; позволяют оценить, в какой степени проекты обеспечили изменения, которые привели к улучшениям в бизнесе за счет создания новых или интеграции имеющихся данных.
  - **Качество операционной деятельности.** Показатели повышения эффективности; примеры включают повышение точности, снижение затрат времени и расходов на исправление ошибок, обусловленных некорректными данными, и т. п.
  - **Улучшения в бизнес-среде.** Примеры включают повышение коэффициента удержания клиентов за счет снижения процента ошибок в данных и снижение числа замечаний контролирующих и надзорных органов к представляемым отчетам.

7. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Ahlemann, Frederik, Eric Stettiner, Marcus Messerschmidt, and Christine Legner, eds. *Strategic Enterprise Architecture Management: Challenges, Best Practices, and Future Developments*. Springer, 2012. Print. Management for Professionals.

Bernard, Scott A. *An Introduction to Enterprise Architecture*. 2nd ed. Authorhouse, 2005. Print.

Brackett, Michael H. *Data Sharing Using a Common Data Architecture*. John Wiley and Sons, 1994. Print.

Carbone, Jane. *IT Architecture Toolkit*. Prentice Hall, 2004. Print.

Cook, Melissa. *Building Enterprise Information Architectures: Re-Engineering Information Systems*. Prentice Hall, 1996. Print.

Edvinsson, Hakan and Lottie Aderinne. *Enterprise Architecture Made Simple Using the Ready, Set, Go Approach to Achieving Information Centricity*. Technics Publications, LCC, 2013. Print.

Executive Office of the President of the United States. *The Common Approach to Federal Enterprise Architecture*, whitehouse.gov, 2012. Web.

Fong, Joseph. *Information Systems Reengineering and Integration*. 2nd ed. Springer, 2006. Print.

Gane, Chris and Trish Sarson. *Structured Systems Analysis: Tools and Techniques*. Prentice Hall, 1979. Print.

Hagan, Paula J., ed. *EABOK: Guide to the (Evolving) Enterprise Architecture Body of Knowledge*, mitre.org MITRE Corporation, 2004. Web.

Harrison, Rachel. *TOGAF Version 8.1.1 Enterprise Edition — Study Guide*. The Open Group. 2nd ed. Van Haren Publishing, 2007. Print. TOGAF.

Hoberman, Steve, Donna Burbank, and Chris Bradley. *Data Modeling for the Business: A Handbook for Aligning the Business with IT using High-Level Data Models*. Technics Publications, LLC, 2009. Print. Take It with You Guides.

Hoberman, Steve. *Data Modeling Made Simple: A Practical Guide for Business and Information Technology Professionals*. 2nd ed. Technics Publications, LLC, 2009. Print.

Hoogervorst, Jan A. P. *Enterprise Governance and Enterprise Engineering*. Springer, 2009. Print. The Enterprise Engineering Ser.

ISO (website), http://bit.ly/2sTp2rA, http://bit.ly/2ri8Gqk

Inmon, W. H., John A. Zachman, and Jonathan G. Geiger. *Data Stores, Data Warehousing and the Zachman Framework: Managing Enterprise Knowledge*. McGraw-Hill, 1997. Print.

Lankhorst, Marc. Enterprise Architecture at Work: Modeling, Communication and Analysis. Springer, 2005. Print.

Martin, James and Joe Leben. *Strategic Information Planning Methodologies*, 2nd ed. Prentice Hall, 1989. Print.

Osterwalder, Alexander and Yves Pigneur. *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers*. Wiley, 2010. Print.

Perks, Col and Tony Beveridge. *Guide to Enterprise IT Architecture*. Springer, 2003. Print. Springer Professional Computing.

Poole, John, Dan Chang, Douglas Tolbert, and David Mellor. *Common Warehouse Metamodel*. Wiley, 2001. Print. OMG (Book 17).

Radhakrishnan, Rakesh. *Identity and Security: A Common Architecture and Framework For SOA and Network Convergence*. Futuretext, 2007. Print.

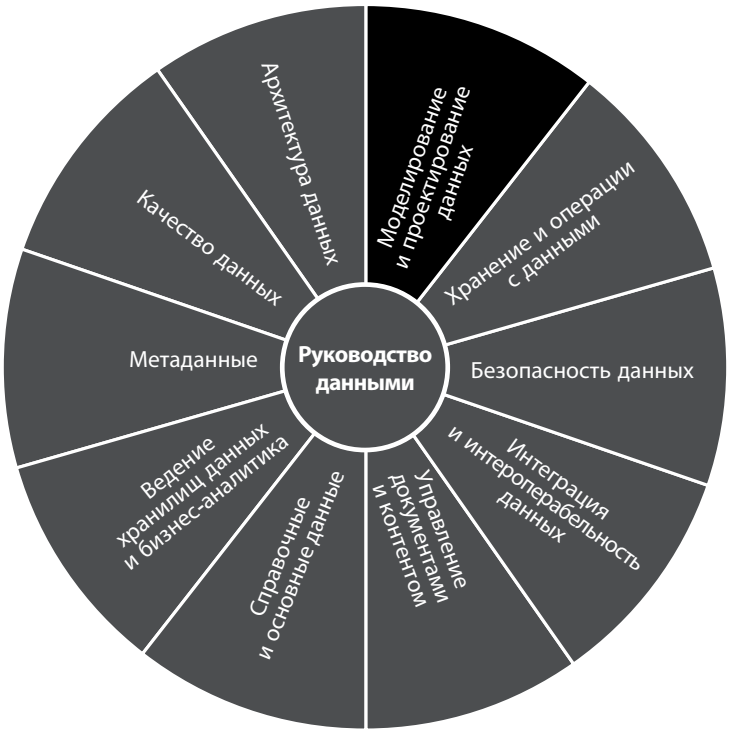
Ross, Jeanne W., Peter Weill, and David Robertson. *Enterprise Architecture As Strategy: Creating a Foundation For Business Execution*. Harvard Business School Press, 2006. Print.

Schekkerman, Jaap. *How to Survive in the Jungle of Enterprise Architecture Frameworks: Creating or Choosing an Enterprise Architecture Framework*. Trafford Publishing, 2006. Print.

Spewak, Steven and Steven C. Hill. *Enterprise Architecture Planning: Developing a Blueprint for Data, Applications, and Technology*. 2nd ed. A Wiley-QED Publication, 1993. Print.

Ulrich, William M. and Philip Newcomb. *Information Systems Transformation: Architecture-Driven Modernization Case Studies*. Morgan Kaufmann, 2010. Print. The MK/OMG Press.

Моделирование и проектирование данных



DAMA-DMBOK2 Data Management Framework

1. ВВЕДЕНИЕ

Моделирование данных заключается в последовательном выявлении, анализе и формулировании основных требований к данным с последующим их представлением и распространением в точно определенной форме, называемой *моделью данных*. Моделирование данных — критически важный компонент управления данными. Процесс моделирования требует от организации выяснения и документирования того, как ее данные соотносятся друг с другом в рамках общей картины. В то же время моделирование само по себе заключается в разработке решений в отношении компоновки данных и их связи (Simsion, 2013). Модели данных отражают и одновременно улучшают понимание организацией информационных активов, которыми она располагает и оперирует.

Рисунок 28. Контекстная диаграмма: моделирование и проектирование данных

Существуют различные схемы представления данных[1]. Наиболее часто используются шесть следующих схем: реляционная, многомерная (dimensional), объектно-ориентированная, основанная на фактах (fact-based), хронологическая (time-based) и NoSQL[2]. Модели данных во всех этих схемах представляются на трех уровнях детализации — концептуальном, логическом и физическом. Каждая модель содержит набор компонентов. Примеры компонентов — сущности, связи, факты, ключи, атрибуты. По завершении построения модели она подлежит согласованию и уточнению, а после утверждения — сопровождению.

Модели данных содержат важные для потребителей данных метаданные. Значительная часть этих метаданных, выявленных в процессе моделирования, необходима другим функциям управления данными. Например, определения, требующиеся для руководства данными, или информация, относящаяся к происхождению данных и используемая при ведении хранилищ данных, а также в бизнес-аналитике.

В настоящей главе описаны: назначение моделей данных; основные понятия и общепринятая терминология, используемые в моделировании данных; цели и принципы моделирования. Для того чтобы наглядно показать, как работают различные модели данных и чем они отличаются друг от друга, приведены примеры из области образования.

1.1 Бизнес-драйверы

Модели данных имеют критическое значение для эффективного управления данными, поскольку они:

- определяют единую общую терминологию во всем, что касается данных;
- собирают и документируют точные знания о данных и информационных системах организации;
- служат основным средством коммуникации в процессе реализации проектов;
- являются отправной точкой при настройке, интеграции или даже замене приложений.

1.2 Цели и принципы

Главная цель моделирования данных — подтвердить и документально зафиксировать понимание различных аспектов организации данных, которое обеспечит создание приложений, наиболее точно соответствующих текущим и будущим потребностям бизнеса, а также заложить фундамент для успешной реализации широкомасштабных инициатив, таких как программы управления основными данными и руководства данными. Правильное моделирование данных приводит к снижению затрат на поддержку и расширяет возможности повторного использования моделей при проведении в жизнь будущих инициатив, способствуя таким образом минимизации затрат на создание новых приложений. Модели данных — важная форма метаданных.

---

[1] В русскоязычной литературе в таких случаях чаще говорится о «типах моделей данных». — *Примеч. науч. ред.*

[2] В данном издании используется устоявшийся термин NoSQL (от *англ.* not only SQL — не только SQL). — *Примеч. науч. ред.*

Подтверждение и документирование понимания различных аспектов организации данных и перспектив в рамках моделирования данных способствует более эффективной деятельности по следующим направлениям.

- **Формализация.** Модель данных документирует краткое и четкое определение структур данных и связей между ними. Она позволяет оценивать, как влияют на данные реализованные бизнес-правила (как для текущих, так и для будущих целевых состояний). Формальное определение вводит строго соблюдаемую структуру данных, что снижает вероятность нарушений при обеспечении доступа к данным и их ведении. Иллюстрируя структуры данных и связи между их элементами, модель данных упрощает их практическое использование.
- **Определение области применения.** Модель данных помогает объяснить границы контекста данных, а также границы внедрения приобретенного программного обеспечения и области охвата проектов, инициатив и существующих систем.
- **Сохранение/документирование знаний.** Модель данных может сохранять корпоративную память о какой-либо системе или проекте, фиксируя знания в четко определенной форме. Она служит документацией для будущих проектов в качестве версии «как есть». Модели данных помогают нам лучше понимать различные аспекты организации или бизнеса, механизмы работы приложений или последствия изменений существующей структуры данных. Таким образом, модель данных становится многократно используемой картой, помогающей профессионалам в области бизнеса, руководителям проектов, аналитикам, специалистам по моделированию и разработчикам лучше понимать структуру данных в контексте среды окружения. Так же как картографы изучают и документируют географический ландшафт, помогая другим осуществлять навигацию, специалисты по моделированию данных помогают другим понять информационный ландшафт (Hoberman, 2009).

1.3 Основные понятия и концепции

В настоящем разделе рассматриваются различные виды данных, которые могут моделироваться, отдельные компоненты моделей данных, а также типы моделей данных и причины, обусловливающие выбор в пользу той или иной модели в зависимости от ситуации. Набор определений в данном разделе весьма обширен отчасти потому, что моделирование данных само по себе в значительной степени является процессом определения. Важно иметь четкое понимание терминологии, которая поддерживает практику.

1.3.1 Моделирование и модели данных

Моделирование данных чаще всего проводится в контексте деятельности по разработке и сопровождению систем, известной как жизненный цикл разработки систем (SDLC). Моделирование данных также может осуществляться в рамках широкомасштабных инициатив (например, инициативы в области архитектуры бизнеса и данных, управления основными данными, руководства

данными), непосредственным результатом которых являются не базы данных, а лучшее понимание данных организации.

Модель — это представление чего-либо, что уже существует, или примерный образец того, что предстоит создать. Модель может содержать одну или несколько диаграмм. В каждой диаграмме используются стандартные символы, обеспечивающие понимание ее смыслового содержания. Примерами широко распространенных моделей являются карты, схемы организационных структур, чертежи зданий.

Модель данных либо описывает данные организации так, как они понимаются на текущий момент, либо отражает то состояние данных, в котором организация хотела бы их видеть. Модель данных содержит набор символов с текстовыми метками, предназначенными для визуального представления требований к данным, в том виде, в котором их сообщили специалисту по моделированию. При этом количество элементов описываемой области данных может варьироваться от небольшого (если рассматривается отдельный проект) до весьма внушительного (если рассматривается организация). Модель данных является формой документирования требований к данным и определений данных. Получаемые в результате процесса моделирования документально оформленные модели — главное средство коммуникации, обеспечивающее передачу требований к данным от сферы бизнеса в блок ИТ, а также (в рамках блока ИТ) от аналитиков, специалистов по моделированию и архитекторов взаимодействующим с ними проектировщикам и разработчикам баз данных.

1.3.2 Виды моделируемых данных

Моделироваться могут данные четырех основных видов (Edvinsson, 2013). Виды данных, моделируемые в конкретной организации, отражают приоритеты организации или проекта, которым требуется модель данных.

- **Информация о категориях.** Данные, используемые для классификации объектов и отнесения их к определенным категориям. Например, клиенты могут классифицироваться по сегментам рынка или направлениям бизнеса; продукты — по модели, цвету и размеру; заказы — по статусу выполнения; и т. д.
- **Информация о ресурсах.** Основные профили ресурсов, необходимых для осуществления операционной деятельности, например: продукт, клиент, поставщик, оборудование, организация, счет. Среди профессионалов в сфере ИТ применительно к сущностям, относящимся к ресурсам, иногда используется термин *справочные данные*.
- **Информация о бизнес-событиях.** Данные, создаваемые в ходе операционной деятельности. Примеры: заказы клиентов, счета поставщиков, платежи, бизнес-встречи. В сфере ИТ применительно к сущностям, относящимся к событиям, иногда используется термин *транзакционные бизнес-данные*.
- **Детальная информация о транзакциях.** Детализированная информация об операциях, поступающая обычно из систем оплаты (в магазинах или онлайн). Сюда же относится

информация, получаемая из социальных медиа, других источников информации об интернет-активности (счетчики посещений и т. п.), собираемая с телеметрических датчиков всевозможных транспортных средств и промышленного оборудования или поступающая с персональных устройств (GPS, RFID, Wi-Fi и т. п.). Подобная детализированная информация может агрегироваться и использоваться с целью получения других данных и анализа тенденций, примерно так же, как используется информация о бизнес-событиях. Данные этого вида (накапливаемые в больших объемах и/или стремительно меняющиеся) принято называть *большими данными*.

Все вышеперечисленные виды данных называют «данными в покое» (data at rest) (статичные данные, находящиеся в местах хранения). Но и «данные в движении» (data in motion) (динамичные, перемещающиеся данные) также можно моделировать: например, с помощью описаний системных решений, включая протоколы, а также специфических схем для систем обмена сообщениями и систем на основе событий (event-based systems).

1.3.3 Компоненты модели данных

Далее в этой главе подробно обсуждаются причины, по которым различные типы моделей данных представляют данные с помощью различных соглашений (см. раздел 1.3.4). Однако основные строительные блоки в большинстве моделей данных одни и те же: сущности, связи, атрибуты и области значений атрибута (домены).

1.3.3.1 СУЩНОСТЬ

В общем смысле — вне контекста моделирования данных — под сущностью понимается предмет, существующий отдельно от других предметов. В рамках моделирования данных сущность — это предмет, о котором организация собирает информацию. Иногда сущности уподобляют «существительным» организации (по аналогии с членами предложения). Действительно, сущность можно рассматривать как ответ на один из фундаментальных вопросов (кто, что, где, когда, почему и как) или сочетание этих вопросов (см. главу 4). Таблица 7 дает определения и приводит примеры общеупотребительных категорий сущностей (Hoberman, 2009).

Таблица 7. Общеупотребительные категории объектов

Категория	Определение	Примеры
Кто	Физическое лицо или организация, представляющие интерес. Тот, *Кто* важен для бизнеса. Категория «*Кто*» часто представляет однородные группы или роли, такие как клиент или поставщик. При этом любое лицо или организация могут быть включены в различные группы и выступать в различных ролях	Работодатель, Пациент, Игрок, Подозреваемый, Клиент, Поставщик, Студент, Пассажир, Конкурент, Автор

Категория	Определение	Примеры
Что	Продукт, услуга, товар или иной предмет интереса предприятия. То, *Что* важно для бизнеса. Обычно используется для определения категорий производимой продукции или услуг, оказываемых организацией. Тут крайне важно четко определять атрибуты категорий, типов и т. п.	Продукт, Услуга, Поставка, Товар, Курс, Саундтрек, Фотография, Книга
Когда	Календарные даты, сроки или периоды, интересующие организацию. *Когда* именно действует то, что важно для бизнеса	Время, Дата, Месяц, Квартал, Год, Расписание, Семестр, Срок, Время отправления
Где	Места локализации интересов организации, включая фактические физические, почтовые и электронные адреса. *Где* ведется бизнес	Почтовый адрес, Пункт выдачи, URL веб-сайта, IP-адрес
Почему	События или транзакции, представляющие интерес для организации и держащие бизнес на плаву. *Почему* и для чего делается то, что делается	Заказ, Возврат, Претензия, Жалоба, Депозит, Отзыв, Запрос, Замена, Рекламация
Как	Документация, относящаяся к событиям, интересующим организацию. Документы служат подтверждением факта того или иного события (например, оформления заказа или его исполнения). *Как* мы определяем, имело ли место то или иное событие	Счет-фактура, Договор, Контракт, Заказ на покупку, Квитанция об уплате штрафа за нарушение, Товарная накладная, Подтверждение сделки
Измерение	Итоги, суммы и т. п. Сводные и контрольные данные по всем остальным категориям	Продажи, Количество единиц товара, Платежи, Баланс

1.3.3.1.1 Альтернативные наименования (aliases) понятия «сущность»

Общий термин *сущность* иногда может фигурировать под иными наименованиями. Чаще всего при этом используется понятие *тип сущности*, как тип чего-то, что должно быть представлено. Например, Джейн (конкретный человек) относится к типу Сотрудник; таким образом, Джейн является сущностью, а Сотрудник является типом сущности. Однако сегодня широко распространенной практикой является использование термина *сущность* в отношении типа Сотрудник, а в отношении конкретного человека — Джейн — используется термин *экземпляр сущности* (*entity instance*).

Таблица 8. Сущность, тип сущности и экземпляр сущности

Использование	Сущность	Тип сущности	Экземпляр сущности
Общепринятое	Джейн	Сотрудник	
Рекомендуемое	Сотрудник		Джейн

Экземпляры сущностей представляют собой их материальные воплощения или значения, которые их описывают. Сущность Студент может быть представлена множеством экземпляров с именами и фамилиями — Боб Джоунс, Джо Джексон, Джейн Смит и т. д. Сущность Предмет (учебный) также может иметь несколько экземпляров, например: Английская литература XVII века, Геология, Основы моделирования данных и т. д.

Используемые для обозначения понятия *сущность* альтернативные наименования зависят также от схемы представления данных (см. раздел 1.3.4). В реляционных схемах обычно применяется сам термин *сущность*; в многомерных схемах используются термины *таблица фактов* (*fact table*) и *таблица измерений* (*dimension table*); в объектно-ориентированных схемах — термины *класс* (*class*) или *объект* (*object*); в хронологических схемах — термины *концентратор* (или *хаб — hub*), *сателлит* (*satellite*) и *связь* (*link*), а в схемах NoSQL — такие термины, как *документ* (*document*) или *узел* (*node*).

Использование альтернативных наименований может зависеть и от уровня детализации (см. раздел 1.3.5). На концептуальном уровне сущность может называться *концептом* (*concept*) или *термином* (*term*), на логическом — собственно *сущностью* (также может использоваться альтернативное наименование, принятое в схеме), а на физическом уровне наименование зависит от технологии реализации базы данных, но чаще всего используется термин *таблица*.

1.3.3.1.2 Графическое представление сущностей

В любых моделях данных сущности обычно отражают в виде прямоугольников (иногда со скругленными углами) с именами внутри. В нижеприведенном примере (см. рис. 29) представлены три сущности: Студент, Предмет и Преподаватель.

Рисунок 29. Сущности

1.3.3.1.3 Определение сущностей

Определение сущностей вносит важный вклад в ценность модели данных для бизнеса. Они являются ключевыми метаданными. Хорошо продуманные (качественные) определения показывают важность бизнес-словаря и придают строгость бизнес-правилам, управляющим взаимоотношениями между сущностями. Они помогают профессионалам в сфере бизнеса и ИТ принимать разумные решения относительно ведения бизнеса и разработки приложений. О качестве определения сущностей судят по их соответствию следующим основным критериям.

- **Ясность.** Определение должно быть легко читаемым, понятным и запоминающимся. Формулировка должна быть четкой и грамотной, не содержать сокращений (кроме общеупотребимых) и неоднозначно трактуемых терминов, таких как «*иногда*» или «*нормальный*».
- **Точность.** Определение является точным и корректным описанием сущности. Для этого полезно согласовывать определения с экспертами в профильных областях бизнеса.

- **Полнота.** Важно не упустить ни единой детали определения. Например, в определение кода следует включать примеры значений кода, а в определение идентификатора — описание границ его уникальности.

1.3.3.2 Связь

Связь (relationship) — это отношение (ассоциация — association) между сущностями (Chen, 1976). Связи фиксируют информацию о взаимодействиях между концептуальными сущностями, детализированные взаимодействия между логическими сущностями и взаимные ограничения при взаимодействии физических сущностей.

1.3.3.2.1 Альтернативные наименования понятия «связь»

Общий термин *связь* может фигурировать под каким-либо иным наименованием. Альтернативное наименование зависит от схемы. В реляционных схемах *связи* так и называются; в многомерных схемах вместо термина *связь* часто используется термин *путь навигации* (*navigation path*), а в схемах NoSQL, к примеру, *ребро* (*edge*) или *ссылка* (*link*). Часто термины, используемые для обозначения связей, зависят и от уровня детализации. На концептуальном и логическом уровнях *связь* так и называется, а вот на физическом уровне связи могут фигурировать и под другими наименованиями: например, *ограничение* (*constraint*) или *ссылка* (*reference*), в зависимости от используемой технологии реализации базы данных.

1.3.3.2.2 Графическое представление связей

Связи на диаграммах моделей данных принято отображать линиями (см. рис. 30).

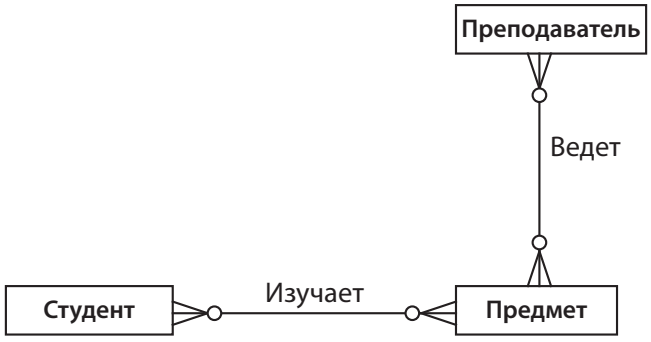


Рисунок 30. Связи

В приведенном примере связи фиксируют два правила: Студенты могут изучать различные Предметы; Преподаватели могут вести занятия по различным Предметам. Символы на концах связей («вилки») отражают их множественность. Обозначаемая ими характеристика называется мощностью (cardinality) связи. Таким образом, правила зафиксированы посредством точного синтаксиса (см. п. 1.3.3.2.3). В реляционных базах данных связи представляются через внешние

ключи (foreign keys), а в базах данных NoSQL — с помощью альтернативных методов: например, через ребра или ссылки.

1.3.3.2.3 Мощность связи

Мощность (*cardinality*) связи между двумя сущностями определяет, сколько экземпляров одной сущности и сколько экземпляров другой могут быть связаны друг с другом. Мощность отображается специальными символами на обоих концах линии связи. Правила в области данных определяются через мощность, без указания которой максимум, что можно сказать о связи, — это то, что она каким-то образом реализована.

Допустимые значения мощности — ноль, один или много («много» означает «больше чем один»). Допускаются произвольные сочетания трех этих значений на противоположных концах связи. Обозначая нулевую или единичную мощность, мы тем самым фиксируем наличие или отсутствие связи. Задавая мощность как один или много, мы определяем точное число экземпляров сущности, участвующих в образовании данной связи.

Ниже представлен пример использования символов мощности связи между объектами Студент и Предмет.

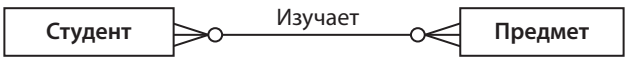


Рисунок 31. Символы мощности связи

Зафиксированные с помощью связи бизнес-правила интерпретируются следующим образом.

- Каждый Студент может изучать один или много Предметов.
- Каждый Предмет может изучаться одним или многими Студентами.

1.3.3.2.4 Арность связей

Число сущностей, участвующих в образовании связи называется «арностью» (arity) связи. На практике широко используются в основном унарные, бинарные и тернарные связи.

1.3.3.2.4.1 Унарная (рекурсивная) связь

Унарная (известная также как рекурсивная) связь соотносит между собой экземпляры одной и той же сущности. Унарная связь одного экземпляра со многими описывает иерархию, а многих со многими — сеть или граф. В иерархии всякий экземпляр сущности не может иметь более одного родительского (или вышестоящего) экземпляра. В реляционных моделях иерархии дочерние экземпляры находятся на множественной (вильчатой) стороне связи, родительские — на одинарной. При сетевой унарной связи любой экземпляр объекта может иметь более одного родителя.

Например, возможность изучения Предмета может быть обусловлена выполнением неких предварительных требований. Предположим, Семинары по биологии могут посещать только

слушатели Лекций по биологии, то есть Лекция по биологии — обязательное условие Семинара по биологии. В реляционных моделях данных, как показано ниже в примерах стандартных обозначений, такую рекурсивную связь можно представить двояко.

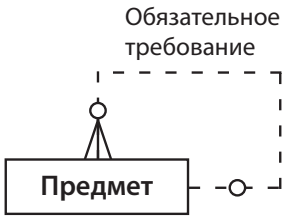


Рисунок 32. Унарная связь (иерархическая)

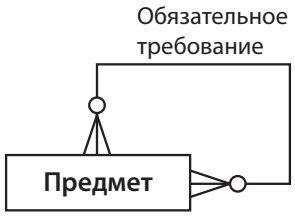


Рисунок 33. Унарная связь (сетевая)

В первом примере (рис. 32) использована иерархическая унарная связь, во втором (рис. 33) — сетевая. В первом случае единственным непременным условием посещения Семинаров по биологии является посещение Лекций по биологии, а во втором — Лекций по биологии и (например) Лекций по химии. В рамках иерархической модели, если Лекции по биологии выбраны в качестве обязательного условия Семинаров по биологии, они не могут использоваться в качестве обязательного условия для изучения других предметов. Во втором случае Лекции по биологии могут быть сделаны обязательным требованием и для допуска к изучению других предметов.

1.3.3.2.4.2 Бинарная связь

Бинарная связь, то есть связь между двумя сущностями, — самая распространенная. Она широко применяется в традиционных диаграммах моделей данных. На рисунке 34 представлена диаграмма классов унифицированного языка моделирования (Unified Modeling Language, UML), отражающая бинарную связь между сущностями Студент и Предмет.

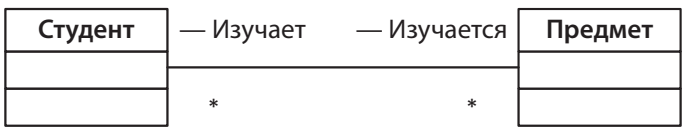


Рисунок 34. Бинарная связь

1.3.3.2.4.3 Тернарная связь

Тернарная связь устанавливается между тремя сущностями. Рисунок 35 содержит пример, относящийся к моделированию на основе фактов (объектно-ролевая нотация). Студент может быть зарегистрирован в качестве изучающего Предмет в течение Семестра.

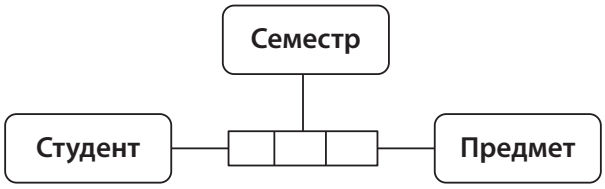


Рисунок 35. Тернарная связь

1.3.3.2.5 Внешний ключ

Внешний ключ (foreign key) используется в физических и (реже) логических реляционных схемах моделирования данных для представления связи. Создаваться внешний ключ может в том числе и неявно, при определении связи между двумя объектами, в зависимости от технологии реализации базы данных или используемого инструмента моделирования данных и наличия или отсутствия взаимозависимостей между двумя объектами.

В следующем примере (рис. 36) объект Регистрация содержит два внешних ключа (ВК) — Студент № из объекта Студент и Код обучающего курса из объекта Обучающий курс. Внешние ключи отражаются в сущности на множественной стороне связи (такую сущность часто называют дочерней). В приведенном примере Студент и Обучающий курс — родительские сущности, а Регистрация — дочерняя сущность.

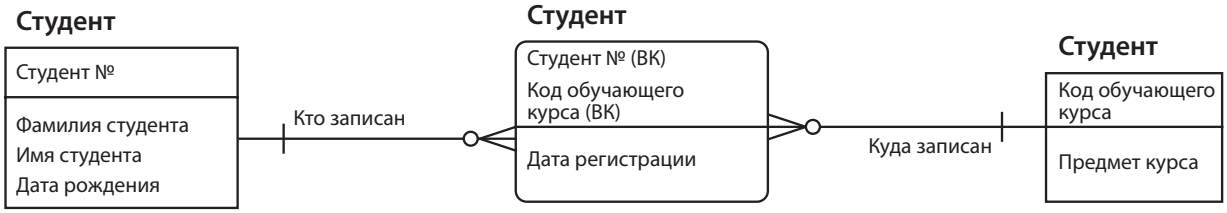


Рисунок 36. Внешние ключи

1.3.3.3 АТРИБУТ

Атрибут (attribute) — это характеристика сущности, позволяющая ее идентифицировать, описать или измерить. Для атрибута может быть определен домен (domain) — совокупность возможных значений (см. п. 1.3.3.4). На физическом уровне атрибуту сущности может соответствовать столбец, поле, тег или узел (место пересечения) в таблице, представлении, документе, графе или файле.

1.3.3.3.1 Графическое представление атрибутов

В моделях данных атрибуты обычно отображаются в виде списка в прямоугольнике сущности. В приведенном примере (рис. 37) атрибутами сущности Студент являются Студент №, Фамилия студента, Имя студента и Дата рождения.

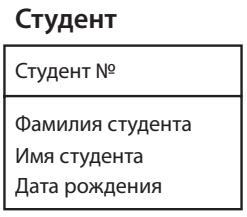


Рисунок 37. Атрибуты

1.3.3.3.2 Идентификатор

Идентификатором (или *ключом — key*) называют атрибут или набор атрибутов, уникальным образом определяющий экземпляр сущности. Далее в этом разделе описаны всевозможные типы ключей, классифицированные по их конструкции (простой, составной, композитный, суррогатный) и функциям (потенциальный, первичный, альтернативный).

1.3.3.3.2.1 Типы ключей в зависимости от конструкции

Простой ключ (*simple key*) — один атрибут, уникальным образом идентифицирующий экземпляр сущности. Примеры простых ключей — универсальные коды продуктов (Universal Product Codes, UPCs) и идентификационные номера транспортных средств (Vehicle Identification Numbers, VINs). Подвидом простого ключа является *суррогатный ключ* (*surrogate key*), используемый в качестве уникального идентификатора для записей таблицы. Генерируемый исключительно автоматически (с помощью счетчика или случайным образом), суррогатный ключ представляет собой целое число, не несущее никакой смысловой нагрузки. (Иными словами, если, к примеру, экземпляру сущности Месяц присвоен идентификатор «1», из этого вовсе не следует, что этот месяц — январь.) Суррогатные ключи выполняют чисто технические функции, и конечным пользователям баз данных видеть их ни к чему. Поэтому они остаются за кадром, помогая обеспечивать целостность данных, оптимизировать переходы между структурами и упрощать интеграцию приложений.

Составной ключ (*compound key*) — сочетание двух или более атрибутов, уникальным образом идентифицирующее экземпляр объекта. Примеры: полный абонентский номер телефона (код страны + код города/оператора + номер телефона); номер кредитной карты (ID эмитента + номер карты + CVV2/CVC2).

Композитный ключ (*composite key*) содержит один составной ключ плюс один или несколько простых и/или составных ключей или иных атрибутов. Примером композитного ключа является ключ для многомерной таблицы фактов, который может содержать несколько составных ключей, простых ключей и (опционально) метку времени загрузки данных.

1.3.3.3.2.2 Типы ключей в зависимости от функций

Суперключ (*super key*) — любой набор атрибутов, уникальным образом идентифицирующий экземпляр сущности. *Потенциальный ключ* (*candidate key*) — единственный атрибут или минимальный набор атрибутов (то есть простой или составной ключ), идентифицирующий экземпляр сущности. «Минимальный» означает, что никакое из подмножеств ключа не позволяет уникальным образом идентифицировать экземпляр сущности. Потенциальных ключей может иметься несколько, отсюда и название. Примеры потенциальных ключей сущности Клиент: e-mail, номер мобильного телефона, номер счета клиента. Потенциальные ключи могут использоваться в качестве бизнес-ключей (иногда их еще называют *естественными ключами — natural keys*). *Бизнес-ключ* (*business key*) — атрибут или набор атрибутов, по которому бизнесмен однозначным образом извлекает из базы данных нужную запись (экземпляр сущности). Суррогатные ключи не могут играть роль бизнес-ключей.

Первичный ключ (*primary key*) — это потенциальный ключ, выбранный в качестве фактического уникального идентификатора объекта. При наличии нескольких потенциальных ключей только один из них назначается первичным. *Альтернативный ключ* (*alternate key*) — любой из потенциальных ключей, не выбранных в качестве первичного. Альтернативный ключ тем не менее может использоваться на практике для поиска экземпляров сущностей. Часто в качестве первичного выбирается суррогатный ключ, а бизнес-ключи остаются альтернативными.

1.3.3.3.2.3 Идентифицирующие и неидентифицирующие связи

Независимой называется сущность, которая содержит первичный ключ и атрибуты, принадлежащие только ей. Зависимые же сущности используют внешние ключ(и) и/или атрибут(ы) из другой сущности или сущностей. В реляционных моделях данных в большинстве случаев принято схематически отображать независимые объекты строгими прямоугольниками, а зависимые — прямоугольниками со скругленными углами.

В следующем примере (рис. 38) Студент и Спецкурс — независимые сущности, а Регистрация — зависимая от них (дочерняя).

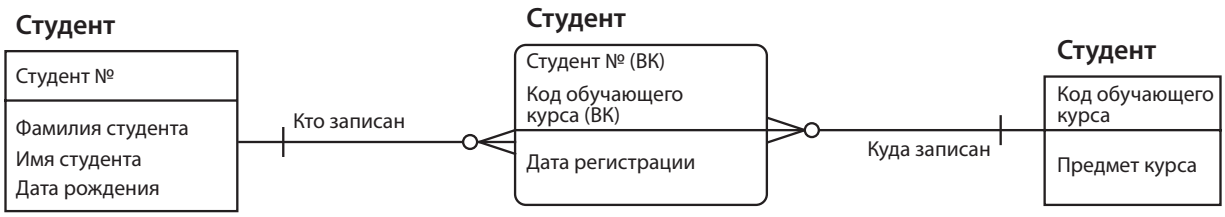


Рисунок 38. Зависимые и независимые сущности

Зависимые сущности имеют как минимум одну идентифицирующую связь, то есть такую, которая назначает первичный ключ родительской сущности внешним ключом дочерней сущности (в примере это перенос первичных ключей из сущностей Студент и Спецкурс в сущность

Регистрация). При определении неидентифицирующей связи первичный ключ из родительской сущности переносится в дочернюю в качестве одного из наследуемых внешних атрибутов, не играющих роли первичного ключа.

1.3.3.4 ДОМЕН

В моделировании данных *доменом* (*domain*) называется исчерпывающим образом описанный набор, диапазон или множество значений, которые могут быть присвоены атрибуту. Домен может быть задан различными способами (см. пункты из списка в конце этого раздела). Определение домена — одно из средств стандартизации характеристик атрибутов. Например, домен Дата, включающий все допустимые значения календарных дат, может задаваться для любого атрибута датировки в логической модели и для любых столбцов/полей дат в физической модели данных, таких как:

- Дата_приема_на_работу
- Дата_поступления_заказа
- Дата_рекламации
- Дата_начала_занятий

Допустимыми считаются все значения, относящиеся к домену атрибута, и только они. Массивы данных не должны содержать ни единого экземпляра сущности со значением какого-либо атрибута вне его домена. Например, домен атрибута Пол_сотрудника следует ограничить двумя допустимыми значениями (М и Ж). А вот домен атрибута Дата_приема_на_работу вполне можно задать как все действительные даты. При таком правиле исключено появление записей, датированных несуществующими календарными числами наподобие 30 февраля любого года.

Можно наложить на домен дополнительные *ограничения* (*constraints*) по формальным и/или логическим признакам. Например, в приведенном примере атрибут Дата_приема_на_работу логично ограничить разумным интервалом, завершающимся текущей датой, во избежание появления в базах данных (в результате опечаток или ошибок ввода данных) сотрудников, трудоустроившихся 10 марта 1899 или 2050 года, поскольку формально это действительные даты. В качестве дополнительной проверки можно ограничить область определения атрибута Дата_приема_на_работу стандартными рабочими днями отдела кадров (например, датами, приходящимися на календарные дни с понедельника по пятницу).

Домены атрибутов задаются наложением ограничений следующих видов.

- **Типы данных.** Домены могут задаваться через стандартные определения допустимых типов данных, например: целое число, текст (до 30 знаков) или дата — это сами по себе области допустимых значений.
- **Форматы данных.** Шаблоны или маски строго заданного формата (телефоны, почтовые индексы и т. п.) или с запретами на ввод определенных символов (только буквенно-цифровые, буквенно-цифровые плюс строго определенные служебные символы (как в адресах e-mail, к примеру) и т. д.) также, по сути, являются явными ограничениями области значений атрибутов сущностей.

- **Списки.** Фиксированные наборы возможных значений; такие домены всем хорошо знакомы по их реализации в интерфейсах в виде предложения опций из раскрывающегося списка или переключения статуса. Например, область определения атрибута Статус_заказа может быть ограничена предопределенными списочными значениями {Открыт, Комплектация, Доставка, Закрыт, Возврат}.
- **Допустимые интервалы.** Ограничения области значений атрибута минимальным и/или максимальным допустимыми значениями. При этом возможна взаимная обусловленность ограничений. Например, область значений атрибута Дата_выполнения_заказа может быть установлена в пределах 90 календарных дней относительно значения атрибута Дата_приема_заказа.
- **Реализация правил.** Ограничения допустимых значений атрибутов, накладываемые в силу установленных правил или в целях обеспечения их соблюдения. В частности, допустимые значения одного атрибута могут ставиться в зависимость от значений других атрибутов. Пример: значение атрибута Отпускная_цена не может быть ниже значения атрибута Себестоимость.

1.3.4 Схемы представления данных при моделировании

Шесть наиболее распространенных схем представления данных — *реляционная, многомерная, объектно-ориентированная, на основе фактов, хронологическая* и *NoSQL*. Для каждой схемы существуют собственные варианты формализованных систем условных обозначений (нотаций — notations) для построения диаграмм (табл. 9).

Таблица 9. Схемы и нотации моделирования

Схема	Примеры нотаций
Реляционная	Информационный инжиниринг (Information Engineering, IE) Описание интеграции для информационного моделирования (Integration Definition for Information Modeling, IDEF1X) Нотация Баркера Нотация Чена
Многомерная	Нотация для многомерного моделирования
Объектно-ориентированная	Унифицированный язык моделирования (UML)
На основе фактов	Объектно-ролевое моделирование (Object Role Modeling — ORM или ORM2) Полностью коммуникационно-ориентированное моделирование (Fully Communication Oriented Modeling, FCO-IM)
Хронологическая	«Свод данных» (Data Vault 1.0 и 2.0) Якорное моделирование (anchor modeling)
NoSQL	Документоориентированная Колоночная Графовая Ключ-значение

В этом разделе будут кратко рассмотрены каждая из этих схем и нотаций. Выбор схемы зависит отчасти от характера создаваемой базы данных, поскольку некоторые из них ориентированы на определенные технологии (табл. 10).

Реляционная схема позволяет строить модели данных всех трех уровней для реляционных систем управления базами данных (Relational Database Management System, RDBMS), однако для баз данных других типов поддерживает создание лишь концептуальной и логической моделей. То же самое касается и схемы на основе фактов. Многомерная схема позволяет строить полные трехуровневые модели как для RDBMS, так и для многомерных систем управления базами данных (Multidimensional Database Management System, MDBMS), а объектно-ориентированная — для RDBMS и объектных баз данных.

Хронологическая схема — это метод физического моделирования данных в первую очередь для построения хранилищ данных в среде. Схема NoSQL сильно зависит от лежащей в основе структуры базы данных (документы, графы или ключи-значения) и потому используется только для физического моделирования данных. Таблица 10 иллюстрирует некоторые важные моменты, в частности тот факт, что даже для нетрадиционных баз данных, например документоориентированных, концептуальная модель данных (Conceptual Data Model, CDM) и логическая модель данных (Logical Data Model, LDM) могут быть построены в соответствии с обычной реляционной схемой, а затем быть дополненными документоориентированной физической моделью данных (Physical Data Model, LDM).

Таблица 10. Сочетаемость схем и типов баз данных

Схема	RDBMS	MDBMS	Объектные базы данных	Документоориентированные	Колоночные	Графовые	Ключ–Значение
Реляционная	CDM LDM PDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM
Многомерная	CDM LDM PDM	CDM LDM PDM					
Объектно-ориентированная	CDM LDM PDM		CDM LDM PDM				

Схема	RDBMS	MDBMS	Объектные базы данных	Документоориентированные	Колоночные	Графовые	Ключ–Значение
Фактографическая	CDM LDM PDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM	CDM LDM
Хронологическая	PDM						
NoSQL			PDM	PDM	PDM	PDM	PDM

1.3.4.1 РЕЛЯЦИОННАЯ СХЕМА

Сформулированная в 1970 году Эдгаром Коддом[1] теория реляционных баз данных предоставляет систематизированную методику организации данных таким образом, чтобы они отражали свое назначение (Codd, 1970). Важным дополнительным эффектом такого подхода стало существенное снижение объемов памяти, занимаемых данными. Главной находкой Кодда стало определение возможности наиболее эффективного управления данными посредством их представления в виде двумерных таблиц — *отношений* (*relations*). Термин *отношение* позаимствован из математической теории множеств (см. главу 6).

Реляционная модель данных позволяет решить две задачи — точного отражения бизнес-данных и соблюдения хранения сведений о каждом факте только в одном месте (устранение избыточности данных). Реляционное моделирование идеально подходит для проектирования систем, обслуживающих операционную деятельность, которые требуют максимально быстрого ввода и безошибочного сохранения данных (Hay, 2011).

Существует несколько различных вариантов нотаций для отражения связи между сущностями в реляционном моделировании, включая информационный инжиниринг (Information Engineering, IE), описание интеграции для информационного моделирования (Integration Definition for Information Modeling, IDEF1X), нотация Баркера, нотация Чена.

Самая распространенная из них — нотация IE с уже знакомыми нам «вилками» (или «трезубцами») или «птичьими лапками», используемыми для отображения мощности связи (см. рис. 39).

[1] Эдгар Ф. Кодд (*англ.* Edgar Frank «Ted» Codd, 1923–2003) — англо-американский математик, создатель реляционной алгебры и построенной на ее основе модели данных, лауреат премии Тьюринга (1981). — *Примеч. пер.*

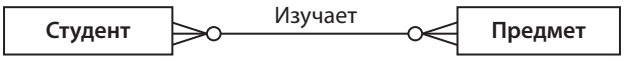


Рисунок 39. Нотация IE

1.3.4.2 МНОГОМЕРНАЯ СХЕМА

Концепция многомерного моделирования была разработана в 1960-х годах в рамках совместного проекта корпорации General Mills и Дартмутского колледжа[1]. В многомерных моделях данные структурированы таким образом, чтобы оптимизировать обработку запросов к базе данных и анализ при работе с большими объемами данных. В этом плане они контрастируют с моделями, которые используются в системах поддержки операционной деятельности, ориентированных на быструю обработку отдельных транзакций.

Многомерные модели позволяют фиксировать данные в проекции на различные аспекты рассматриваемого бизнес-процесса (бизнес-вопросы). На рисунке 40 приведен пример представления с помощью многомерной модели процесса Приема абитуриентов. В виде осей отражено распределение числа зачисленных абитуриентов по Географии проживания, Учебным заведениям, Календарным периодам обучения и статусу получения Пособий. Вдоль осей может осуществляться навигация (с целью получения ответов на вопросы): от Района до Страны, от Семестра до Учебного года, от Названия до Уровня учебного заведения.

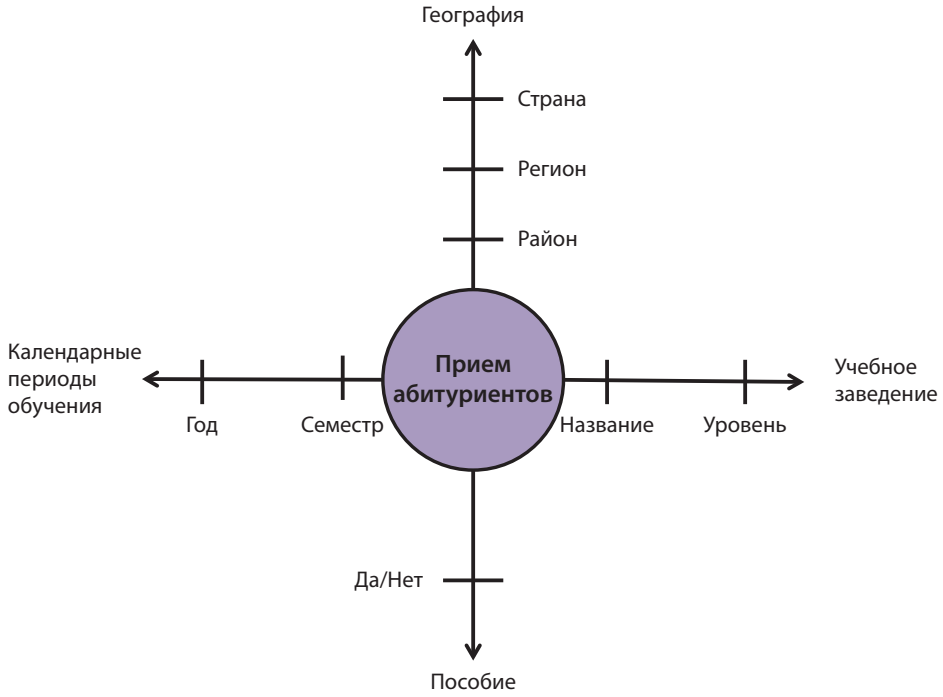


Рисунок 40. Осевая нотация для представления многомерных моделей

[1] http://bit.ly/2tsSP7w

Графическая нотация, использованная для построения этой модели, — «осевая нотация» — может быть очень эффективным инструментом коммуникации для тех, кто предпочитает не вчитываться в традиционный синтаксис моделирования данных.

Как реляционная, так и многомерная концептуальные модели данных могут быть основаны на одном и том же бизнес-процессе (как в вышеприведенном примере, где моделируется Прием абитуриентов). Разница между ними — в смысловом значении связей: в реляционной модели они отражают бизнес-правила, а в многомерной — навигационные пути к ответам на бизнес-вопросы.

1.3.4.2.1 Таблицы фактов

Центральное место в многомерной схеме занимает таблица фактов, строки которой являются числовыми и соответствуют значениям отдельных показателей, таких как объем, количество или численность. Значения некоторых показателей могут рассчитываться по алгоритмам, и в этом случае для их правильного понимания и использования критически важны метаданные. Таблицы фактов занимают основной объем базы данных (обычное эмпирическое правило оценки объема — около 90%) и имеют тенденцию к накоплению огромного числа строк.

1.3.4.2.2 Таблицы измерений

Таблицы измерений представляют информацию о важных для бизнеса объектах и содержат преимущественно текстовые описания. Измерения (dimensions) служат первоисточниками ограничивающих условий в запросах поиска данных («запрос по») или при формировании отчетов («отчет по»), выполняя роль точек входа или ссылок на данные в таблицах фактов. Данные в таблицах измерений, как правило, сильно денормализованы и составляют обычно около 10% от общего объема данных.

Каждая строка таблицы измерений должна иметь уникальный идентификатор. Два основных подхода к реализации этого требования — идентификация измерений с помощью суррогатных или естественных ключей.

Измерения имеют атрибуты, которые изменяются с различной скоростью. Для обеспечения контроля проведения изменений используется механизм медленно меняющихся измерений (Slowly Changing Dimensions, SCDs), поддерживающий изменения различных типов. Основными типами (иногда их называют ORC) являются следующие.

- **Перезапись (Overwrite) (тип 1).** Новое значение записывается поверх старого, замещая его.
- **Новая строка (Row) (тип 2).** Добавляется строка с новыми значениями, а старая помечается как неактуальная.
- **Новый столбец (Column) (тип 3).** В строке фигурирует фиксированное конечное множество значений, и при записи нового значения оно становится первым в ряду, при этом остальные значения смещаются на одну позицию, чтобы освободить место для нового значения, а последнее (самое старое) значение отбрасывается.

1.3.4.2.3 Схема снежинки

Наиболее простой схемой многомерной модели (схемой звезды) является плоская структура, состоящая из таблицы фактов в окружении таблиц измерений. Эта схема превращается в так называемую *схему снежинки* (*snowflake*), когда таблицы измерений в целях нормализации заменяются иерархическими или сетевыми структурами.

1.3.4.2.4 Уровень гранулированности

Под *уровнем гранулированности* (*grain*) подразумевается степень подробности (детализации) описания факта в строке таблицы фактов. Определение уровня гранулированности таблицы фактов — ключевой этап создания многомерной модели. Например, в многомерной модели, описывающей процесс записи студентов на обучающие курсы, уровень гранулированности можно определить так: студент, дата, курс.

1.3.4.2.5 Согласованные измерения

Согласованные измерения (conformed dimensions) строятся с целью обеспечения возможности их использования в масштабах всей организации, а не только в рамках конкретного проекта, что позволяет применять одни и те же измерения во множестве многомерных моделей благодаря согласованной терминологии и значениям. Например, если измерение Календарные периоды обучения является согласованным, многомерная модель, построенная для процесса учета поступающих в учебное заведение по Семестрам, будет содержать те же определения и значения Семестра, что и модель учета выпускников.

1.3.4.2.6 Согласованные факты

Согласованные факты (conformed facts) описываются с помощью стандартизованных определений терминов во всех отдельно взятых моделях. Различные категории бизнес-пользователей могут по-разному понимать один и тот же термин. «Прирост числа клиентов» может пониматься иначе, нежели «большой прирост» или «уточненный прирост». Разработчикам нужно это тонко чувствовать и не забывать: одно и то же название может в реальности относиться к совершенно разным понятиям в понимании различных подразделений организации, или, напротив, по-разному называемые на разных участках работы вещи на самом деле могут означать ровным счетом одно и то же.

1.3.4.3 ОБЪЕКТНО-ОРИЕНТИРОВАННОЕ МОДЕЛИРОВАНИЕ И ЯЗЫК UML

Унифицированный язык моделирования (Unified Modeling Language, UML) — графический язык, предназначенный для моделирования программного обеспечения. Одна из представленных в UML нотаций (модель классов) предназначена для моделирования баз данных. Модель определяет классы (типы сущностей) и взаимосвязи между ними, описывающие отношения различного типа (Blaha, 2013).

Рисунок 41 иллюстрирует характеристики модели классов UML.

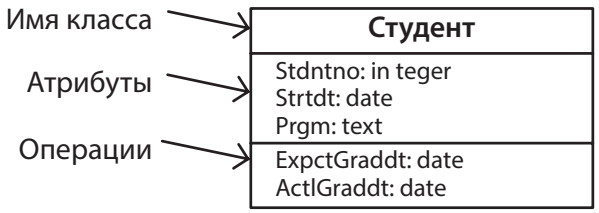


Рисунок 41. Модель классов UML

- Диаграмма классов строится аналогично классической ER-диаграмме (диаграмме «сущность-связь» — entity-relationship), но классы также включают раздел операций или методов, отсутствующий в диаграмме ER.
- Ближайшим аналогом операций в ER-моделировании являются хранимые процедуры (stored procedures).
- Типы атрибутов (дата, время в минутах и т. п.) указываются применительно к языку программирования, на котором реализуется приложение, а не в терминах физической модели базы данных.
- В нотации могут (опционально) отображаться значения по умолчанию.
- Доступ к данным осуществляется через открытый для класса интерфейс. Инкапсуляция (encapsulation) или возможность скрывать данные основаны на «эффекте локализации». Доступ к классу и соответствующим ему сущностям обеспечивается посредством операций.

Операции или методы класса называют также «поведением» (behavior). Поведение класса слабо соотносится с бизнес-логикой, поскольку нуждается в дополнительном определении последовательности выполнения действий и сроков. В терминах ER-моделирования поведение соответствует хранимым в таблице процедурам и триггерам (triggers).

Операции класса могут быть:

- открытыми (public): видимые извне;
- внутренними (internally visible): видимые только для дочерних объектов;
- частными (private): скрытые.

Для сравнения: физические ER-модели предлагают только открытый доступ ко всем данным со стороны процессов, запросов и операций.

1.3.4.4 МОДЕЛИРОВАНИЕ НА ОСНОВЕ ФАКТОВ (FBM)

Моделирование на основе фактов (Fact-Based Modeling, FBM) — общий подход, охватывающий семейство языков для концептуального моделирования, который зародился в конце 1970-х годов. Такие языки основаны на анализе естественной вербализации (выявлении стандартных наборов формулировок) деятельности, относящейся к сфере бизнеса. Языки моделирования на

основе фактов трактуют мир в терминах существующих объектов, фактов, относящихся к этим объектам или характеризующим их, и строго определенных ролей рассматриваемых объектов в происхождении каждого факта. При таком подходе обширная система мощных семантических ограничений опирается на гибкую автоматическую проверку вербализированных формулировок на предмет их совпадения с конкретными примерами. Основанные на фактах модели не предусматривают использования атрибутов, что значительно снижает потребность в интуитивных или экспертных суждениях, поскольку позволяет определять отношения между объектами напрямую (как между сущностями, так и между значениями). Самым распространенным вариантом FBM является логическая схема объектно-ролевого моделирования, предложенная в 1989 году Терри Халпином[1].

1.3.4.4.1 Объектно-ролевое моделирование (ORM или ORM2)

Объектно-ролевое моделирование (Object Role Modeling, ORM) — подход к проектированию на основе моделей, при котором проектирование начинается с изучения типичных примеров требуемой информации или запросов, которые классифицируются и обобщаются до понятных пользователям формулировок, а затем вербализируются на концептуальном уровне в терминах простых фактов, выражаемых на контролируемом естественном языке. Этот язык представляет собой ограниченную версию естественного языка, исключающую любую неоднозначность трактовок, с понятной людям семантикой. Являясь строго формализованным, он пригоден для автоматического отображения структур на более низкие уровни, обеспечивающие реализации (Halpin, 2015).

Рисунок 42 содержит иллюстративный пример модели ORM.

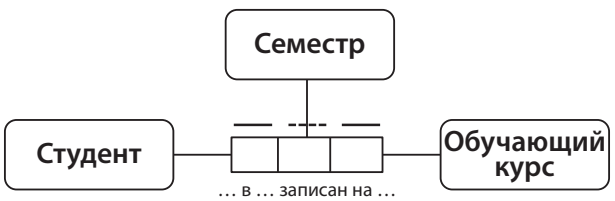


Рисунок 42. Модель ORM

1.3.4.4.2 Полностью коммуникационно-ориентированное моделирование (FCO-IM)

Полностью коммуникационно-ориентированное моделирование (Fully Communication Oriented Modeling, FCO-IM) по подходу и нотации весьма похоже на ORM. В приведенном примере (рис. 43) каждая из цифр является ссылкой на набор вербализированных фактов. Например, цифра 2 может отсылать к нескольким вербальным характеристикам студента, включая, например: «Студент 1234 носит имя Билл».

[1] Терри Халпин (*англ.* Terence Aidan «Terry» Halpin, p. 1950) — австралийско-американский специалист по информатике и программированию; описываемая логическая схема получила широкое распространение после ее реализации в 1994 г. в программном продукте Asymetrix InfoModeler. — *Примеч. пер.*

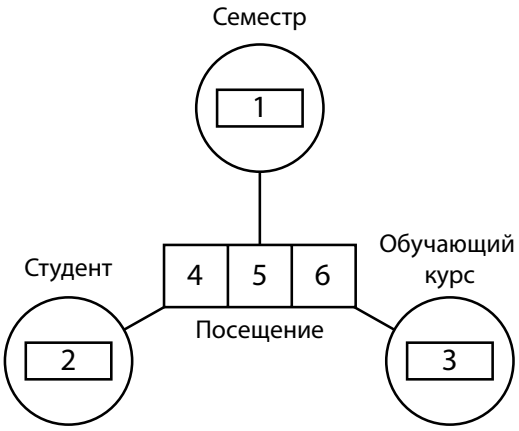


Рисунок 43. Модель FCO-IM

1.3.4.5 ХРОНОЛОГИЧЕСКИЕ СХЕМЫ

Хронологические (time-based) схемы представления данных используются при необходимости их упорядочения в хронологическом порядке с привязкой к конкретным моментам времени.

1.3.4.5.1 Метод моделирования Data Vault

Data Vault («свод данных») представляет собой набор уникальным образом связанных между собой нормализованных таблиц с детализированными и упорядоченными по времени данными, предназначенными для использования в различных функциональных областях бизнеса. В этой модели реализован гибридный подход, объединяющий лучшие свойства третьей нормальной формы (3NF, см. раздел 1.3.6) и звездообразной схемы. Модели Data Vault создаются с учетом специфических нужд хранилищ данных предприятия. В модели выделяются три типа сущностей: концентраторы (или хабы — hubs), связи (links) и сателлиты (satellites). Структура строится вокруг функциональных областей бизнеса, каждой из которых соответствует свой хаб, содержащий первичные ключи. Связи обеспечивают транзакционную интеграцию между хабами. Наконец, сателлиты предоставляют контекст для первичных ключей хаба (Linstedt, 2012).

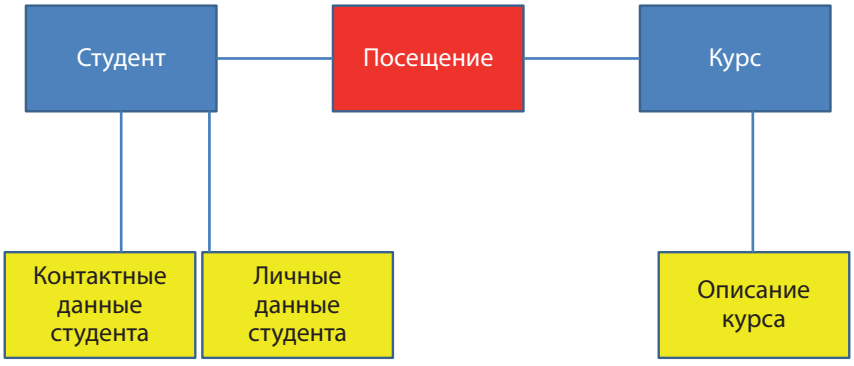


Рисунок 44. Модель Data Vault

В приведенном примере (рис. 44) Студент и Курс — хабы, представляющие с помощью ключей основные понятия в рассматриваемой предметной области; Посещение — связь между хабами; Контактные данные студента, Личные данные студента и Описание курса — сателлиты, содержащие описательную информацию для понятий, представленных в хабах, и, кроме того, обеспечивающие возможность ведения различных типов данных об истории.

1.3.4.5.2 Анкерное моделирование

Анкерное (якорное) моделирование (anchor modeling) хорошо подходит для проектирования динамических баз данных, информация в которых со временем меняется не только по содержанию, но и по структуре. Этот подход предоставляет графическую нотацию для концептуального моделирования, отчасти похожую на традиционную, но с расширениями, которые позволяют работать с данными, изменяющимися с течением времени. Четыре основных понятия, используемых в анкерном моделировании, — якоря, атрибуты, связи (ties) и узлы (knots). Якоря соответствуют сущностям и событиям, атрибуты моделируют свойства якорей, связи — отношения между якорями, а узлы — общие свойства, такие как состояния.

В примере анкерной модели на рисунке 45: Студент, Курс и Посещение — якоря; серые ромбы — связи; кружки — атрибуты.

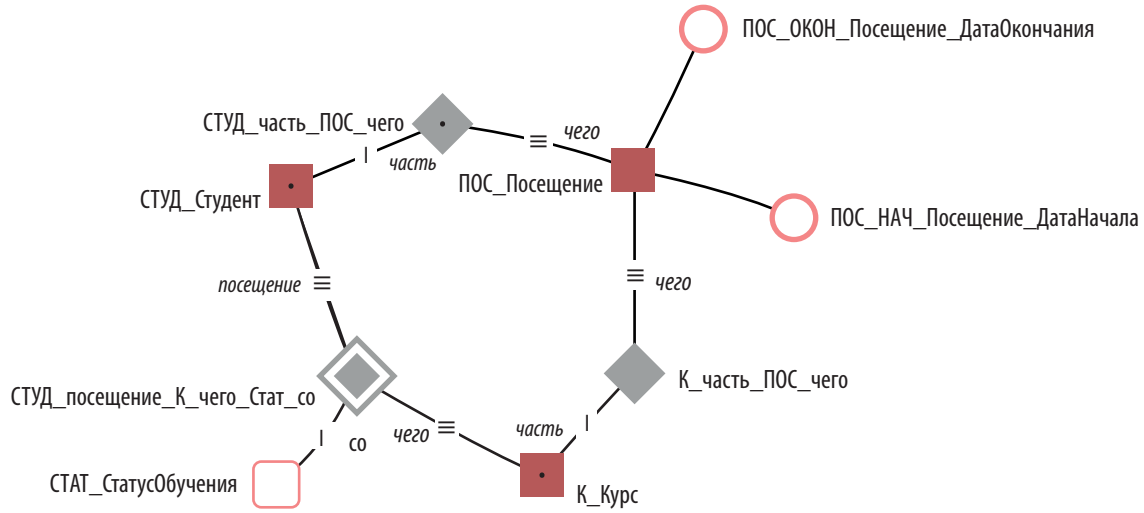


Рисунок 45. Анкерная модель

1.3.4.6 БАЗЫ ДАННЫХ NOSQL

Акронимом NoSQL называют категорию баз данных, созданных на основе нереляционных технологий. Название NoSQL нравится не всем, поскольку в реальности речь идет не столько о том, как организуются запросы к базе данных (без использования языка SQL), сколько о том, как данные хранятся (без использования реляционных структур). Четыре основных типа баз данных NoSQL: документоориентированная, ключ-значение, колоночная и графовая.

1.3.4.6.1 Документоориентированная база данных

Совокупность данных о рассматриваемой предметной области бизнеса не разбивается на множество реляционных структур, а часто представляется в виде набора одиночных структур, называемых *документами*. Например, вместо создания реляционных структур Студент, Обучающий курс и Семестр все относящиеся к ним данные могут быть объединены с помощью создания единого документа Регистрация.

1.3.4.6.2 База данных «ключ-значение»

База данных «ключ-значение» позволяет приложениям сохранять свои данные в двух колонках («ключ» и «значение»), при этом «значения», в зависимости от приложения, могут быть как простейшими по структуре (например, даты, числа, коды), так и представлять собой сложные информационные объекты (неформатированный текст, видео- и аудиоинформация, документы, фотографии и т. п.).

1.3.4.6.3 Колоночная база данных

Из четырех типов баз данных NoSQL колоночные (column-oriented) базы данных наиболее близки к реляционным. Данные в этой модели также представлены в виде таблиц с колонками и строками, однако, в отличие от реляционных систем управления базами данных (СУБД), работающих с предопределенной структурой и простейшими типами данных (числами, датами и т. п.), колоночные СУБД, такие как Cassandra, поддерживают работу с более сложными типами данных, включая неформатированный текст и изображения. Кроме того, в таких базах данных каждая колонка хранится как отдельная структура.

1.3.4.6.4 Графовая база данных

Графовые базы данных предназначены для хранения данных, структуру которых удобно представлять в виде множества узлов (вершин графа) с неопределенным числом попарных связей между ними (ребер). Примеры процессов и систем, идеально моделируемых с помощью графов, включают социальные отношения (где узлы — отдельные индивидуумы), схемы маршрутов общественного транспорта (узлы — станции или остановки), дорожные карты (узлы — перекрестки и развязки). Графовые СУБД максимально упрощают выполнение запросов на минимальное по числу ребер соединение двух вершин или нахождение ближайшего узла, соответствующего критерию запроса, что позволяет без труда реализовывать такие функции, как поиск кратчайшего маршрута, ближайшей АЗС и т. п., требующие громоздких и затратных по времени алгоритмов в традиционных реляционных СУБД. Примеры СУБД на основе графов включают Neo4J, Allegro и Virtuoso.

1.3.5 Уровни детализации модели данных

В 1975 году Комитет по планированию стандартов Американского национального института стандартов (ANSI/SPARC) опубликовал трехуровневую модель архитектуры систем управления данными. К архитектурным уровням в рамках архитектуры ANSI/SPARC относятся следующие.

- **Концептуальный уровень.** Наиболее полное представление о функционировании предприятия, для которого создается база данных, в реальных условиях. Отражает текущее понимание «лучших практик» или «оптимальной работы» предприятия.
- **Внешний уровень.** Различным категориям пользователей базы данных открывается доступ к подмножествам данных, описываемых некоторым частичным набором компонентов общего представления, которые нужны пользователям каждой категории. Эти подмножества данных и субкомпонентов модели отображаются в виде «внешних схем».
- **Внутренний уровень.** «Машинный ракурс» рассмотрения данных, описываемый «внутренней схемой» распределения информации предприятия по хранилищам (Hay, 2011).

Трем этим архитектурным уровням обычно соответствуют концептуальный, логический и физический уровни детализации модели. На практике концептуальное и логическое моделирование данных относятся к категории работ по планированию и анализу требований, а физическое моделирование данных — к проектным работам. В настоящем разделе представлены обзоры содержания работ по концептуальному, логическому и физическому моделированию. Кроме того, каждый из трех уровней проиллюстрирован на примерах двух схем СУБД — реляционной и многомерной.

1.3.5.1 КОНЦЕПТУАЛЬНАЯ МОДЕЛЬ

Концептуальная модель данных (Conceptual Data Model, CDM) фиксирует высокоуровневые требования к данным как к набору взаимосвязанных понятий. Она содержит только базовые и критически важные для бизнеса сущности в рассматриваемой функциональной области с описанием каждой сущности и связей между ними.

Например, если мы моделируем связи между студентами и учебными заведениями, то в рамках реляционной концептуальной модели они могут быть представлены (в стандартной нотации IE), например, следующим образом (рис. 46).

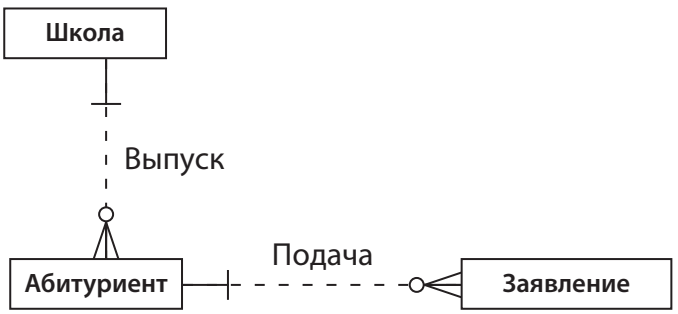


Рисунок 46. Реляционная концептуальная модель данных

В числе Абитуриентов может присутствовать произвольное число выпускников каждой Школы, при этом каждый Абитуриент является выпускником строго одной Школы. Кроме того,

каждый Абитуриент может подать произвольное число Заявлений в вузы, но каждое Заявление подается строго одним Абитуриентом.

В реляционной модели данных бизнес-правила отражаются типами линий связи. В приведенном примере отдельно взятый абитуриент, например Иванов, может являться выпускником любой из школ, занесенных в базу данных, но не может являться выпускником одновременно двух или более школ. Кроме того, Иванов может подать много заявлений в разные вузы или на разные факультеты или вовсе не подавать заявления, но каждое заявление подается строго одним абитуриентом, а не нулем или двумя.

А теперь вспомним рисунок 40, воспроизведенный ниже в уменьшенном виде (рис. 47). На нем с использованием осевой нотации представлена концептуальная четырехмерная модель Приема абитуриентов в учебные заведения различного уровня.

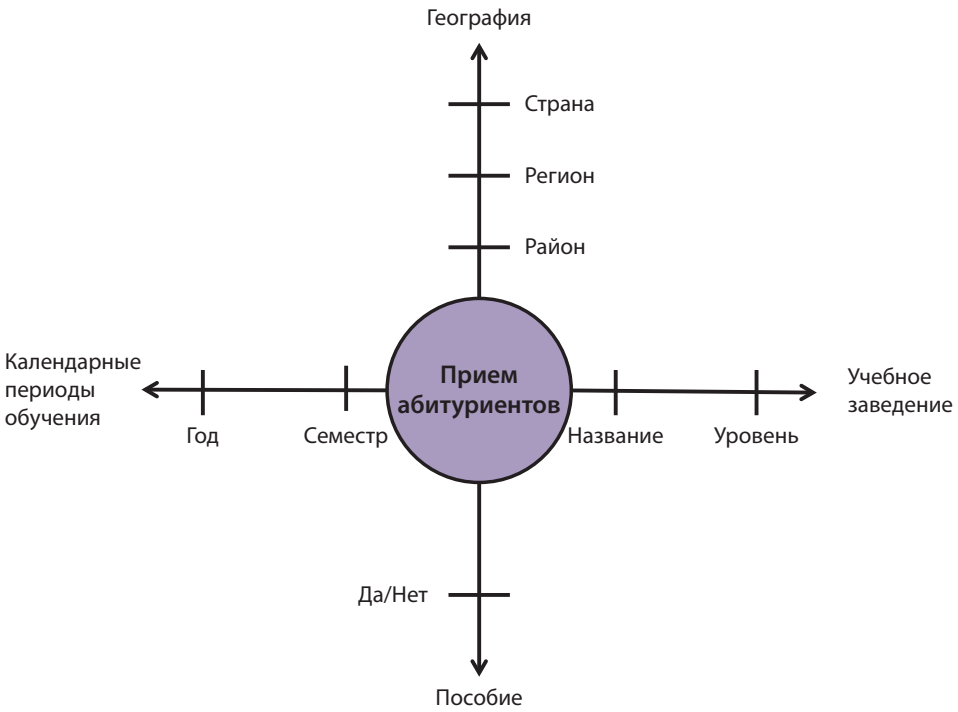


Рисунок 47. Многомерная концептуальная модель данных

1.3.5.2 ЛОГИЧЕСКАЯ МОДЕЛЬ

Логическая модель данных (Logical Data Model, CDM) детально отражает требования к данным, обычно в контексте их конкретного применения — например, с точки зрения потребностей в данных пользовательских приложений. На логическом уровне модель данных всё еще независима от каких-либо технологических ограничений, которые возникают и учитываются лишь на стадии реализации. Обычно логическая модель, по крайней мере поначалу, строится как детализирующее расширение концептуальной модели данных.

В реляционных схемах логическая модель данных строится путем добавления атрибутов к объектам концептуальной модели. Атрибуты присваиваются в процессе нормализации (см. раздел 1.3.6). Каждый атрибут строго привязан к первичному ключу таблицы объекта, в которой находится (см. рис. 48). Например, атрибут Название школы строго привязан к Коду школы. То есть первичному ключу Код школы не может соответствовать более одного значения атрибута Название школы.

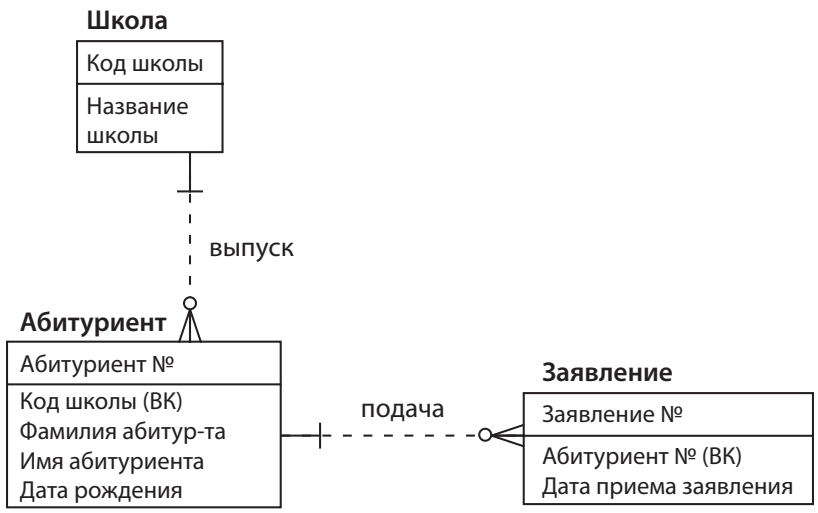


Рисунок 48. Реляционная логическая модель данных

Многомерная логическая модель данных, как видно из примера (рис. 49 на следующей странице), во многих случаях представляет собой дополненную набором атрибутов концептуальную модель. В отличие от логической модели реляционной базы данных, фиксирующей бизнес-правила применительно к бизнес-процессам, логическая модель многомерной базы данных фиксирует показатели здоровья и эффективности бизнес-процессов.

Всего зачислено — счетчик показателя, отвечающего на главный бизнес-вопрос, который располагается в центре координат в виде сущности Прием абитуриентов. По осям вокруг него в порядке нарастания детализации распределены контекстные сущности, позволяющие рассматривать показатель числа зачислений на разных уровнях гранулированности: например, Семестр или Год.

1.3.5.3 ФИЗИЧЕСКАЯ МОДЕЛЬ

Физическая модель данных (Physical Data Model, PDM) отражает детализированное техническое решение, за основу которого обычно берется логическая модель данных, а затем доводится до состояния полной совместимости с комплексом аппаратного и программного обеспечения и сетевого оборудования. Физические модели данных разрабатываются в расчете на конкретные технологии. Реляционные базы данных, например, проектируются с учетом функциональной

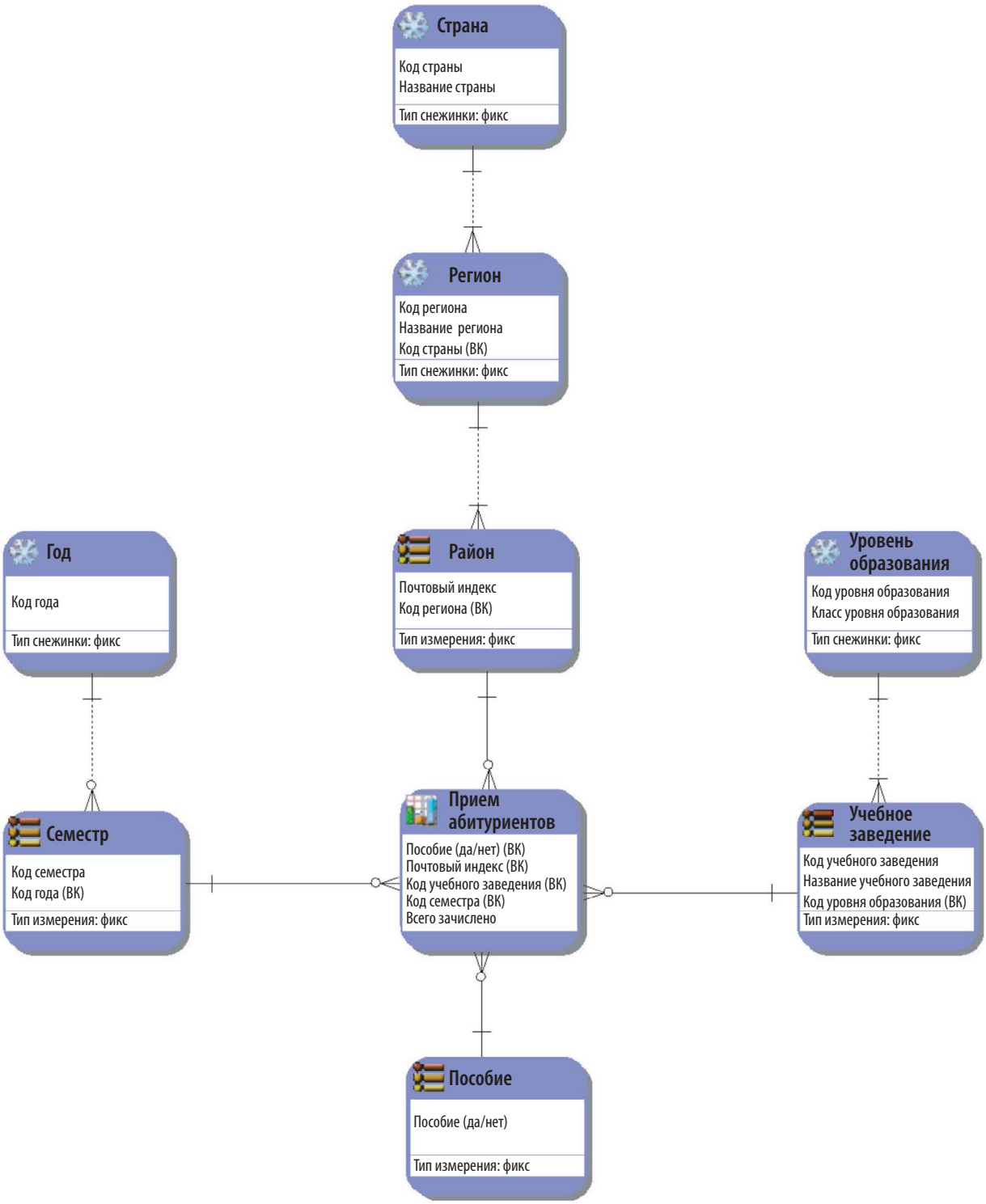


Рисунок 49. Многомерная логическая модель данных

специфики СУБД, которую планируется использовать (IBM DB2, UDB, Oracle, Teradata, Sybase, Microsoft SQL Server или Microsoft Access).

Рисунок 50 содержит пример реляционной физической модели данных. Обратите внимание на денормализацию: логический объект Школа из структуры связей удален посредством включения данных о школе в объект Абитуриент в качестве атрибутов. Поскольку сделано это умышленно, можно предположить, что на практике при любом запросе данных об абитуриенте запрашиваются и данные о школе, которую он окончил. Следовательно, хранение записей о школе в таблице Абитуриент способствует повышению производительности СУБД (ускорению обработки запросов) по сравнению с двухтабличной реализацией.

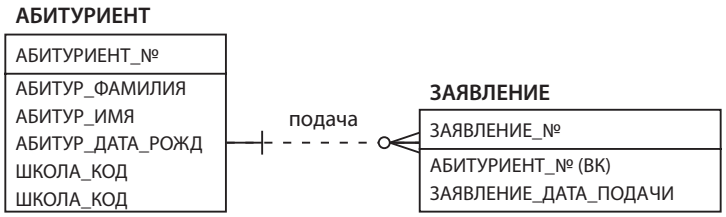


Рисунок 50. Реляционная физическая модель данных

Поскольку физическая модель данных строится с учетом технологических ограничений, объединение связанных структур (денормализация) с целью повышения производительности используется достаточно часто.

Рисунок 51 иллюстрирует денормализованную многомерную физическую модель данных (обычно на физическом уровне «снежинка» свертывается до простой звездообразной схемы с одной таблицей по каждому измерению).

Как и в реляционной модели, на физическом уровне структура данных модифицирована по сравнению с логической моделью в целях обеспечения максимальной производительности (с учетом конкретной технологии) и ускорения получения ответов на бизнес-вопросы.

1.3.5.3.1 Каноническая модель

Разновидностью физической модели данных является так называемая каноническая модель, которая используется для обеспечения совместимости различных систем с точки зрения обмена данными. В рамках этой модели структура данных, передаваемых из системы в систему, описывается на уровне пакетов или сообщений. При обмене данными через веб-сервисы, корпоративную сервисную шину (Enterprise Service Bus, ESB) или средства интеграции корпоративных приложений (Enterprise Application Integration, EAI) каноническая модель описывает, какую структуру данных должны использовать отправляющие и принимающие сервисы. Структура должна быть предельно неспецифичной, чтобы обеспечить повторное использование и упростить требования к интерфейсу.

Такая структура может быть реализована в виде буфера или очереди на основе системы обмена сообщениями (промежуточное программное обеспечение — middleware), с целью временного сохранения содержимого сообщений.

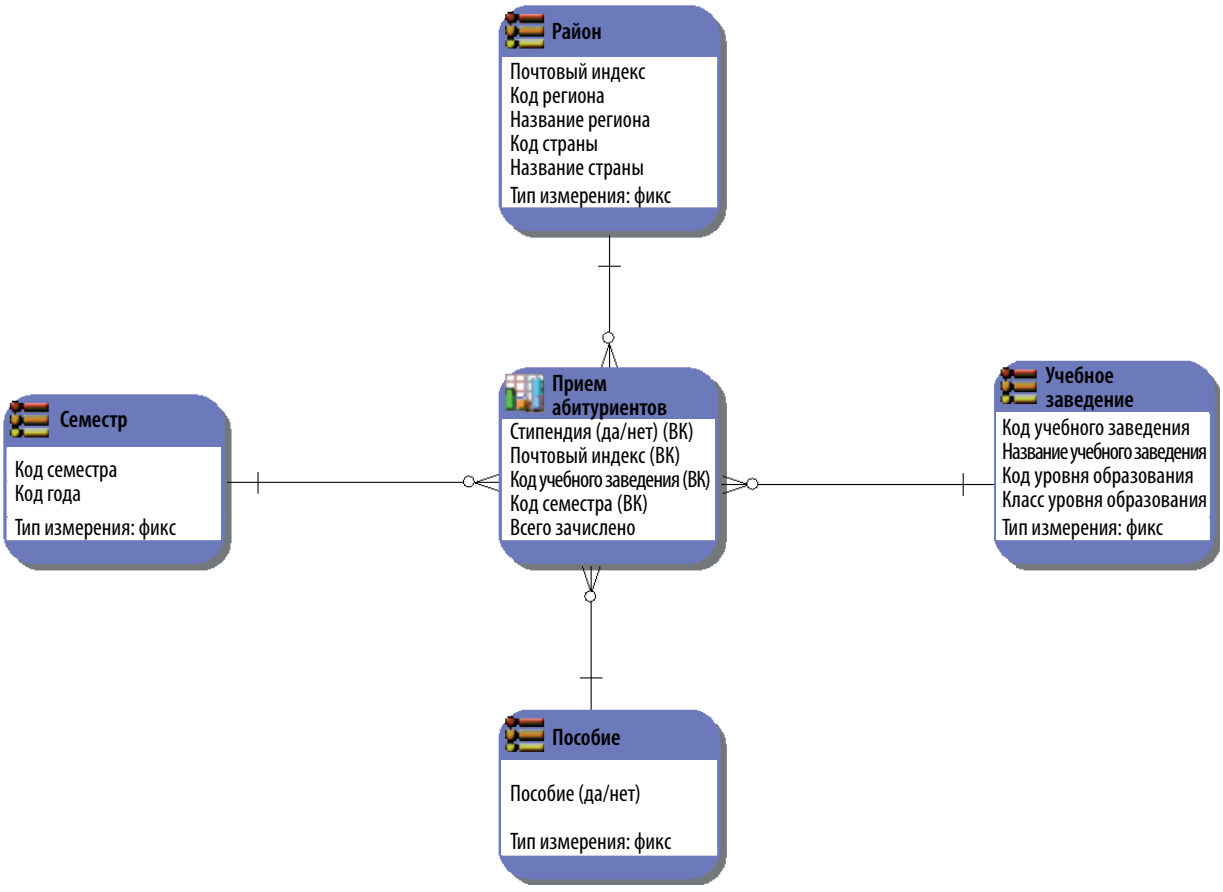


Рисунок 51. Многомерная физическая модель данных

1.3.5.3.2 Представления

Представлением (view) называют виртуальную таблицу, служащую средством просмотра данных из одной или многих таблиц, содержащих фактические атрибуты и/или ссылки на них. Обычное представление в процессе его использования запускает SQL-запросы к базе данных, обеспечивающие получение и отображение текущих значений входящих в представление атрибутов. Материализованные (materialized) представления сохраняют у себя итоговые результаты запросов и автоматически их обновляют, значительно увеличивая скорость выдачи данных. Представления используются для упрощения структуры запросов, контроля доступа к данным и переименования колонок без риска нарушить ссылочную целостность из-за нарушения нормализации.

1.3.5.3.3 Секционирование

Секционированием (partitioning) называют разделение таблицы с целью упрощения архивирования или ускорения извлечения данных. Таблицы могут разделяться вертикально (по столбцам) или горизонтально (по строкам).

- **Вертикальное разделение.** Используется для ускорения обработки запросов за счет уменьшения объема просматриваемых данных. Например, таблица с данными о клиентах разделяется на две подтаблицы, одна из которых содержит колонки со статическими и редко изменяемыми данными, а другая — колонки с данными, которые регулярно изменяются (для ускорения загрузки и индексирования). Кроме того, можно вынести в отдельную подтаблицу столбцы с часто запрашиваемыми данными (для ускорения сканирования таблицы).
- **Горизонтальное разделение.** Позволяет ускорять обработку запросов за счет выделения в подтаблицы подмножеств строк, используя значения в какой-либо колонке в качестве дифференцирующего параметра. Таким способом можно, например, создавать отдельные таблицы клиентов по регионам.

1.3.5.3.4 Денормализация

Денормализацией (*denormalization*) называют намеренное внесение в физические таблицы, создаваемые на основе нормализованной логической модели, избыточных или дублирующих друг друга полей данных. Иными словами, денормализация — умышленное размещение одного и того же атрибута в двух или более местах. Причины для этого могут иметься разные. Но первая и самая распространенная заключается в намерении повысить производительность за счет использования следующих приемов.

- Предварительная заготовка сводных таблиц данных из множества других таблиц во избежание необходимости затратного по времени и аппаратным ресурсам многократного повторного объединения одних и тех же данных во время выполнения приложения.
- Создание предварительно отфильтрованных выборочных копий данных с целью снижения затрат времени на операционные расчеты и/или сканирование больших таблиц.
- Предварительное проведение и сохранение результатов ресурсоемких расчетов с использованием базовых данных с целью снижения конкуренции процессов за системные ресурсы во время выполнения.

Денормализация также может использоваться для обеспечения безопасности пользователей и защиты данных посредством их разделения на множественные представления или создания отдельных копий таблиц в зависимости от практических целей доступа.

С риском ошибок, обусловленных дублированием данных, такой подход не сопряжен. Поэтому денормализация часто выбирается в качестве альтернативного решения в тех ситуациях, когда не удается эффективно реализовать представления или секционирование таблиц. При этом желательно внедрить систему обязательной проверки качества данных денормализованных атрибутов и их корректного сохранения. В общем случае денормализацию рекомендуется использовать исключительно в целях оптимизации обработки запросов к базам данных или обеспечения безопасности пользователей.

Хотя термин *денормализация* используется в настоящем разделе, это не означает, что аналогичные процедуры неприменимы к другим моделям (не являющимся реляционными). Вполне можно денормализовать документоориентированную базу данных, просто процедура будет называться как-то иначе — например, *встраивание* (*embedding*).

В многомерном моделировании данных денормализация называется свертыванием (collapsing) или объединением (combining). Если по каждому из измерений данные свернуты в единую структуру, получается модель данных, построенная по *схеме звезды* (см. рис. 51). Если данные хотя бы по одному измерению оставлены в развернутом виде, получается модель данных, построенная по *схеме снежинки* (см. рис. 49).

1.3.6 Нормализация

Нормализация (*normalization*) заключается в применении наборов правил, позволяющих упорядочить всё разнообразие необходимых для ведения бизнеса данных в стабильные структуры. Главная цель оптимизации, говоря простым языком, — сделать так, чтобы каждый атрибут содержался строго в одном месте во избежание избыточности данных и, как следствие, их возможной противоречивости. Для проведения нормализации требуется глубокое понимание каждого атрибута и его отношения к первичному ключу.

Правила нормализации разделяют и организуют атрибуты в соответствии с первичными и внешними ключами. Правила последовательно распределяются по уровням, и на каждом следующем уровне повышается степень детализации и добавляются новые требования по учету специфики сущностей при подборе корректных первичных и внешних ключей. Каждому уровню соответствует отдельная нормальная форма (normal form). При переходе к следующей нормальной форме свойства предыдущих нормальных форм сохраняются. Уровни нормализации определяются следующим образом.

- **Первая нормальная форма (1NF).** Обеспечивает наличие у каждой сущности корректного первичного ключа и зависимость каждого атрибута от первичного ключа, отсутствие повторяющихся групп, однозначность и атомарность атрибутов (то есть отсутствие скрытых атрибутов внутри каждого атрибута). Приведение к 1NF включает разрешение связей «многие-ко-многим» с помощью дополнительных сущностей, часто называемых ассоциативными (associative entity).
- **Вторая нормальная форма (2NF).** Обеспечивает минимизацию полного первичного ключа каждой сущности (устранение, по возможности, составных ключей) и зависимость каждого атрибута только от полного ключа (отсутствие зависимости от части ключа) в случае сохранения составных ключей.
- **Третья нормальная форма (3NF).** Обеспечивает отсутствие у сущностей скрытых (неявных) первичных ключей и отсутствие у каждого атрибута зависимости от любых неключевых атрибутов объекта (правило «атрибут зависит от ключа, полного ключа и только ключа»).

- **Нормальная форма Бойса — Кодда (BCNF).** Разрешает проблему функциональной зависимости потенциальных составных ключей. К «потенциальным составным ключам» относятся все возможные альтернативные составные ключи, включая первичный ключ, если он не является простым, а под «функциональной зависимостью» понимается наличие каких-либо скрытых бизнес-правил, связывающих эти ключи.
- **Четвертая нормальная форма (4NF).** Устраняет все попарные связи «многие-ко-многим» между атрибутами за счет дальнейшей декомпозиции до предельно возможного уровня детализации атрибутов.
- **Пятая нормальная форма (5NF).** Детализирует все зависимости внутри всех сущностей до уровня базовых попарных зависимостей атрибутов от компонентов первичных ключей.

Под *нормализованной моделью* обычно понимают данные, приведенные в форму 3NF. Ситуации, требующие нормализации до уровней BCNF, 4NF и 5NF, на практике встречаются редко.

1.3.7 Абстрагирование

Абстрагированием называется удаление из модели излишних деталей, с тем чтобы ее можно было применять к максимально широкому классу ситуаций, при сохранении всех важных свойств, проистекающих от природы, концепции и/или предметов модели. Примером абстрагирования является использование в структуре модели обобщенных категорий, таких как Участник или Роль, которые можно использовать для моделирования всевозможных функций и взаимодействий людей и организаций (под ними могут с равным успехом, в зависимости от контекста, пониматься, например, партнеры, сотрудники, клиенты и т. п.). Не все разработчики моделей данных считают нужным оперировать абстрактными категориями, да и не все на это способны. Еще на концептуальном уровне нужно взвесить все «за» и «против», сопоставив издержки разработки и ведения абстрактных структур с потенциальными трудозатратами на переделку излишне конкретизированной модели в будущем, когда, возможно, понадобится ее модифицировать под другое применение (Giles, 2011).

Абстрагирование включает *обобщение* и *специализацию*. При обобщении общие атрибуты сущностей группируются в виде сущности *супертипа*; при специализации выявляются характерные отличительные признаки сущности и определяются *подтипы*. Выявление отличительных признаков обычно происходит в процессе анализа экземпляров сущностей, которые могут быть отнесены к тому или иному подтипу в зависимости от значений отдельных атрибутов.

Подтипы также могут создаваться посредством использования *ролей* или *классификации* с целью разделения экземпляров сущности на группы по функциям. Пример: сущность Участник может иметь подтипы Физическое лицо и Юридическое лицо.

Выделение подтипов подразумевает, что подтип наследует все свойства супертипа. В приведенном примере реляционной модели (рис. 52) Университет и Средняя школа — подтипы супертипа Учебное заведение.

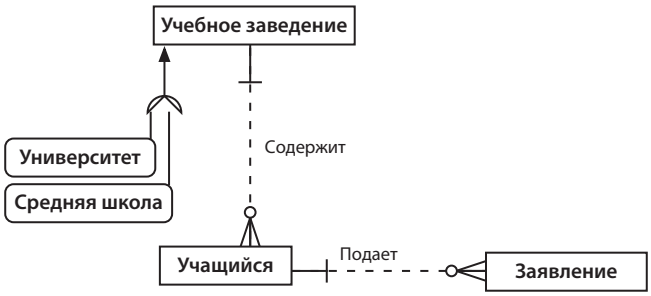


Рисунок 52. Отношения между супертипом и подтипами

Использование подтипов снижает избыточность модели и способствует лучшему пониманию общности характеристик сущностей, которые внешне могут представляться качественно различными.

2. ПРОВОДИМЫЕ РАБОТЫ

В этом разделе будут кратко описаны этапы построения концептуальной, логической и физической моделей данных, а также их сопровождения и пересмотра. Кроме того, рассматриваются основные аспекты прямого и обратного проектирования.

2.1 План проведения работ по моделированию данных

План проведения работ по моделированию данных должен предусматривать решение таких задач, как оценка требований организации, разработка стандартов и определение места хранения моделей.

Результаты процесса моделирования данных делятся на следующие виды.

- **Диаграмма.** Модель данных содержит одну или несколько диаграмм. Требования на диаграмме должны быть отражены в точной форме. Указывается выбранный уровень детализации модели (концептуальная, логическая или физическая), схема (реляционная, многомерная, объектно-ориентированная, на основе фактов, хронологическая или NoSQL) и нотация (IE, UML, объектно-ролевая модель и т. п.).
- **Определения.** Определения сущностей, свойств/атрибутов и отношений/связей в полном объеме, необходимом для обеспечения точной реализации модели данных.
- **Проблемные и нерешенные вопросы.** Часто в процессе моделирования данных возникают вопросы и проблемы, не поддающиеся решению на фазе моделирования. Кроме того, зачастую ответы или решения зависят не от проектировщиков модели данных, а от иных структурных или функциональных подразделений. Поэтому часто подготавливается специальный документ со списком нерешенных вопросов. В примере из сферы образования нерешенный вопрос может формулироваться, например, следующим образом: «Если Студент восстанавливается

после академического отпуска, как его или ее учитывать — с прежним или новым первичным ключом Студент №?».

- **Происхождение.** На физическом, а иногда и на логическом уровне моделирования бывает важно знать, откуда именно берутся данные. Часто достаточно формального представления отображения «источник/получатель», фиксирующего попарное соотнесение атрибутов в системе-источнике и системе-получателе. Также происхождение может отслеживать переход компонентов модели с концептуального уровня на логический и с логического на физический в рамках одного проекта. Имеются как минимум две веские причины, по которым следует обязательно фиксировать происхождение при моделировании. Во-первых, разработчик модели данных получает очень хорошее понимание требований к данным и, как следствие, оказывается в наилучшей позиции для определения атрибутов исходных данных. Во-вторых, определение атрибутов исходных данных может послужить эффективным средством проверки точности модели и правильности отображения (то есть проверки реалистичности модели).

2.2 Построение модели данных

Проектируя модели, разработчики часто опираются на богатый практический опыт анализа и моделирования данных — как собственный, так и накопленный коллегами. Они могут изучать существующие модели и базы данных, выверять свои действия по опубликованным стандартам, рассматривать и учитывать всевозможные требования к данным. Изучив все входные материалы подобного рода, проектировщики приступают непосредственно к построению модели. Моделирование — в очень большой мере процесс итерационный (см. рис. 53). Подготовив проект модели, разработчики обращаются к бизнес-профессионалам и аналитикам за разъяснением реальных условий и бизнес-правил. Затем, доработав и уточнив модель, они формулируют новые вопросы и снова обращаются за ответами на них к практикам и аналитикам (Hoberman, 2014).

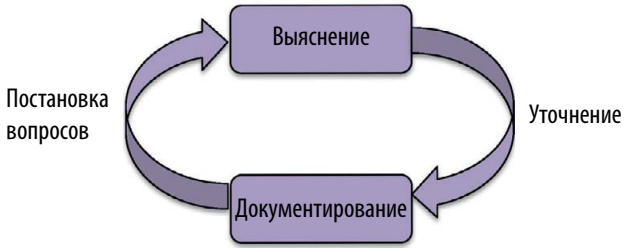


Рисунок 53. Моделирование как итерационный процесс

2.2.1 Прямое проектирование

Прямое проектирование (forward engineering) представляет собой процесс построения нового приложения, начиная с выяснения предъявляемых к нему требований. Сначала создается CMD, чтобы понять границы и состав предстоящих работ, выработать и согласовать ключевую терминологию. Затем создается LMD, документирующая бизнес-решение, и наконец — PMD, документирующая техническое решение.

2.2.1.1 КОНЦЕПТУАЛЬНОЕ МОДЕЛИРОВАНИЕ ДАННЫХ

Создание CMD включает следующие этапы.

- **Выбор схемы.** Решите, в соответствии с какой схемой — реляционной, многомерной, на основе фактов, хронологической или NoSQL — будет строиться модель данных. Руководствуйтесь описанными выше особенностями и критериями выбора различных схем (см. раздел 1.3.4).
- **Выбор нотации.** Выбрав схему, выберите нотацию для ее формального отображения, например IE или объектно-ролевую. Выбор нотации зависит от стандартов организации и знакомства пользователей с конкретными методиками.
- **Создание исходной CMD.** Первоначальная версия CMD должна отражать точку зрения некой группы пользователей. Это позволит избежать излишних осложнений и задержки процесса проектирования из-за необходимости согласовывать все позиции (других подразделений или организации в целом).
  - Определите совокупность самых высокоуровневых понятий, которыми оперирует организация (существительные). Самые распространенные концептуальные категории отражают Дату/Время, Географию (местонахождения/адреса), Клиентов/Участников, Продукты/Услуги и Транзакции.
  - Определите совокупность действий, связывающих эти понятия (глаголы). Отношения (связи) могут быть односторонними, двусторонними (обоюдными) или многосторонними, то есть связывающими более двух понятийных категорий. Примеры: одному Клиенту может соответствовать несколько Местонахождений (по домашнему адресу, месту работы и т. д.); а одному и тому же Местонахождению — много Клиентов. Транзакции соответствуют Время и Функция, а Клиенту при продаже — Продукт или Услуга.
- **Учет корпоративной терминологии.** Зафиксировав представление пользователей в виде прямоугольников и линий, проектировщик модели должен взглянуть на результат с позиции организации в целом и удостовериться в соответствии терминологии модели корпоративной терминологии и правилам. Могут потребоваться некоторые корректировки, если, например, в концептуальной модели, составленной со слов целевой аудитории, сущность названа Клиентом, а на уровне предприятия в целом и согласно его документации то же самое понятие имеет название Заказчик.
- **Окончательное согласование.** По завершении разработки CMD обязательно представьте модель на рецензирование, дабы удостовериться, что она соответствует лучшим практикам моделирования и вполне удовлетворяет предъявляемым требованиям. Обычно достаточно бывает подтверждения точности модели по электронной почте.

2.2.1.2 ЛОГИЧЕСКОЕ МОДЕЛИРОВАНИЕ ДАННЫХ

В логической модели данных находят отражение детализированные требования к данным, предусмотренные CMD.

2.2.1.2.1 Анализ информационных потребностей

Для определения информационных потребностей необходимо прежде всего выявить, какие именно данные нужны бизнесу, в контексте одного или нескольких бизнес-процессов. На входе любого бизнес-процесса могут использоваться информационные продукты других бизнес-процессов, а результаты обработки данных передаваться в следующие. Названия всех информационных продуктов в таких цепочках часто отражают важнейшие понятия из бизнес-словаря, которые следует идентифицировать и брать за основу терминологии, используемой в моделировании данных. Вне зависимости от того, моделируются ли данные и процессы последовательно (в любом порядке) или параллельно, для эффективного их анализа и проектирования модели требуется достаточно сбалансированное представление о данных (существительные) и процедурах (глаголы), в равной мере соответствующее реалиям процессов и требованиям моделирования.

Анализ информационных потребностей включает выявление, упорядочение, документирование, изучение, уточнение, согласование и контроль изменений бизнес-требований к данным. Часть этих требований напрямую определяет, какие именно данные и информация нужны для бизнеса. Спецификации информационных потребностей должны находить отражение в предельно четких словесных формулировках и графических диаграммах.

Логическое моделирование — важное средство явного выражения потребностей бизнеса в данных. Пословица «лучше один раз увидеть, чем сто раз услышать» с предельной точностью отражает специфику восприятия очень многих людей. Есть, однако, и такие, кому наглядных картинок недостаточно, поскольку они лучше воспринимают отчеты и таблицы, создаваемые с помощью программных средств моделирования данных, и их информационные потребности также подлежат удовлетворению.

Во многих организациях установлены строгие формальные требования. Проектирование может происходить под руководством начальства, требующего соблюдения правил вплоть до порядка слов в формулировках, например: «Система должна поддерживать…» В таких случаях полезно управлять письменными инструкциями и спецификациями требований к данным с помощью специализированных программных средств управления требованиями. Спецификации, собранные и обобщенные в результате анализа содержания всевозможной регламентирующей документации, должны тщательно синхронизироваться с требованиями к данным, зафиксированными в моделях, что позволяет значительно упростить анализ влияния и получение ответов на вопросы вроде: «Где именно в моей модели учтено или реализовано требование X?» или «На каком основании в модель включена сущность Y и почему именно в этом месте?»

2.2.1.2.2 Анализ имеющейся документации

Для быстрого старта полезно проанализировать имеющиеся артефакты, включая описания уже созданных моделей данных и реализованных баз данных, на предмет возможности их использования. Даже если модели данных в целом устарели, отдельные их части вполне можно взять за основу новой модели. Только не забудьте удостовериться, проконсультировавшись с экспертами в предметных областях, что найденные наработки и артефакты соответствуют текущему положению дел.

Компании часто внедряют у себя пакеты приложений, такие как системы планирования ресурсов предприятия (ERP), в которых применяются собственные модели данных. При создании LDM эти модели данных должны, во-первых, учитываться, а во-вторых, использоваться, если это возможно и целесообразно, или увязываться с новой моделью данных предприятия. Кроме того, среди существующих артефактов могут найтись полезные модельные структуры — например, стандартные способы представления ролей участников. Кроме того, имеется ряд обобщенных правил представления данных, принятых в определенных областях деятельности вне зависимости от конкретной отрасли, таких как товарное производство или продажи, и их также следует в полной мере учитывать при разработке логической модели. Впоследствии эти общепринятые или отраслевые модели можно будет адаптировать под нужды конкретного проекта или инициативы.

2.2.1.2.3 Добавление ассоциативных сущностей

Ассоциативные сущности (associative entity) используются для развернутого описания связей типа «многие-ко-многим» (или «многие-ко-многим-ко-многим» и т. п.) с целью их декомпозиции. В ассоциативную сущность переносятся идентифицирующие атрибуты объектов, участвующих в описываемой связи. При необходимости ассоциативная сущность дополняется новыми атрибутами — например, датами, описывающими срок действия. Ассоциативные сущности могут иметь более двух родительских сущностей. В графовых базах данных ассоциативные сущности порой играют роль узлов. В многомерных моделях ассоциативные сущности, как правило, превращаются в таблицы фактов.

2.2.1.2.4 Добавление атрибутов

Добавление атрибутов к концептуальным сущностям в дальнейшем продолжается в логической модели и должно производиться вплоть до ее атомарного представления. Каждый атрибут должен соответствовать строго одному (и только одному) элементу данных (факту), не допускающему возможности дальнейшего дробления. Например, концептуальный атрибут «номер телефона» разбивается на все возможные логические атрибуты, описывающие тип номера телефона (домашний, рабочий, мобильный, факс и т. д.) и его структуру (код страны, код города/оператора, прямой номер, дополнительный номер и т. п.).

2.2.1.2.5 Определение доменов

Области значений атрибутов (домены), о которых говорилось в разделе 1.3.3.4, позволяют обеспечивать согласованность форматов, настроек и ограничений на значения однотипных данных в рамках связанных между собою проектов. Например, Стоимость обучения в год и Ставка оплаты преподавателей привязываются к домену Валюта, который является стандартным.

2.2.1.2.6 Определение ключей

Присваиваемые сущностям атрибуты могут быть как описательными, так и ключевыми. Во втором случае атрибут позволяет выделять единственный экземпляр сущности из общей совокупности

либо сам по себе, либо в сочетании с другими атрибутами. Не относящиеся к категории ключевых атрибуты описывают экземпляры объекта, но не позволяют их идентифицировать. Обязательно определяйте первичный и альтернативные ключи.

2.2.1.3 ФИЗИЧЕСКОЕ МОДЕЛИРОВАНИЕ ДАННЫХ

Логические модели данных требуют модификаций и адаптации с целью получения итогового проектного решения, обеспечивающего эффективную работу в среде конкретной СУБД. Например, адаптация LMD для Microsoft Access потребует внесения одних изменений, а для Teradata — совсем других. Далее в этом разделе термин *таблица* используется собирательно для обозначения таблиц, файлов и схем; термин *столбец* — для обозначения столбцов (колонок), полей и элементов; термин *строка* — для обозначения строк, записей и экземпляров в терминологии различных СУБД.

2.2.1.3.1 Разрешение логических абстракций

Логические абстрактные сущности (супертипы и подтипы) на стадии создания физического проекта базы данных преобразуются в отдельные объекты одним из двух способов.

- **Поглощение (absorption) подтипа.** Атрибуты подтипа сущности включаются в таблицу супертипа в виде столбцов, допускающих пустые поля (NULL).
- **Разделение (partition) супертипа.** Атрибуты супертипа сущности распределяются по отдельным таблицам, каждая из которых соответствует одному из подтипов.

2.2.1.3.2 Добавление детальной информации об атрибутах

В физической модели к атрибутам добавляются такие детали, как технические имена таблиц и столбцов (в реляционных базах данных), файлов и полей (в нереляционных БД) или схем и элементов (в базах данных XML).

Определите физические домены, тип и длину столбца или поля физической базы данных. Добавьте необходимые ограничения (например, по допустимости неопределенных значений) и значения по умолчанию для столбцов или полей, уделяя особое внимание обязательному указанию в явном виде требований по недопустимости неопределенных значений (NOT NULL).

2.2.1.3.3 Добавление объектов справочных данных

Небольшие наборы справочных данных, указанные на логическом уровне, в PMD могут быть реализованы тремя распространенными способами.

- **Соответствие отдельной таблицы кодов каждому объекту.** В некоторых моделях такой подход приводит к неуправляемому нагромождению множества таблиц.
- **Создание сводной совместно используемой таблицы кодов.** Помогает решить проблему избыточного числа отдельных таблиц кодов посредством их сворачивания в одну таблицу;

однако это означает, что единственное изменение в спецификации какого-либо справочного списка будет приводить к изменению всей таблицы. Кроме того, такой подход требует предельной внимательности во избежание конфликтующих значений кодов.

- **Включение правил или допустимых значений кодов в определения объектов.** При проектировании объекта в его определение физической реализации добавляется правило ограничения по кодам или список допустимых кодов. Такое решение подходит в тех случаях, когда список кодов используется применительно к единственному объекту PMD.

2.2.1.3.4 Определение суррогатных ключей

Часто в PMD бывает удобно использовать суррогатные ключи. Значения таких ключей должны быть скрыты от бизнес-пользователей, а кроме того, они не должны нести никакой смысловой нагрузки и не могут быть каким-либо образом связаны с данными, которым соответствуют. Этот этап не является обязательным, поскольку суррогатный ключ используется лишь в тех случаях, когда естественный ключ слишком велик, имеет сложносоставную структуру или содержит атрибуты, которые со временем могут меняться.

Если суррогатный ключ назначается первичным ключом таблицы, убедитесь, что на основе исходного первичного ключа определен альтернативный ключ. Например, если в LMD первичный ключ объекта Студент состоял из Фамилии студента, Имени студента и Даты рождения студента (то есть использовался составной первичный ключ), в PMD первичным ключом таблицы Студент может быть назначен суррогатный ключ ID студента. В таком случае должен быть определен и альтернативный ключ, основанный на исходном первичном ключе из Фамилии студента, Имени студента и Даты рождения студента.

2.2.1.3.5 Повышение производительности за счет денормализации

При определенных обстоятельствах денормализация, или привнесение избыточных данных, способна ускорить обработку запросов настолько радикально, что повышение производительности перевешивает издержки дублирования данных в хранилищах и их синхронизации. Основным средством преднамеренной денормализации является привнесение в базы данных многомерных структур.

2.2.1.3.6 Повышение производительности за счет индексирования

Индексирование данных — альтернативное средство оптимизации и ускорения обработки запросов к базам данных. Повысить производительность СУБД за счет индексирования данных удается во многих случаях. Но для этого администратору или разработчику базы данных нужно правильно выбрать тип индексирования таблиц. Основные коммерческие реляционные СУБД поддерживают разнообразные типы индексирования. Индексы могут быть уникальными или неуникальными, кластеризованными или некластеризованными, секционированными или несекционированными, одностолбцовыми или многостолбцовыми, на основе сбалансированных бинарных деревьев (b-tree), битовыми (bitmap) или хешированными (hashed). Без надлежащей

индексации СУБД вынуждена будет всякий раз заново сканировать всю таблицу в поисках запрошенных данных. В случае больших таблиц это очень затратно. По возможности индексируйте большие таблицы хотя бы по часто запрашиваемым столбцам с ключами (первичными, альтернативными и внешними).

2.2.1.3.7 Повышение производительности за счет секционирования

Очень внимательно следует подходить к стратегии секционирования (partitioning) в рамках общей модели данных (если проектируется многомерная модель), особенно если в таблицах фактов много необязательных (optional) ключей измерений. В идеале рекомендуется секционирование по ключам даты. Если же это нереализуемо, требуется дальнейшее изучение структуры данных на основе профилирования результатов и анализа рабочей нагрузки с целью предложения подходящей модели секционирования, которую впоследствии можно было бы улучшать.

2.2.1.3.8 Создание представлений

Представления можно использовать для контроля доступа к определенным элементам данных либо для того, чтобы задавать встроенные сочетания условий или фильтров с целью стандартизации отображения общих объектов или запросов. Сами представления следует проектировать исходя из бизнес-требований. Во многих случаях их нужно разрабатывать параллельно с проектированием LMD и PMD.

2.2.2 Обратное проектирование

Обратное проектирование (или реверс-инжиниринг — reverse engineering) — это процесс документирования существующей базы данных. Первым делом составляется PMD с целью понять техническое устройство имеющейся системы, затем создается LMD с целью документирования решаемых ею бизнес-задач, и, наконец, подготавливается CMD для документирования области применения системы и используемой терминологии. Большинство инструментов моделирования данных поддерживают реверс-инжиниринг самых разнообразных баз данных; однако для создания читабельного представления элементов моделей так или иначе потребуется специалист по моделированию данных. За основу для начала берется одна из стандартных схем представления (ортогональная, многомерная или иерархическая), но контекстное упорядочение модели (группировка сущностей по предметным областям или функциям) по-прежнему производится в основном вручную.

2.3 Проверка и оценка качества моделей данных

Как и любые другие результаты деятельности в сфере ИТ, модели данных нуждаются в контроле качества. В организации должна быть внедрена практика их непрерывного совершенствования. Технически это может быть реализовано с использованием различных методик, которые могут оценивать такие аспекты, как время реализации выгод (time-to-value), стоимость поддержки моделей или их качество. Одной из подобных методик является, например, карта балльной оценки

Data Model Scorecard® (Hoberman, 2009). Все методики в той или иной мере позволяют оценивать модели данных на предмет их корректности, полноты и непротиворечивости. По завершении создания CMD, LMD и PMD они становятся крайне полезными средствами обеспечения правильного понимания модели всеми заинтересованными лицами — от бизнес-аналитиков до разработчиков систем.

2.4 Сопровождение моделей данных

По завершении проектирования и построения моделей данных их нужно поддерживать в актуальном состоянии. Модель данных должна обновляться при всяком изменении требований и правил, а часто и при изменении каких-либо бизнес-процессов. В рамках конкретного проекта часто бывает так, что внесение изменений в модель на более низком уровне влечет за собой необходимость пересмотра и соответствующей высокоуровневой модели. Например, добавление нового столбца в таблицу PMD часто требует добавления соответствующего атрибута в объект LMD. Хорошей практикой является завершение каждой итерации разработки проведением операций реверс-инжиниринга новой версии физической модели данных с целью подтверждения, что она соответствует последней версии LMD. Многие инструменты моделирования данных позволяют автоматизировать процедуру сравнения физической и логической моделей.

3. ИНСТРУМЕНТЫ

Имеется много различных инструментов, призванных помочь специалистам по моделированию данных в их работе, включая инструменты моделирования, отслеживания происхождения данных, профилирования данных, а также репозитории метаданных.

3.1 Инструменты моделирования данных

Проектировщикам доступен широкий спектр программных продуктов, позволяющих автоматизировать многие задачи моделирования данных. Инструментарий базового уровня позволяет создавать диаграммы моделей данных с использованием палитры объектов и связей. Базовые инструменты также поддерживают *эластичное соединение* (*rubber banding*), при котором линии связей при перемещении сущностей перемещаются вслед за ними и перерисовываются. Более развитые программные средства поддерживают весь цикл прямого проектирования, включая создание CMD, LMD, PMD и генерирование структур на языке описания данных (Data Definition Language, DDL). Большинство приложений подобного уровня поддерживают и операции реверс-инжиниринга, начиная от описания базы данных до подготовки концептуальной модели. Часто эти сложные инструменты включают и такую функциональность, как проверка соблюдения стандартов именования, проверка правописания, хранение метаданных (например, определений и происхождения), а также поддержка совместного использования (в частности, веб-публикаций).

3.2 Инструменты для отслеживания происхождения данных

Инструменты для отслеживания происхождения — это программное обеспечение, позволяющее собирать и вести в структурированном виде сведения об источниках данных для каждого атрибута, включенного в модель. Эти же инструменты позволяют проводить анализ последствий изменений, то есть прогнозировать, каким именно образом то или иное изменение в одной из систем или подсистем скажется на других системах. Например, атрибут Валовый объем продаж может рассчитываться по исходным данным, поступающим из нескольких приложений, — и инструменты для отслеживания происхождения данных всю эту информацию фиксируют и сохраняют. Часто в качестве такого инструмента используется Microsoft Excel. При всей простоте и относительной дешевизне этого решения его недостаток заключается в том, что Excel не позволяет реализовать анализ последствий изменений и к тому же вынуждает управлять метаданными вручную. Информация о происхождении данных также часто фиксируется инструментами моделирования данных, в репозиториях метаданных или инструментами интеграции данных (см. главы 11 и 12).

3.3 Инструменты профилирования данных

Инструменты профилирования помогают исследовать содержание данных, сверять его с имеющимися метаданными и контролировать качество данных, выявляя пробелы и/или недостатки как в части качества самих данных, так и в части таких артефактов, как логические и физические модели, DDL и описания моделей. Например, если совмещение должностей бизнес-правилами не допускается, а некий сотрудник в какой-то период времени числится в системе на двух или более должностях, этот факт будет зафиксирован как аномалия данных (см. главы 8 и 13).

3.4 Репозитории метаданных

Репозиторий метаданных — программное средство для хранения описательной информации о модели данных, включая диаграмму модели и сопутствующий текст (например, определения), а также метаданных, импортированных из других программных средств и процессов (инструментов разработки программного обеспечения, систем управления бизнес-процессами (business process management, BPM), системных каталогов и т. д.). Репозиторий должен поддерживать функции интеграции метаданных и обмена метаданными. Обеспечение совместного доступа к метаданным — задача не менее важная, чем хранение метаданных. Репозитории метаданных должны предусматривать доступные и удобные способы просмотра содержимого хранилища и навигации. Средства моделирования данных обычно включают небольшие репозитории, но их возможности ограниченны (см. главу 13).

3.5 Шаблоны моделей данных

Шаблоны моделирования данных (data model patterns) — это повторно используемые типовые структуры для использования в моделях, применимые к широкому классу ситуаций. Различают элементарные (базовые), сборочные (модульные) и интеграционные шаблоны моделей данных.

Элементарные шаблоны — это, по сути, наборы заготовок, «винтов и гаек», из которых собираются модели данных. Они включают способы разрешения связей «многие-ко-многим» и построения масштабируемых иерархий. Сборочные шаблоны представляют собой строительные блоки для применения в самых различных моделях бизнес-процессов и данных. Важно, что они понятны представителям сферы бизнеса, поскольку оперируют такими понятиями, как активы, документы, категории людей, организации и т. п. Не менее важно, что они часто становятся предметами публикаций, из которых проектировщики моделей черпают проверенные, надежные, расширяемые и реализуемые конструктивные модули. Интеграционные шаблоны предлагают проектировщикам рамочные структуры для соединения сборочных шаблонов общепринятыми способами (Giles, 2011).

3.6 Отраслевые модели данных

Отраслевые модели данных представляют собой готовые модели, разработанные для универсального применения в масштабах целых отраслей экономики, таких как здравоохранение, телекоммуникации, страхование, банковское дело или промышленное производство. Такие модели часто весьма широки по области применения и одновременно очень детализированы. В некоторых отраслевых моделях данных содержатся тысячи объектов и атрибутов. Распространяются такие модели либо на коммерческой основе — разработчиками, либо через отраслевые ассоциации, такие как ARTS (торговля), SID (связь) или ACORD (страхование).

Любая покупная модель данных требует настройки и адаптации под нужды организации, поскольку разрабатывалась она на основе обобщения нужд и потребностей множества других организаций. Объемы работ по конфигурированию и перенастройке коммерческой модели будут зависеть от того, насколько близко она соответствует нуждам организации, а также от степени детализации важнейших ее компонентов. В одних случаях купленная модель может послужить эталоном и основой для ведущихся в организации работ по проектированию собственной модели данных, в других — полезным дополнением. Наконец, коммерческая модель может быть использована проектировщиками и просто в качестве позволяющего сэкономить время источника аннотированных готовых элементов.

4. ЛУЧШИЕ ПРАКТИКИ

4.1 Лучшие практики в области соглашений об именовании

Международный стандарт ISO 11179 «Регистры метаданных» содержит разделы, посвященные стандартным рекомендациям в части представления данных, включая имена и письменные определения различных элементов данных[1].

[1] См.: ГОСТ Р ИСО/МЭК 11179-5-2012 «Информационная технология (ИТ). Регистры метаданных (РМД). Часть 5. Принципы наименования и идентификация». — *Примеч. пер.*

Стандарты в области моделирования данных и проектирования баз данных определяют руководящие принципы, соблюдение которых необходимо для удовлетворения потребностей бизнеса в данных, обеспечения согласованности с корпоративной архитектурой и архитектурой данных (см. главу 4) и поддержки высокого уровня качества данных (см. главу 14). Архитекторы и аналитики данных, а также администраторы баз данных должны совместными усилиями разрабатывать эти стандарты. Причем стандарты данных должны дополнять соответствующие ИТ-стандарты, а не вступать с ними в противоречие.

Модель данных и стандарты присвоения имен элементам данных для каждого типа моделируемого объекта и для каждого объекта базы данных должны быть опубликованы. Стандартизация наименований особенно важна для сущностей, таблиц, атрибутов, ключей, представлений и индексов. Имя каждого такого элемента должно быть уникальным и в то же время максимально информативным.

В логической модели имена должны быть осмысленными с точки зрения бизнес-пользователей, поэтому следует по возможности избегать любых сокращений (за исключением общепринятых), используя полные и отражающие содержание слова. В физической модели имена не должны превышать максимально допустимой для выбранной СУБД длины, поэтому сокращайте их по мере необходимости. В то время как в именах элементов логических моделей слова разделяются пробелами, в физических моделях пробелы заменяются знаком подчеркивания.

Стандарты наименований должны быть по возможности едиными для всех рабочих сред (environments). Следует минимизировать рассогласование имен элементов в различных средах — тестовой (test), обеспечения качества (Quality Assurance, QA) или эксплуатационной (production). Для этого достаточно избегать присвоения элементам имен, указывающих на конкретную среду. В то же время для простоты разграничения сущностей и атрибутов, названий таблиц и столбцов используйте слова, определяющие классы (class words), которые стоят последними в именах атрибутов, таких как Количество (Quantity), Имя (Name), Код (Code). Они же помогают понимать, какого типа данные — количественные или качественные — описывает атрибут/столбец.

4.2 Лучшие практики проектирования баз данных

В процессе проектирования и построения базы данных ее администратору следует постоянно иметь в виду следующие принципы (запомните акроним PRISM).

- **Производительность (Performance) и простота использования.** Нужно обеспечить быстрый и простой доступ авторизованных пользователей к необходимым им данным, которые должны выдаваться в пригодной и ориентированной на потребности бизнеса форме, обеспечивая тем самым максимальную отдачу от приложений и данных.
- **Возможность повторного использования (Reusability).** Структура базы данных должна обеспечивать возможность использования данных по мере необходимости различными приложениями и в различных целях (например, для бизнес-анализа, повышения качества, стратегического планирования, управления отношениями с клиентами, совершенствования

процессов и т. п.). Избегайте привязки базы данных, структуры данных или объектов данных к единственному приложению.

- **Целостность (Integrity).** Все без исключения данные должны быть корректными и непротиворечивыми вне зависимости от контекста, а также обязаны точно отражать фактическую ситуацию в бизнесе. Ограничения, призванные обеспечивать целостность данных, должны быть максимально направлены на сами данные, а их соблюдение — постоянно контролироваться, чтобы гарантировать незамедлительное выявление любых нарушений целостности и уведомление о них.
- **Безопасность (Security).** Достоверные и точные данные должны быть всегда доступны авторизованным пользователям, но надежно защищены от несанкционированного доступа. Необходимо обеспечить надежную защиту конфиденциальных данных всех заинтересованных сторон, включая клиентов и бизнес-партнеров, и соблюдение всех требований регулирующих органов. Защита данных, как и обеспечение их целостности, должна быть реализована как можно ближе к самим данным с целью незамедлительного выявления нарушений требований информационной безопасности и уведомления о них.
- **Удобство сопровождения (Maintainability).** Весь комплекс работ по сопровождению данных должен быть окупаемым, то есть суммарные затраты на создание, хранение, ведение, использование и ликвидацию данных должны быть ниже суммарной оценки выгод, которые эти данные приносят организации. Также следует обеспечивать максимально оперативный отклик на возможные изменения в бизнес-процессах и бизнес-среде, включая удовлетворение новых потребностей бизнеса.

5. РУКОВОДСТВО МОДЕЛИРОВАНИЕМ И ПРОЕКТИРОВАНИЕМ ДАННЫХ

5.1 Управление качеством моделей и проектных решений

Аналитики данных, разработчики моделей и баз данных выступают в роли посредников между потребителями информации (теми, кто определяет нужны бизнеса в данных) и производителями данных (теми, кто фиксирует данные в пригодной для использования форме). Профессионалы в области данных должны обеспечивать баланс при учете требований к данным от потребителей информации и требований к приложениям от производителей данных.

Профессионалы в области данных также должны обеспечивать баланс при учете краткосрочных и долгосрочных интересов бизнеса. Потребителям информации нужны своевременные оперативные данные для выполнения своих обязанностей по текущему управлению бизнесом и реализации возможностей. Команды проектов по созданию систем должны укладываться в заданные временны́е и бюджетные рамки. Они же должны учитывать долгосрочные интересы всех заинтересованных сторон, обеспечивая размещение данных организации в безопасных и надежных хранилищах, защищенных системами резервного копирования и обеспечивающих совместный доступ к данным и их повторное использование, а также корректность, актуальность,

релевантность и максимальное удобство использования данных с точки зрения пользователей. Следовательно, модели и проектные решения по организации базы данных должны быть разумно сбалансированы таким образом, чтобы учитывать как краткосрочные, так и долгосрочные нужды организации.

5.1.1 Разработка стандартов моделирования и проектирования данных

Как уже отмечалось (в разделе 4.1), стандарты моделирования данных и проектирования баз данных определяют основополагающие принципы, позволяющие удовлетворять потребности бизнеса в данных, обеспечивать согласованность с корпоративной архитектурой и архитектурой данных и поддерживать высокий уровень качества данных. Стандарты моделирования данных и проектирования баз данных должны включать следующее.

- Перечень и описание стандартных результатов моделирования данных и проектирования баз данных.
- Перечень стандартных имен, допустимых сокращений и правил определения сокращений в частных случаях, распространяющийся на все объекты модели данных.
- Перечень стандартных форматов имен для всех объектов моделей данных, включая слова, определяющие классы (class words), используемые в названиях атрибутов и столбцов.
- Перечень и описание стандартных методов создания и сопровождения результатов моделирования и проектирования.
- Перечень и описание ролей и обязанностей специалистов по моделированию данных и проектированию баз данных.
- Перечень и описание свойств всех метаданных, фиксируемых в процессе моделирования и проектирования, включая бизнес-метаданные и технические метаданные; например, может быть приведена рекомендация по регистрации в модели сведений о происхождении данных для каждого атрибута.
- Требования и рекомендации к качеству метаданных (см. главу 13).
- Руководства по использованию инструментов моделирования данных.
- Руководства по подготовке и проведению проверки и оценки моделей и проектных решений.
- Руководства по управлению версиями моделей данных.
- Описание практик, не рекомендуемых к применению.

5.1.2 Проверка и оценка качества моделей данных и проектных решений

Проектные команды должны регулярно проводить обзорные проверки выполнения требований в сопоставлении с концептуальной и логической моделями данных, а также физическим проектом базы данных. На рабочих встречах, посвященных таким проверкам, должны рассматриваться исходная модель (если она существовала), вносимые в модель изменения и дополнения (включая рассмотренные ранее и отвергнутые варианты), а также вопросы соответствия новой модели действующим стандартам в области моделирования и архитектуры данных.

Обзорные проверки следует проводить с участием экспертов в предметных областях, обладающих различным опытом и навыками и представляющих различные профессиональные интересы, ожидания и мнения. Не исключено, что для привлечения к участию в таких рабочих встречах экспертов необходимого уровня потребуется получить официальное согласие руководства. Участники должны иметь возможность высказывать и обсуждать различные точки зрения и приходить к групповому консенсусу без персональных конфликтов, поскольку все они преследуют общую цель — содействие выработке наиболее практичных и эффективных проектных решений. Для организованного проведения встреч должен быть определен ответственный координатор, выступающий в роли лидера. Этот сотрудник обязан заранее планировать повестку, обеспечивать подготовку и распространение среди участников всех необходимых документов, запрашивать их мнения, поддерживать порядок и конструктивный ход встреч, а в конце подводить итоги и фиксировать принятые решения. Не лишним бывает и протоколирование хода встреч с целью не упустить важные моменты дискуссии.

Если при проведении проверок какие-то результаты проектирования не получили одобрения, проектировщики должны их пересмотреть и устранить выявленные проблемы. Если силами проектировщиков разрешить проблемы невозможно, последнее слово остается за владельцем системы, для которой создаются модель и проект базы данных.

5.1.3 Управление версиями и интеграцией моделей данных

Модели данных и прочие проектные спецификации подлежат тщательному контролю в части внесения каких-либо изменений, равно как и спецификации требований и другие результаты, получаемые в ходе жизненного цикла разработки систем. Следует отмечать каждое изменение, вносимое в модель данных, чтобы сохранить информацию о его происхождении. Если изменение затрагивает логическую модель данных — например, вследствие корректировки требований к бизнес-данным, — аналитик или архитектор данных должен рассмотреть и согласовать вносимые в модель изменения.

Каждое изменение должно сопровождаться следующими пояснениями.

- **Почему** проект или ситуация потребовали внесения изменения.
- **Что** и **Как** именно было изменено, включая точное описание добавленных, удаленных или измененных столбцов и связанных с ними изменений и т. д.
- **Когда** было утверждено решение о внесении изменения и когда оно было внесено в модель (когда изменение было реализовано, в системе фиксировать не обязательно).
- **Кто** внес изменение.
- **Где** внесено изменение (в каких моделях).

Некоторые инструменты моделирования данных включают репозитории, которые поддерживают функциональность управления версиями и интеграции. Если таковые отсутствуют, сохраняйте модели данных в экспортируемых файлах DDL или XML и ведите их учет в системе управления репозиторием исходного кода наравне с исходным кодом приложений.

5.2 Метрики моделирования данных

К объективной оценке качества моделей данных можно подходить с различными наборами мерок и критериев, но в любом случае необходима эталонная база сравнения. В качестве примера приведем лишь одну из возможных методик проверки соответствия модели данных общепринятым стандартам качества, а именно — шаблон ведомости оценки модели данных Data Model Scorecard® (Hoberman, 2015). Шаблон предусматривает оценку качества модели данных по десяти категориям критериев с подсчетом итоговой суммы баллов и процентов набранных баллов по каждой категории (табл. 11).

Таблица 11. Шаблон ведомости оценки модели данных Data Model Scorecard®

№	Категория	Максимум (баллы)	Оценка модели (баллы)	%	Комментарии
1	Насколько хорошо в модели отражены требования к данным?	15			
2	Является ли модель достаточно полной?	15			
3	Насколько хорошо модель согласуется со схемой представления данных?	10			
4	Насколько хорошо модель проработана структурно?	15			
5	Насколько эффективно модель использует преимущества обобщенных структур?	10			
6	Соблюдаются ли в модели стандарты именования?	5			
7	Является ли модель читабельной?	5			
8	Насколько хорошо сформулированы определения?	10			
9	Насколько модель согласуется с текущей корпоративной практикой представления данных?	5			
10	Достаточно ли хорошо метаданные описывают данные?	10			
	СУММА БАЛЛОВ	100			

В столбце «Оценка модели» проверяющий модель эксперт выставляет свою оценку ее соответствия критерию, указанному в столбце «Категория» в пределах от нуля до максимального числа баллов, определяющего удельный вес этого критерия в суммарной оценке. Например, если эксперт выставляет оценку в 10 баллов из 15 возможных по категории «Насколько хорошо в модели отражены требования к данным?», эти 10 баллов добавляются к суммарной оценке, а в столбце степени соответствия (%) появляется значение 66%. В столбце «Комментарии» проверяющий

должен объяснять причины снижения оценки относительно максимальной и/или давать рекомендации по устранению недостатков. В последней строке подсчитываются сумма баллов и средний процент соответствия модели критериям по всем десяти категориям.

Далее приводится краткое описание каждой категории.

1. **Насколько хорошо в модели отражены требования к данным?** Здесь мы проверяем степень учета в модели требований к данным. Если есть требование по регистрации заказов, проверяем наличие в модели структур для сбора информации о заказах. Если есть требование по обеспечению возможности просмотра данных о **Количестве студентов** по **Семестрам** и **Специализации**, необходимо убедиться, что модель поддерживает такой запрос.

2. **Является ли модель достаточно полной?** Здесь полнота рассматривается в двух аспектах: полнота соблюдения требований и полнота метаданных. Первый аспект подразумевает, что в модели все запрошенные требования учтены достаточно подробно. Этот же аспект подразумевает отсутствие в модели избыточных данных. Конечно, нет ничего проще, чем добавлять не запрашиваемые структуры данных — на тот случай, если они понадобятся в ближайшем будущем. Проверка призвана такие структуры выявлять и снижать за них оценку, поскольку они усложняют проект и задерживают срок его сдачи из-за того, что проектировщик тратит время на моделирование данных, которые не требуются. Следует считаться с издержками от реализации потенциальных будущих требований в случае, если эти требования так и не возникнут. Полнота метаданных означает наличие исчерпывающей описательной информации по всем элементам модели; например, если мы проверяем физическую модель данных, мы вправе ожидать наличия описаний форматов и информации о допустимости неопределенных значений по всем полям.

3. **Насколько хорошо модель согласуется со схемой представления данных?** Здесь мы проверяем соответствие уровня детализации (концептуальный, логический или физический) и схемы (например, реляционная, многомерная, NoSQL) рассматриваемой модели определениям, используемым для моделей соответствующего типа.

4. **Насколько хорошо модель проработана структурно?** Здесь мы проверяем соблюдение формальных правил моделирования, с тем чтобы гарантировать возможность построения реальной физической базы данных на основе изучаемой модели. Важно вовремя выявлять и устранять такие ошибки, как наличие двух одноименных атрибутов у одной и той же сущности или первичного ключа с атрибутом, допускающим неопределенное значение.

5. **Насколько эффективно модель использует преимущества обобщенных структур?** Здесь мы удостоверяемся в достаточном уровне и правильности абстрагирования понятий. Например, заменяя атрибуты типа Местонахождение заказчика обобщенными атрибутами типа Местонахождение, проектировщик получает возможность впоследствии использовать те же сущности в качестве шаблонов-заготовок для описания местонахождения других физических объектов: например, складов, центров выдачи заказов и т. п.

6. **Соблюдаются ли в модели стандарты именования?** Здесь проверяется последовательное соблюдение стандартов именования в масштабах модели. Следует обратить внимание прежде всего на соответствие стандартам в части структуры, терминологии и стиля. Под стандартными требованиями в части структуры понимается включение в названия сущностей, атрибутов и связей определенных составных элементов. Например, стандарт может предусматривать включение в имя каждого атрибута сущности имени самой сущности, такого как Клиент или Продукт. Стандартные требования в части терминологии предписывают именовать сущности или атрибуты определенного типа строго определенными словами (например, Клиент, а не Покупатель или Заказчик). К требованиям в части терминологии относятся также требования по соблюдению правил орфографии и сокращения слов. Под требованиями в части стиля понимаются, например, требования по выбору регистра имен всех элементов наименований в соответствии со стандартной практикой.

7. **Является ли модель читабельной?** Казалось бы, читабельность — отнюдь не самый важный из десяти рассматриваемых вопросов. Однако если модель читается с трудом, то, разбираясь в ней, можно упустить важные аспекты, относящиеся к другим категориям оценочной ведомости. Размещение дочерних сущностей строго под родительскими, отображение связанных сущностей по соседству друг с другом и минимизация длины связующих линий — всё это значительно повышает читабельность моделей.

8. **Насколько хорошо сформулированы определения?** Здесь мы проверяем, насколько ясными, полными и точными являются определения.

9. **Насколько модель согласуется с текущей корпоративной практикой представления данных?** Здесь мы должны убедиться, что все структуры в модели данных представлены в широком и согласованном контексте общей терминологии и правил, созвучных принятому в организации языку. Все структуры, включенные в модель данных, должны терминологически и понятийно согласовываться с аналогичными структурами, встречающимися в других моделях данных, и, в идеале, с корпоративной моделью данных, если она существует.

10. **Насколько метаданные соответствуют описываемым данным?** Здесь мы подтверждаем, что модель адекватно отражает реальные данные, которые будут храниться в физической базе данных, построенной на основе модели. То есть мы проверяем, например, столбец Фамилия_клиента и удостоверяемся, что там действительно будут храниться исключительно фамилии и именно клиентов. Эта категория проверки призвана устранить риск неприятных сюрпризов, когда вдруг выясняется, что структура базы данных не соответствует тем данным, которые должны в ней храниться.

Таким образом, рассмотренный шаблон позволяет проводить всестороннюю оценку качества модели и выявлять конкретные области ее дальнейшего совершенствования.

6. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Ambler, Scott. *Agile Database Techniques: Effective Strategies for the Agile Software Developer*. Wiley and Sons, 2003. Print.

Avison, David and Christine Cuthbertson. *A Management Approach to Database Applications*. McGraw-Hill Publishing Co., 2002. Print. Information systems ser.

Blaha, Michael. *UML Database Modeling Workbook*. Technics Publications, LLC, 2013. Print.

Brackett, Michael H. *Data Resource Design: Reality Beyond Illusion*. Technics Publications, LLC, 2012. Print.

Brackett, Michael H. *Data Resource Integration: Understanding and Resolving a Disparate Data Resource*. Technics Publications, LLC, 2012. Print.

Brackett, Michael H. *Data Resource Simplexity: How Organizations Choose Data Resource Success or Failure*. Technics Publications, LLC, 2011. Print.

Bruce, Thomas A. *Designing Quality Databases with IDEF1X Information Models*. Dorset House, 1991. Print.

Burns, Larry. *Building the Agile Database: How to Build a Successful Application Using Agile Without Sacrificing Data Management*. Technics Publications, LLC, 2011. Print.

Carlis, John and Joseph Maguire. *Mastering Data Modeling — A User-Driven Approach*. Addison-Wesley Professional, 2000. Print.

Codd, Edgar F. «A Relational Model of Data for Large Shared Data Banks». *Communications of the ACM*, 13, № 6 (June 1970).

DAMA International. *The DAMA Dictionary of Data Management. 2nd Edition: Over 2,000 Terms Defined for IT and Business Professionals*. 2nd ed. Technics Publications, LLC, 2011. Print.

Daoust, Norman. *UML Requirements Modeling for Business Analysts: Steps to Modeling Success*. Technics Publications, LLC, 2012. Print.

Date, C. J. *An Introduction to Database Systems*. 8th ed. Addison-Wesley, 2003. Print.

Date, C. J. and Hugh Darwen. *Databases, Types and the Relational Model*. 3d ed. Addison Wesley, 2006. Print.

Date, Chris J. *The Relational Database Dictionary: A Comprehensive Glossary of Relational Terms and Concepts, with Illustrative Examples*. O'Reilly Media, 2006. Print.

Dorsey, Paul. *Enterprise Data Modeling Using UML*. McGraw-Hill Osborne Media, 2009. Print.

Edvinsson, Håkan and Lottie Aderinne. *Enterprise Architecture Made Simple: Using the Ready, Set, Go Approach to Achieving Information Centricity*. Technics Publications, LLC, 2013. Print.

Fleming, Candace C. and Barbara Von Halle. *The Handbook of Relational Database Design*. Addison Wesley, 1989. Print.

Giles, John. *The Nimble Elephant: Agile Delivery of Data Models using a Pattern-based Approach*. Technics Publications, LLC, 2012. Print.

Golden, Charles. *Data Modeling 152 Success Secrets — 152 Most Asked Questions On Data Modeling — What You Need to Know*. Emereo Publishing, 2015. Print. Success Secrets.

Halpin, Terry, Ken Evans, Pat Hallock, and Bill McLean. *Database Modeling with Microsoft Visio for Enterprise Architects*. Morgan Kaufmann, 2003. Print. The Morgan Kaufmann Series in Data Management Systems.

Halpin, Terry. *Information Modeling and Relational Databases*. Morgan Kaufmann, 2001. Print. The Morgan Kaufmann Series in Data Management Systems.

Halpin, Terry. *Information Modeling and Relational Databases: From Conceptual Analysis to Logical Design*. Morgan Kaufmann, 2001. Print. The Morgan Kaufmann Series in Data Management Systems.

Harrington, Jan L. *Relational Database Design Clearly Explained*. 2nd ed. Morgan Kaufmann, 2002. Print. The Morgan Kaufmann Series in Data Management Systems.

Hay, David C. *Data Model Patterns: A Metadata Map*. Morgan Kaufmann, 2006. Print. The Morgan Kaufmann Series in Data Management Systems.

Hay, David C. *Enterprise Model Patterns: Describing the World (UML Version)*. Technics Publications, LLC, 2011. Print.

Hay, David C. *Requirements Analysis from Business Views to Architecture*. Prentice Hall, 2002. Print.

Hay, David C. *UML and Data Modeling: A Reconciliation*. Technics Publications, LLC, 2011. Print.

Hernandez, Michael J. *Database Design for Mere Mortals: A Hands-On Guide to Relational Database Design*. 2nd ed. Addison-Wesley Professional, 2003. Print.

Hoberman, Steve, Donna Burbank, Chris Bradley, et al. *Data Modeling for the Business: A Handbook for Aligning the Business with IT using High-Level Data Models*. Technics Publications, LLC, 2009. Print. Take It with You Guides.

Hoberman, Steve. *Data Model Scorecard*. Technics Publications, LLC, 2015. Print.

Hoberman, Steve. *Data Modeling Made Simple with ER/Studio Data Architect*. Technics Publications, LLC, 2013. Print.

Hoberman, Steve. *Data Modeling Made Simple: A Practical Guide for Business and IT Professionals*. 2nd ed. Technics Publications, LLC, 2009. Print.

Hoberman, Steve. *Data Modeling Master Class Training Manual*. 7th ed. Technics Publications, LLC, 2017. Print.

Hoberman, Steve. *The Data Modeler's Workbench. Tools and Techniques for Analysis and Design*. Wiley, 2001. Print.

Hoffer, Jeffrey A., Joey F. George, and Joseph S. Valacich. *Modern Systems Analysis and Design*. 7th ed. Prentice Hall, 2013. Print.

IIBA and Kevin Brennan, ed. *A Guide to the Business Analysis Body of Knowledge (BABOK Guide). International Institute of Business Analysis*, 2009. Print.

Kent, William. *Data and Reality: A Timeless Perspective on Perceiving and Managing Information in Our Imprecise World*. 3rd ed. Technics Publications, LLC, 2012. Print.

Krogstie, John, Terry Halpin, and Keng Siau, eds. *Information Modeling Methods and Methodologies: Advanced Topics in Database Research*. Idea Group Publishing, 2005. Print. Advanced Topics in Database Research.

Linstedt, Dan. *Super Charge Your Data Warehouse: Invaluable Data Modeling Rules to Implement Your Data Vault*. Amazon Digital Services. 2012. Data Warehouse Architecture Book 1.

Muller, Robert. J. *Database Design for Smarties: Using UML for Data Modeling*. Morgan Kaufmann, 1999. Print. The Morgan Kaufmann Series in Data Management Systems.

Needham, Doug. *Data Structure Graphs: The structure of your data has meaning*. Doug Needham Amazon Digital Services, 2015. Kindle.

Newton, Judith J. and Daniel Wahl, eds. *Manual for Data Administration*. NIST Special Publications, 1993. Print.

Pascal, Fabian. *Practical Issues in Database Management: A Reference for The Thinking Practitioner*. Addison-Wesley Professional, 2000. Print.

Reingruber, Michael. C. and William W. Gregory. *The Data Modeling Handbook: A Best-Practice Approach to Building Quality Data Models*. Wiley, 1994. Print.

Riordan, Rebecca M. *Designing Effective Database Systems*. Addison-Wesley Professional, 2005. Print.

Rob, Peter and Carlos Coronel. *Database Systems: Design, Implementation, and Management*. 7th ed. Cengage Learning, 2006. Print.

Schmidt, Bob. *Data Modeling for Information Professionals*. Prentice Hall, 1998. Print.

Silverston, Len and Paul Agnew. *The Data Model Resource Book, Volume 3: Universal Patterns for Data Modeling*. Wiley, 2008. Print.

Silverston, Len. *The Data Model Resource Book, Volume 1: A Library of Universal Data Models for All Enterprises*. Rev. ed. Wiley, 2001. Print.

Silverston, Len. *The Data Model Resource Book, Volume 2: A Library of Data Models for Specific Industries*. Rev. ed. Wiley, 2001. Print.

Simsion, Graeme C. and Graham C. Witt. *Data Modeling Essentials*. 3rd ed. Morgan Kaufmann, 2004. Print.

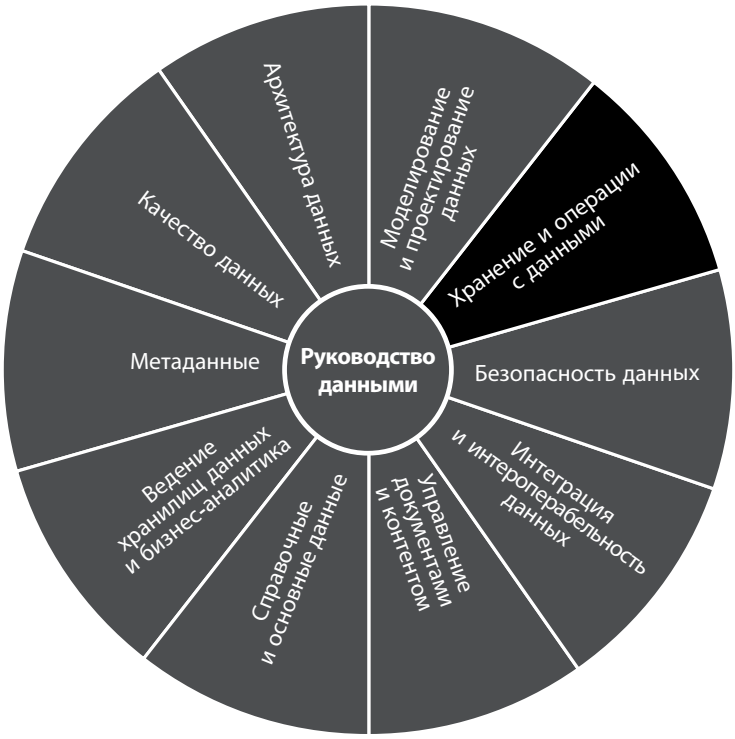
Simsion, Graeme. *Data Modeling: Theory and Practice*. Technics Publications, LLC, 2007. Print.

Teorey, Toby, et al. *Database Modeling and Design: Logical Design*, 4th ed. Morgan Kaufmann, 2010. Print. The Morgan Kaufmann Series in Data Management Systems.

Thalheim, Bernhard. *Entity-Relationship Modeling: Foundations of Database Technology*. Springer, 2000. Print.

Watson, Richard T. *Data Management: Databases and Organizations*. 5th ed. Wiley, 2005. Print.

Хранение и операции с данными



DAMA-DMBOK2 Data Management Framework

1. ВВЕДЕНИЕ

Хранение и операции с данными включают проектирование и реализацию решений для хранения, а также сопровождение хранимых данных с целью получения от них максимальной выгоды на протяжении всего их жизненного цикла (см. главу 1). Работы в этой области ведутся по двум основным направлениям.

- **Сопровождение баз данных** объединяет работы, относящиеся к жизненному циклу данных, включая первоначальную реализацию рабочей среды базы данных (database environment), получение данных, а также их резервное копирование и удаление. Сюда же относится обеспечение оптимальной производительности (мониторинг и настройка — критически важные элементы сопровождения).

Рисунок 54. Контекстная диаграмма: хранение и операции с данными

- **Технологическая поддержка баз данных** включает определение технических требований, соответствующих информационным потребностям организации, определение технической архитектуры, развертывание и администрирование технологических решений, а также разрешение проблемных вопросов, связанных с технологиями.

Ключевую роль в каждом из этих направлений деятельности играют администраторы баз данных (АБД) (Database administrators, DBAs). АБД — самая устоявшаяся и общепринятая профессиональная роль в сфере управления данными, а практические аспекты администрирования баз данных, вероятно, наиболее проработанная и зрелая область практики управлении данными. Помимо работ, описываемых в настоящей главе, АБД принимают активное участие в деятельности по обеспечению безопасности данных (см. главу 7).

1.1 Бизнес-драйверы

В процессе своей операционной деятельности компании постоянно используют собственные информационные системы. С учетом этого обстоятельства хранение и операции с данными являются жизненно важными аспектами деятельности организаций, бизнес которых зависит от данных. Таким образом, обеспечение непрерывности бизнеса является главным драйвером усилий в рассматриваемой области управления данными. Если база данных становится недоступной, текущая операционная деятельность организации осуществляется с задержками или полностью останавливается. Надежная инфраструктура хранения данных, обеспечивающая проведение операций, позволяет свести к минимуму риск подобных сбоев.

1.2 Цели и принципы

Цели хранения и операций с данными включают:

- управление доступностью данных на протяжении всего их жизненного цикла;
- обеспечение целостности информационных активов;
- управление эффективностью проведения информационных транзакций.

Хранение и операции с данными представляют сугубо технические аспекты управления данными. Администраторы БД и другие лица, задействованные в этой работе, будут лучше справляться со своими должностными обязанностями и помогут улучшить управление данными в целом, если станут придерживаться следующих руководящих принципов.

- **Выявление и использование любых возможностей для автоматизации рабочих процессов.** По возможности автоматизируйте процессы разработки баз данных, работу с инструментальными средствами, а также любые процессы, которые позволяют сократить каждый цикл разработки, свести к минимуму ошибки и переделки, ослабляя тем самым нагрузку на команду разработчиков. Придерживаясь этого принципа, администраторы БД получают

возможность перехода к более итеративным и гибким (agile) подходам к разработке приложений. Работа по продвижению в этом направлении должна вестись в тесном сотрудничестве со специалистами по моделированию и архитекторами данных.

- **Построение с учетом повторного использования.** Планируйте и поддерживайте применение абстрагированных и повторно используемых объектов данных, которые ослабляют тесную привязку приложений к конкретным схемам баз данных (и возникающую в связи с этим так называемую проблему «объектно-реляционного несоответствия» — object-relational impedance mismatch[1]). Для достижения этой цели имеется целый ряд инструментов и механизмов, включая представления (views), триггеры, функции и хранимые процедуры, объекты данных приложений и слои доступа к данным (data access layers), языки XML и XSLT, типизированные наборы данных ADO.NET и веб-сервисы. Администраторы БД должны также уметь выбирать оптимальный подход к виртуализации данных. Конечная цель — сделать использование базы данных как можно более быстрым, простым и безболезненным процессом.
- **Знание и разумное использование лучших практик.** Администраторы БД должны всячески способствовать введению требований по применению стандартов и лучших практик, сохраняя достаточную гибкость для того, чтобы отходить от них в разумных пределах, когда для этого имеются веские основания. Стандарты в области баз данных не должны служить препятствием для успешной реализации проекта.
- **Увязка стандартов в области баз данных с требованиями по сопровождению.** Например, соглашение об уровне обслуживания (Service Level Agreement, SLA) может отражать рекомендованные АБД и поддержанные разработчиками методы обеспечения целостности и безопасности данных. Соглашение должно также отражать перенос ответственности с АБД на команду разработчиков в случаях, когда последние станут разрабатывать собственные процедуры обновления данных или слоя доступа к данным. Такой подход позволяет избежать безальтернативного отношения к использованию стандартов по принципу «всё или ничего».
- **Определение ожиданий в отношении роли АБД при выполнении проекта.** Проектная методология должна предусматривать обязательное участие АБД в мероприятиях, относящихся к фазе определения проекта, — это может существенно помочь на всех последующих этапах жизненного цикла разработки системы. АБД заранее получит представление о потребностях проекта и требованиях по сопровождению, что повысит эффективность коммуникаций за счет четкого понимания ожиданий проектной команды от деятельности группы сопровождения данных. Участие в процессе анализа и проектирования специально выделенного главного АБД и его заместителя позволяет прояснить, что потребуется от АБД в отношении задач, стандартов, рабочих усилий и сроков в процессе разработки. Команды также должны четко сформулировать свои ожидания относительно сопровождения после ввода в эксплуатацию.

[1] «Объектно-реляционное несоответствие» — совокупность концептуальных и технических проблем, которые обычно возникают, когда программное приложение для взаимодействия с реляционной СУБД создается с помощью объектно-ориентированного языка программирования. Термин «impedance mismatch» (рассогласование импедансов) позаимствован из электротехники. — *Примеч. науч. ред.*

1.3 Основные понятия и концепции

1.3.1 Терминология баз данных

Терминология баз данных относится к разряду специальной и носит технический характер. Работая в качестве АБД или совместно с АБД, важно понимать специфику этого технического языка.

- **База данных (БД).** Любая совокупность хранимых данных, вне зависимости от их структуры или содержания. По отношению к некоторым большим базам данных могут также использоваться термины «экземпляр» или «схема».
- **Экземпляр (instance).** Совокупность исполняющихся программ работы с базой данных, контролирующих доступ к определенной области памяти. В одной организации обычно бывает запущено много экземпляров БД, выполняющихся параллельно и использующих различные области памяти. Каждый экземпляр независим от всех остальных.
- **Схема (schema).** Подмножество объектов базы данных, содержащихся в базе данных или экземпляре. Схемы используются для организации объектов в лучше управляемые части базы данных или экземпляра. Как правило, у схемы имеется владелец (owner) и список доступа (access list), который зависит от содержимого схемы. Обычно схемы как раз и применяются для того, чтобы изолировать объекты, содержащие данные ограниченного доступа (чувствительные данные — sensitive data), от объектов с общим доступом, или для отделения представлений, доступных в режиме «только для чтения», от таблиц реляционной БД, на основе которых они созданы. Схема также может применяться для организации работы с совокупностью структур базы данных, содержащих какие-либо данные, предназначенные для общего пользования.
- **Узел (node).** Отдельный компьютер, предоставляющий свои ресурсы для обработки данных или их хранения в рамках распределенной базы данных.
- **Абстрагирование базы данных.** Обеспечение возможности обращения к функциям для работы с базой данных с помощью интерфейса прикладного программирования (Application Programming Interface, API), например для того, чтобы приложения могли подключаться к различным базам данных без необходимости для программистов знать особенности всех вызовов функций для всех возможных баз данных. Примером API, обеспечивающего абстрагирование базы данных, является открытый интерфейс доступа к базам данных (Open Database Connectivity, ODBC). Очевидное преимущество абстрагирования базы данных — переносимость; к недостаткам можно отнести невозможность использования специфичных функций, которые не являются общими для всех баз данных.

1.3.2 Управление жизненным циклом данных

АБД обеспечивают точность и согласованность данных на протяжении всего их жизненного цикла, в процессе проектирования, внедрения и использования любых систем, которые осуществляют хранение, обработку или поиск данных. Администратор ведет надзор за любыми изменениями

базы данных. Многие заинтересованные стороны могут запрашивать внесение различных изменений, но лишь АБД вправе определять точный состав и содержание изменений, вносимых в БД, планировать и контролировать их реализацию.

Управление жизненным циклом данных включает внедрение политик и процедур, регламентирующих получение, перемещение, хранение, контроль сроков использования и ликвидацию данных по мере их устаревания. Разумным в этом контексте представляется использование заранее подготовленных чек-листов, которые помогают контролировать качество выполнения всех этих задач. Администраторы должны обеспечивать контролируемый, документируемый и проверяемый процесс последовательного проведения изменений в базах данных приложений сначала в среде проверки качества — QA (Quality Assurance or Certification) environment, а затем в среде эксплуатации (или производственной — production environment). Обычно такой процесс инициируется с помощью подтвержденного менеджером запроса на обслуживание или запроса на изменение. В любом случае у администратора должен всегда иметься план возврата к исходному состоянию (back out plan) в случае возникновения проблем.

1.3.3 Администраторы

Администратор базы данных (АБД) — самая устоявшаяся и общепринятая профессиональная роль в сфере управления данными. АБД несут основную ответственность за хранение и операции с данными и, кроме того, играют критически важную роль в обеспечении их безопасности (см. главу 7), а также при обсуждении физических аспектов в процессе моделирования данных и при разработке физического проекта базы данных (см. главу 5). Наконец, АБД обеспечивают поддержку баз данных как в средах разработки, тестирования и проверки качества, так и в среде эксплуатации.

АБД не проводят все работы по хранению и операциям с данными самостоятельно. Кроме них в этой деятельности принимают участие распорядители данных, архитекторы данных, сетевые администраторы, аналитики данных и специалисты по информационной безопасности. Специалисты каждой группы вносят свой вклад в работы по планированию производительности, сохранению данных, их восстановлению. Эти же команды могут участвовать в получении и обработке данных из внешних источников.

Часто у администраторов БД имеется узкопрофильная специализация: например, администратор базы данных среды эксплуатации, прикладной администратор, процедурный администратор, администратор разработки. В некоторых организациях также предусмотрены роли администраторов сетевых систем хранения данных (Network Storage Administrators, NSA), которые специализируются на сопровождении сетевых хранилищ, рассматриваемых отдельно от остальных приложений или структур, обеспечивающих хранение данных.

В некоторых организациях АБД различных специализаций относятся к различным организационным системам блока ИТ. Например, администраторы эксплуатации могут входить в состав организационной системы, обеспечивающей поддержку эксплуатационной инфраструктуры, или же в состав групп поддержки приложений. Прикладные и процедурные администраторы,

а также администраторы базы данных среды эксплуатации иногда относятся к организационным системам разработки приложений. Администраторы сетевых систем хранения часто включаются в организационные системы поддержки эксплуатационной инфраструктуры.

1.3.3.1 АБД СРЕДЫ ЭКСПЛУАТАЦИИ

Администраторы баз данных среды эксплуатации несут основную ответственность за управление операциями с данными, включая:

- обеспечение оптимальной производительности и надежности базы данных посредством проведения настройки и мониторинга, формирования и анализа отчетов об ошибках, а также выполнения других необходимых работ;
- реализацию механизмов резервного копирования и восстановления данных в случае их утраты или повреждения по каким бы то ни было причинам и при любых обстоятельствах;
- реализацию механизмов кластеризации и автоматического аварийного переключения на резервную базу данных в тех случаях, когда требуется бесперебойный доступ к данным;
- выполнение других работ по сопровождению базы данных, таких как реализация механизмов архивирования данных.

Выходными результатами деятельности АБД среды эксплуатации являются следующие.

- Эксплуатационная среда базы данных, включая экземпляр СУБД, функционирующий на сервере, который имеет достаточную для поддержки необходимой производительности вычислительную мощность и емкость. Сервер должен быть сконфигурирован таким образом, чтобы обеспечивать надлежащий уровень защиты, надежности и доступности данных; за работоспособность СУБД отвечают системные администраторы базы данных.
- Механизмы и процессы контролируемого внесения изменений в базы данных среды эксплуатации.
- Механизмы обеспечения доступности, целостности и восстанавливаемости данных, реализованные с учетом любых обстоятельств, допускающих потерю или повреждение данных.
- Механизмы выявления ошибок при функционировании базы данных, СУБД и сервере и формирования отчетов о них.
- Соответствие уровней доступности, восстанавливаемости и производительности базы данных соглашению об уровне обслуживания.
- Механизмы и процессы мониторинга изменения производительности базы данных при изменении рабочей нагрузки и объема данных.

1.3.3.2 ПРИКЛАДНЫЕ АБД

Прикладной администратор базы данных отвечает за одну или несколько баз данных во всех средах (разработки/тестирования, проверки качества и эксплуатационной), в противоположность

системным администраторам базы данных каждой из этих сред. Иногда прикладные АБД подотчетны подразделениям организации, отвечающим за разработку и сопровождение приложений, которые работают с их базами данных. Наличие специально выделенных прикладных АБД имеет свои плюсы и минусы.

Прикладные АБД рассматриваются как внутренние участники команды поддержки приложения. Поскольку их внимание сосредоточено на конкретной базе данных, они могут оказать более качественную помощь разработчикам приложений. Однако такие администраторы рискуют зациклиться на решении узких задач, утратив при этом понимание информационных потребностей организации в целом и отойдя от принятых в организации практик администрирования баз данных. Прикладные АБД работают в тесном контакте с аналитиками данных, разработчиками моделей и архитекторами данных.

1.3.3.3 ПРОЦЕДУРНЫЕ АБД И АБД РАЗРАБОТКИ

Процедурные администраторы базы данных являются ведущими исполнителями в части анализа и администрирования процедурных объектов базы данных. Процедурный АБД специализируется на разработке и поддержке процедурной логики (представленных в виде программного кода алгоритмов), выполнение которой осуществляет СУБД, включая хранимые процедуры, триггеры и пользовательские функции (User Defined Functions, UDFs). Процедурный администратор отвечает за то, чтобы процедурная логика была запланирована, реализована, протестирована и предоставлена для совместного (и повторного) использования.

Администратор базы данных разработки основные усилия направляет на создание и обслуживание проектных решений в части специализированных баз данных, таких как различного рода экспериментальные базы («песочницы» — sandbox) или области для исследования (exploration areas).

Во многих случаях две эти функции объединяются в одну позицию.

1.3.3.4 АДМИНИСТРАТОРЫ СЕТЕЙ ХРАНЕНИЯ ДАННЫХ

Администраторы сетей хранения занимаются вопросами аппаратного и программного обеспечения систем, которые объединяют устройства хранения данных, распределенные по сети. Многочисленные сетевые системы хранения данных имеют специфические требования в отношении мониторинга и администрирования, отличающиеся от простых систем баз данных.

1.3.4 Типы архитектур баз данных

Прежде всего, базы данных подразделяются на централизованные и распределенные. Централизованная система управляет одной базой данных, в то время как распределенная система управляет множеством баз данных, реализованных во множестве систем. Распределенные системы можно разделить на два класса по степени автономности входящих в них компонентов: федеративные (автономные компоненты) и не федеративные (неавтономные компоненты). Рисунок 55 иллюстрирует различие между централизованной и распределенной архитектурами.

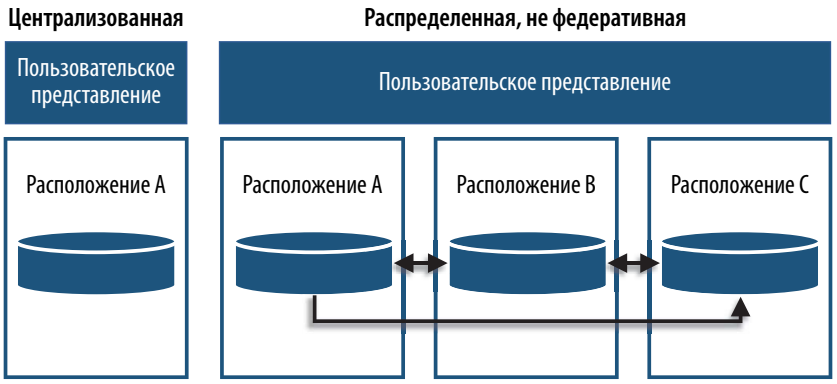


Рисунок 55. Централизованная и распределенная архитектуры базы данных

1.3.4.1 ЦЕНТРАЛИЗОВАННЫЕ БАЗЫ ДАННЫХ

В централизованных базах все данные хранятся в одной системе в единственном месте. Все пользователи обращаются к данным с помощью одной системы. Для некоторых видов данных с ограниченным кругом пользователей такой подход идеален, а вот в случае данных, которые должны быть широко доступны, подобная централизация сопряжена с серьезными рисками. Например, если централизованная система выходит из строя, то альтернативных вариантов получения необходимых данных нет.

1.3.4.2 РАСПРЕДЕЛЕННЫЕ БАЗЫ ДАННЫХ

Распределенные базы данных обеспечивают возможность быстрого доступа к данным через множество узлов. Популярные технологии распределенных баз данных основаны на использовании стандартных аппаратных решений широкого применения (commodity) и позволяют масштабировать архитектуру, начиная с единичных серверов и заканчивая тысячами и тысячами машин, поддерживающих локализованное хранение и вычисления. Вместо обеспечения высокой степени доступности данных за счет наращивания аппаратных мощностей программное обеспечение управления распределенными базами ориентировано на реплицирование данных по серверам, что обеспечивает высокодоступный сервис на основе компьютерного кластера. Большое внимание при разработке таких программ уделяется также развитию возможностей по эффективному выявлению и нейтрализации последствий сбоев: при потенциальном отказе любого из компьютеров в распределенной сети система в целом остается работоспособной.

В некоторых распределенных базах данных реализована модель распределенных вычислений MapReduce, способствующая еще большему повышению производительности за счет дробления запросов к базе данных на множество фрагментов, каждый из которых обрабатывается отдельно любым из компьютеров — узлов кластера. Кроме того, дополнительный выигрыш дает размещение данных параллельно на многих вычислительных узлах, что обеспечивает крайне высокую скорость (ширину полосы) обработки данных в целом по кластеру. При этом и файловая система, и приложения разработаны так, чтобы сбои на уровне узлов обрабатывались автоматически.

1.3.4.2.1 Федеративные базы данных

Федеративная архитектура позволяет предоставлять пользователям данные из различных источников без приложения дополнительных усилий по их подготовке или дублирования массивов источников данных. В федеративной системе информация из множества автономных баз данных отображается в одну федеративную базу данных. Входящие в федеративную структуру базы иногда географически разнесены и соединяются с помощью компьютерной сети. Они остаются автономными, но одновременно предоставляют федеративной системе контролируемый доступ к определенной части своих данных. Федеративная архитектура — хорошая альтернатива объединению (слиянию — merging), когда речь идет о разнородных по структуре базах данных. Фактической интеграции данных, хранящихся в составляющих федеративную структуру автономных базах, не происходит; вместо этого за счет интероперабельности обеспечивается представление этих баз как одного большого объекта (см. главу 8). В не федеративных системах баз данных, напротив, отдельные СУБД интегрируются и утрачивают свою автономность, поскольку ими начинает управлять центральная СУБД.

Федеративные базы данных лучше всего подходят для проектов по интеграции гетерогенных и распределенных систем, таких как интеграция корпоративной информации, виртуализация данных, согласование схем и управление основными данными.

Архитектуры федеративных систем баз данных различаются в зависимости от уровней интеграции с локальными базами данных и объема предлагаемых услуг. В целом федеративные СУБД можно разделить на слабо связанные (loosely coupled) и сильно связанные (tightly coupled).

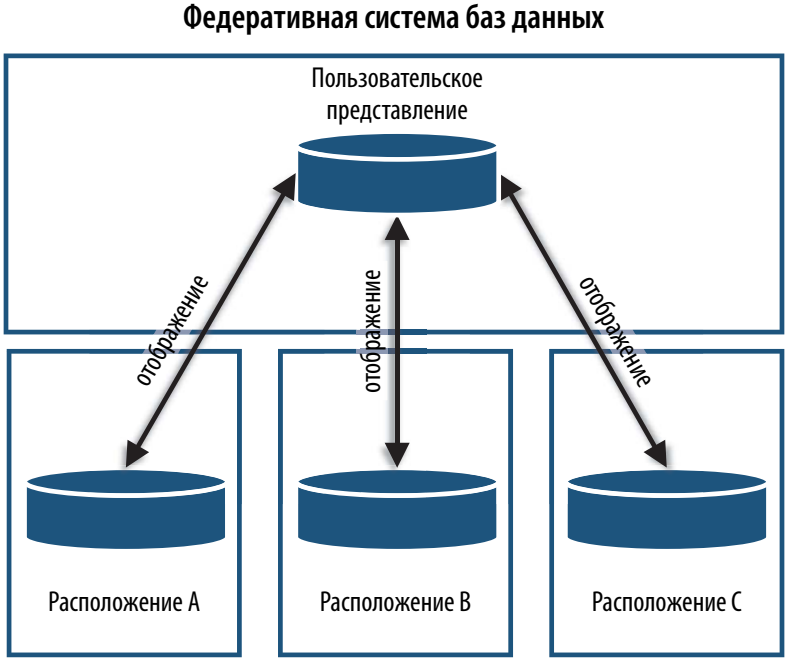


Рисунок 56. Федеративная система баз данных

В слабо связанных федеративных системах для каждой локальной базы данных требуется построение ее собственной федеративной схемы и применяются специальные способы именования для доступа к объектам локальных баз. Пользователь обычно получает доступ к данным систем-компонентов посредством запросов на общем для всех локальных баз языке мультибазы данных (multi-database), но это приводит к исчезновению прозрачности относительно фактического места расположения данных, то есть пользователь должен знать непосредственно схему объединения баз в федерацию. В таких системах на глобальном уровне допускается только выборка данных. Пользователь импортирует все требующиеся данные из входящих в федерацию баз данных, а затем самостоятельно интегрирует их, формируя пользовательскую федеративную схему.

В сильно связанных системах на уровне локальных систем реализованы независимые процессы построения и публикации интегрированной федеративной схемы (см. рис. 57). Одна и та же схема может применяться ко всем входящим в федерацию базам данных, при этом отсутствует необходимость в репликации.

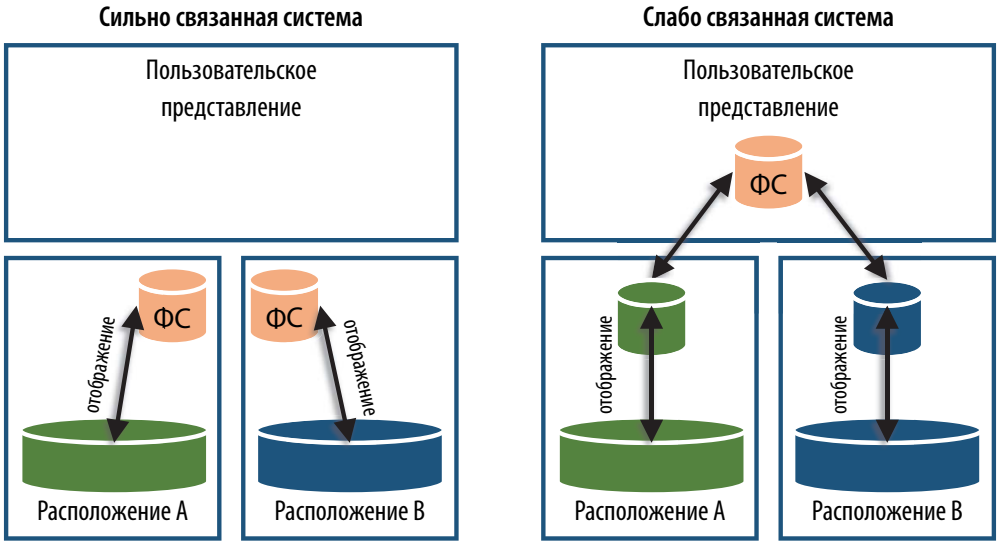


Рисунок 57. Схемы связывания баз данных в федеративную систему

1.3.4.2.2 Базы данных блокчейн

Базы данных блокчейн («цепочка блоков» — blockchain) представляют собой разновидность федеративной базы данных, широко использующуюся для безопасного управления финансовыми транзакциями. Они могут также использоваться при управлении контрактами или для обмена конфиденциальными сведениями в здравоохранении. Схема блокчейн включает структуры двух типов — индивидуальные записи и блоки. Каждой транзакции соответствует запись. В базе данных создаются цепи из хронологически упорядоченных групп транзакций (блоков), в которых каждый блок содержит закодированную информацию о предыдущем блоке в цепи. Закодированная информация о зарегистрированных в блоке транзакциях формируется с помощью алгоритмов

хеширования. Как только создается новый блок, хеш-код в предыдущем блоке (который до того момента был последним в цепочке) фиксируется и больше не подлежит пересчету. Это означает, что данные о транзакциях, содержащиеся во всех блоках цепочки кроме последнего, изменяться не могут. Любое изменение (подделка) данных о транзакциях или блоке будет обнаружено, поскольку хеш-код не будет соответствовать измененным данным.

1.3.4.3 ВИРТУАЛИЗАЦИЯ / ОБЛАЧНЫЕ ПЛАТФОРМЫ

Виртуализация (также называемая «облачными вычислениями» — cloud computing) позволяет оказывать услуги по проведению вычислений, использованию программного обеспечения, предоставлению доступа к данным и их хранению таким образом, что конечному пользователю не требуются знания о физическом местонахождении и конфигурации систем, обеспечивающих предоставление этих услуг. Облачные вычисления часто сравнивают с энергосетями: потребителям электроэнергии ведь также без разницы, где и как она вырабатывается и по каким инфраструктурным сетям поступает, — они просто используют ее как платную услугу. Однако, в отличие от энергосетей, виртуализация может быть не только распределенной (off-premises), но и локальной (on-premise).

Облачные вычисления — естественный этап эволюции практики повсеместного применения виртуализации, сервис-ориентированных архитектур и коммунальных вычислений (utility computing). Ниже кратко описаны некоторые методы реализации баз данных в облаке.

- **Образ виртуальной машины.** Облачные платформы предоставляют пользователям возможность арендовать экземпляры виртуальных машин и использовать их для работы со своими базами данных. Пользователи могут либо загружать на них собственные образы машины с развернутой базой данных, либо использовать предлагаемые провайдерами готовые образы машин с предустановленными и настроенными СУБД.
- **База данных как услуга.** Некоторые облачные платформы предлагают возможность использования базы данных как услуги (Database-as-a-Service, DaaS) без запуска экземпляра виртуальной машины. В такой конфигурации владельцы приложения вовсе избавлены от необходимости устанавливать и поддерживать базу данных. Провайдер услуги DaaS сам устанавливает и поддерживает базу данных, а владельцы приложения пользуются ею за абонентскую плату.
- **Управляемый облачный хостинг базы данных.** При таком варианте база данных не предлагается в качестве услуги; вместо этого провайдер облачного сервиса размещает ее у себя в облаке и осуществляет управление базой данных по поручению и в интересах собственника приложения.

Администраторы БД совместно с сетевыми и системными администраторами должны выработать комплексный подход к реализации проектов, который предусматривал бы требования к стандартизации, консолидации, виртуализации, автоматизации функций резервного копирования и восстановления данных, а также обеспечения безопасности этих функций.

- **Стандартизация/консолидация.** Консолидация позволяет сократить количество мест расположения хранилищ данных, используемых в организации, а также уменьшить количество хранилищ и процессов, действующих в центре обработки данных (ЦОД). Основываясь на политике руководства данными, архитекторы данных совместно с администраторами БД могут разрабатывать стандартные процедуры, включающие выявление особо важных данных, определение сроков хранения данных, порядок шифрования, а также политики в области репликации данных.
- **Виртуализация серверов.** Технологии виртуализации позволяют заменять или объединять оборудование, в частности серверы, установленные в разных ЦОДах. Виртуализация приводит к снижению как капитальных, так и эксплуатационных затрат, включая энергопотребление. Эти технологии позволяют создавать и виртуальные рабочие столы (virtual desktops), которые можно размещать на серверах ЦОДов и сдавать в аренду за абонентскую плату. Аналитическое агентство Gartner считает виртуализацию катализатором модернизации (Bittman, 2009). Виртуализация обеспечивает большую гибкость при выполнении операций по хранению данных за счет предоставления ресурсов хранения в локальной или облачной среде.
- **Автоматизация.** Автоматизация обработки данных включает организацию автоматического выполнения таких задач, как сбор/предоставление данных, конфигурирование, проведение обновлений, управление версиями и обеспечение соблюдения правил.
- **Безопасность.** Меры по обеспечению безопасности данных в виртуальных системах должны быть объединены с действующими мерами по обеспечению безопасности существующей физической инфраструктуры (см. главу 7).

1.3.5 Типы подходов к обработке данных

Известны два основных типа подходов к обработке данных — ACID и BASE. Они являются двумя полюсами спектра возможных промежуточных вариантов. Эти акронимы легко запоминаются в силу их созвучия полюсам спектра значений показателя pH, характеризующего кислотно-щелочной баланс[1]. Для того чтобы оценить степень соответствия распределенной системы подходу ACID или BASE, используется теорема CAP.

1.3.5.1 ПОДХОД ACID

Акроним ACID появился в начале 1980-х годов[2] и с тех пор используется для обозначения четырех обязательных требований, соблюдение которых необходимо для надежного выполнения транзакций в СУБД. На протяжении десятилетий они служат прочным фундаментом, на котором должны строиться механизмы обработки транзакций.

[1] Имеются в виду созвучия ACID — acid (кислота) и BASE — base (основание). Как известно из курса химии, щелочь является основанием — соединением, химически противоположным кислоте. — *Примеч. науч. ред.*

[2] Саму концепцию сформулировал еще в 1970-х годах американский теоретик вычислительных систем Джим Грей (*англ.* James Nicholas «Jim» Gray), а термин ACID впервые был введен в работе Haerder and Rueter (1983). — *Примеч. науч. ред.*

- **Неделимость (атомарность — Atomicity).** Выполняются либо все операции транзакции, либо ни одна из них; иными словами, сбой выполнения любой части транзакции означает сбой выполнения всей транзакции целиком.
- **Согласованность (Consistency).** Транзакция должна обеспечивать соответствие базы данных всем правилам, определенным в системе (переводить базу данных из одного допустимого состояния в другое допустимое состояние); не завершенные по какой-либо причине транзакции аннулируются.
- **Изолированность (Isolation).** Любая транзакция и ее результаты не оказывают влияния на параллельно выполняемые транзакции.
- **Устойчивость (Durability).** Завершенная транзакция необратима и не может быть отменена.

Технологии, основанные на принципах ACID, являются основным инструментом, применяемым в реляционных СУБД; большинство из них использует в качестве интерфейса язык SQL.

1.3.5.2 ПОДХОД BASE

Беспрецедентный рост объемов данных и степени их изменчивости, потребность документирования и хранения неструктурированных данных, необходимость оптимизации рабочих нагрузок систем обработки данных по скорости считывания и вытекающие отсюда требования по обеспечению большей гибкости в отношении масштабирования, организации проектных решений, снижения затрат и обеспечения аварийного восстановления данных привели к развитию альтернативного и диаметрально противоположного ACID комплекса требований, получившего сокращенное название BASE.

- **Базовая доступность (Basic Availability).** Система гарантирует сохранение некоторого уровня доступности к данным даже при сбое в работе части узлов. Предоставленные системой данные могут быть устаревшими, но все ответы на запросы будут выданы.
- **Гибкое состояние (Soft State).** Данные пребывают в состоянии постоянного изменения; полученный ответ на запрос не гарантирует соответствия предоставляемых сведений действительности.
- **Отложенная согласованность (Eventual Consistency).** В конечном итоге данные будут согласованы по всем узлам и во всех базах системы, но непротиворечивость данных для всех транзакций и в любой момент времени не гарантирована.

Системы, ориентированные на подход BASE, широко распространены в средах обработки больших данных. Крупные интернет-компании и социальные сети часто используют в своих системах BASE-реализации, поскольку исчерпывающей точности всех элементов данных в каждый заданный момент времени им не требуется. Таблица 12 отражает основные различия между подходами ACID и BASE.

Таблица 12. Сравнение подходов ACID и BASE

Характеристика	ACID	BASE
Перестраиваемость структуры данных	Схема данных обязательна	Динамическая структура данных
	Табличная структура	Оперативно подстраиваемая структура
	Столбцы однотипных данных	Хранение разнородных данных
Согласованность	Строгая согласованность	Строгая, отложенная или отсутствует
Основной способ обработки данных	Транзакционная обработка	Обработка пар «ключ-значение»
Основной способ хранения данных	Строки/Столбцы	Неформатированные столбцы
История возникновения	СУБД 1970-х годов	Неструктурированные БД 2000-х годов
Масштабирование	В зависимости от продукта	Автоматическое распределение данных по стандартным серверам широкого применения
Политика в области доступности исходных кодов	Различные политики	ПО с открытым кодом
Поддержка транзакций	Обязательна	Возможна

1.3.5.3 ТЕОРЕМА CAP

Теорема CAP (она же теорема Брюера[1]) стала ответом на наметившийся переход ко всё более распределенным вычислительным системам (Brewer, 2000). Положения теоремы говорят о том, что для распределенной системы не может быть обеспечено одновременное выполнение всех требований ACID в любой момент времени. Более того, чем крупнее система, тем хуже эти требования соблюдаются. А потому в распределенных системах приходится искать компромиссные решения, позволяющие сбалансировать три следующих свойства.

- **Согласованность (Consistency).** Система должна работать корректно и предсказуемо в любой момент времени.
- **Доступность (Availability).** Система должна бесперебойно принимать запросы и отвечать на них.
- **Устойчивость к разделению (Partition Tolerance).** Система должна сохранять работоспособность в случаях частичной потери данных или отказов отдельных компонентов.

[1] Эрик Брюер (*англ.* Eric Allen Brewer, р. 1964) — специалист по информатике из Калифорнийского университета в Беркли, с 2011 г. — руководитель проектов развития инфраструктуры Google. Описываемый принцип сформулирован Брюером в 1999 г. в качестве гипотезы. Формальное доказательство CAP-теоремы для ряда частных случаев получено коллегами Брюера из Массачусетского технологического института в 2002 г. (см. doi:10.1145/564585.564601). — *Примеч. пер.*

Теорема CAP постулирует, что в любой распределенной системе обработки данных может поддерживаться не более двух из вышеперечисленных свойств. Обычно это утверждение формулируется в виде правила «два из трех» (см. рис. 58).

Рисунок 58. Теорема CAP

Интересное применение эта теорема нашла в лямбда-архитектуре (lambda architecture), которая будет описана подробно в главе 14. Лямбда-архитектура предусматривает два уровня работы с данными: уровень ускорения и уровень пакетной обработки. Первый уровень применяется в тех случаях, когда более предпочтительны доступность и устойчивость к разделению, а второй — когда необходимы доступность и согласованность.

1.3.6 Средства хранения данных

Данные могут храниться на всевозможных носителях, включая жесткие диски, энергозависимые запоминающие устройства и флэш-накопители. Некоторые системы используют различные сочетания устройств хранения различных типов. Чаще всего используются дисковые накопители и сети хранения данных (Storage Area Network, SAN), решения в оперативной памяти (in-memory), решения на основе технологии сжатия по колонкам (columnar compression colutions), виртуальные сети хранения данных (Virtual Storage Area Network, VSAN), решения на основе технологии радиочастотной идентификации (Radio Frequency IDentification, RFID), цифровые кошельки (digital wallets), ЦОДы и частные, публичные и гибридные облачные хранилища (см. главу 14).

1.3.6.1 ДИСКОВЫЕ НАКОПИТЕЛИ И СЕТИ ХРАНЕНИЯ ДАННЫХ

Жесткие диски обеспечивают надежное долгосрочное хранение данных. В одной системе могут совместно использоваться диски различных типов. Данные по дискам можно распределять в зависимости от характера их использования. Например, редко запрашиваемые данные можно хранить на относительно дешевых дисках с медленным доступом, а часто используемые — на высокопроизводительных дисковых накопителях.

Массивы дисков способны объединяться в сети хранения данных (SAN). При этом обмен данными в SAN может не требовать сети, поскольку данные способны перемещаться с помощью шины.

1.3.6.2 БАЗЫ ДАННЫХ В ОПЕРАТИВНОЙ ПАМЯТИ

Базы данных в оперативной памяти (In-Memory Databases, IMDB) загружаются из хранилища в энергозависимую (оперативную) память при включении системы, и все процедуры обработки данных происходят без обращения к жестким дискам, что значительно ускоряет работу системы по сравнению с системами, всякий раз считывающими данные с дисков. Большинство IMDB также имеют конфигурируемую защиту от потери данных в случае внезапного отключения питания.

Если приложению для работы требуется объем данных, гарантированно помещающийся в оперативную память, использование IMDB приводит к существенной оптимизации работы. Базы данных в оперативной памяти обеспечивают более предсказуемую скорость считывания данных, что выгодно отличает их от дисковых хранилищ, однако такие решения являются весьма дорогостоящими. IMDB предоставляют функциональные возможности для реализации аналитических процессов в режиме реального времени и, как правило, только для этих целей и планируются в силу инвестиционных ограничений.

1.3.6.3 РЕШЕНИЯ СО СЖАТИЕМ ПО КОЛОНКАМ

Колоночные базы данных созданы специально для обработки наборов данных с множеством повторяющихся значений. Например, из таблицы, содержащей 256 колонок, для поиска запрошенного значения в одном из них будут поочередно извлекаться все строки, чтобы проверить единственное из 256 полей в каждой из них (при этом, возможно, потребуется обращение к диску). Однако при использовании колоночной базы данных можно значительно уменьшить описанные потоки ввода/вывода за счет хранения колонок с применением сжатия данных. Заключается сжатие в том, что в каждом столбце сохраняется не само значение, а указатель на одно из допустимых значений, содержащихся в отдельной таблице значений. Объем данных в основной таблице за счет этого сжимается многократно.

1.3.6.4 ФЛЭШ-ПАМЯТЬ

Последние достижения в области технологий хранения данных сделали флэш-память, или твердотельные накопители (Solid State Drives, SSDs), привлекательной альтернативой жестким дискам. По скорости доступа они теперь практически не уступают решениям в оперативной памяти, а по надежности и долговечности — дисковым хранилищам.

1.3.7 Среды баз данных

Базы данных на протяжении жизненного цикла разработки систем используются в различных средах. Для обеспечения тестирований изменений АБД должны участвовать в проектировании структур данных в среде разработки. Команда АБД непосредственно отвечает за реализацию

изменений в среде проверки качества (QA environment), и никто, кроме нее, не имеет права вносить какие-либо изменения в среду эксплуатации. Изменения в эксплуатационной среде должны производиться в строгом соответствии со стандартными процессами и процедурами.

В то время как большинство технологий управления данными основаны на использовании программного обеспечения, установленного на оборудовании общего применения, в отдельных случаях уникальные требования к данным обусловливают необходимость разработки и использования узкоспециализированного аппаратного обеспечения, включая серверы, построенные специально для выполнения операций по преобразованию и распространению данных. Такие серверы интегрируются с существующей инфраструктурой либо напрямую, как модульное расширение, либо как внешнее устройство — с помощью сетевого соединения.

1.3.7.1 СРЕДА ЭКСПЛУАТАЦИИ

Среда эксплуатации (или производственная — production environment) — это рабочая техническая среда, в которой протекают все бизнес-процессы. Она является критически важным элементом обеспечения выполнения организацией своей миссии. Если среда эксплуатации придет в нерабочее состояние, все бизнес-процессы остановятся, а это повлечет за собой убытки и негативные последствия для клиентов, утративших возможность доступа к поддерживаемым системой сервисам. Для экстренных служб или систем жизнеобеспечения внезапное прекращение функционирования может повлечь катастрофические последствия.

С точки зрения бизнеса существует только среда эксплуатации. Однако для ее надежного функционирования необходимо иметь и правильно использовать и другие среды. Например, разработка и тестирование никак не могут проводиться в эксплуатационной среде, поскольку это поставило бы под угрозу производственные процессы и сохранность данных.

1.3.7.2 ПРЕДЭКСПЛУАТАЦИОННЫЕ СРЕДЫ

Для проведения разработки, тестирования и обкатки изменений до их реализации в эксплуатационной среде используются различные предэксплуатационные среды (pre-production environments). В таких средах проблемы, связанные с планируемыми изменениями, выявляются и устраняются без ущерба для текущих бизнес-процессов. Однако для того, чтобы потенциальные проблемы действительно выявлялись, конфигурация предэксплуатационных сред должна быть максимально приближенной к конфигурации реальной эксплуатационной среды.

Из-за громоздкости и дороговизны элементов инфраструктуры очень сложно полностью воспроизвести условия эксплуатационной среды в предэксплуатационной. Однако чем ближе к фазе передачи в эксплуатацию находится фаза цикла разработки, обслуживаемая предэксплуатационной средой, тем точнее эта среда должна воспроизводить эксплуатационную. Любое несоответствие по составу и конфигурации оборудования может вызвать дополнительные проблемы, никак не связанные с реализуемыми изменениями, что осложняет процесс исследования и устранения проблем.

Основные типы предэксплуатационных сред включают: среду разработки, среду тестирования и среды специального применения.

1.3.7.2.1 Среда разработки

Среда разработки обычно представляет собой облегченную версию эксплуатационной среды. Как правило, в ней задействован меньший объем физической и оперативной памяти, меньшее число процессоров и т. д. В этой среде разработчики создают и отлаживают коды новых версий отдельных компонентов, которые затем собираются в среде проверки качества для проведения полного интеграционного тестирования. Среда разработки может содержать много копий моделей данных эксплуатационной среды, в зависимости от того, как осуществляется управление проектами разработки. Более крупные организации могут выделить для отдельных разработчиков персональные среды разработки со всеми необходимыми правами по управлению этими средами.

Любые исправления и обновления должны прежде всего тестироваться в среде разработки. Эта среда должна быть изолирована от эксплуатационной среды и функционировать на другом физическом оборудовании. В связи с необходимостью такой изоляции может потребоваться копирование данных из систем, находящихся в эксплуатации, в среду разработки. Однако во многих отраслях на данные информационных систем организаций распространяются требования регулирующих органов по защите информации, и переносить их запрещено. Поэтому, прежде чем предпринимать какие-то действия в отношении данных в эксплуатационной среде, следует убедиться, что предполагаемые операции не нарушают каких-либо запретов (см. главу 7).

1.3.7.2.2 Среда тестирования

В среде тестирования могут производиться проверки качества, пользовательское приемочное тестирование и, в некоторых случаях, тестирование производительности или стрессовое тестирование. Во избежание получения искаженных результатов тестирования из-за различий программно-аппаратных конфигураций, в идеале в среде тестирования должно использоваться программное и аппаратное обеспечение, идентичное эксплуатационной среде. В случае тестирования производительности соблюдение данного требования является особенно важным. Тестовая среда может быть подключена через сеть к эксплуатационной среде с целью считывания реальных данных (это не обязательно). Однако она *ни при каких условиях* не должна осуществлять запись данных в системы, находящиеся в эксплуатации.

Среды тестирования могут использоваться в различных целях, включая следующие.

- **Проверка качества.** Тестирование на предмет соответствия систем функциональным требованиям.
- **Интеграционное тестирование.** Тестирование как единого целого собранных частей системы, которые были разработаны или модернизированы по отдельности.
- **Пользовательское приемочное тестирование (User Acceptance Testing, UAT).** Тестирование функций системы с точки зрения соответствия требованиям пользователей. Как правило, проводится по специально подготовленным сценариям использования (use-cases).
- **Тестирование производительности.** Такое тестирование дает возможность проверить производительность системы в условиях больших объемов данных или повышенной сложности

операций в любое время, не дожидаясь ее плановой остановки на профилактику и без риска ухудшения ее производительности за счет пиковых нагрузок.

1.3.7.2.3 Экспериментальные среды («песочницы»)

«Песочницей» называют предоставляемую в распоряжение пользователей альтернативную среду, допускающую подключение к данным находящихся в эксплуатации систем только на чтение. «Песочницы» используются для экспериментирования со всевозможными вариантами разработок, проверки гипотез относительно данных, объединения данных информационных систем организации с данными, создаваемыми пользователями, или вспомогательными данными из внешних источников. Также полезно использовать «песочницы» для проверки различного рода концепций (Proof-of-Concept).

Среда «песочницы» может представлять собой либо подмножество систем эксплуатационной среды, изолированное от реальных рабочих процессов, либо полностью независимую среду. Пользователи «песочницы» часто имеют права на выполнение операций CRUD — создание, чтение, обновление, удаление данных (Create, Read, Update, Delete) внутри выделенной им области, что позволяет быстро проверять идеи и обкатывать варианты изменений, планируемых для реализации в системе. Администраторы БД обычно выполняют в экспериментальных средах минимальный объем работ, поскольку их роль ограничивается выделением ресурсов и развертыванием среды, выдачей прав на доступ и мониторингом использования. Если области «песочницы» выделены непосредственно в эксплуатирующихся системах, их следует тщательно изолировать во избежание негативного влияния на выполнение реальных операций. Эти среды не должны иметь возможность записи данных в действующие рабочие системы.

Если у организации есть средства на приобретение соответствующих лицензий на программное обеспечение, «песочницы» можно разворачивать на базе виртуальных машин.

1.3.8 Организация баз данных

Системы хранения данных предоставляют возможность включения в программы готовых инструкций по записи данных на диски и управлению их обработкой, так что разработчики при реализации функций манипулирования данными могут просто использовать эти инструкции. Три основных класса моделей организации баз данных — иерархическая, реляционная и нереляционная. Полностью взаимоисключающими эти классы не являются (см. рис. 59). Некоторые СУБД поддерживают чтение и запись данных, организованных как в реляционные, так и нереляционные структуры, а иерархические базы данных могут отображаться на реляционные таблицы. Плоские (неструктурированные) файлы с разделителями строк могут построчно считываться в таблицы, а для хранения данных из этих строк могут быть организованы один или несколько столбцов.

1.3.8.1 Иерархические БД

Иерархическая организация базы данных — старейшая и самая жестко-структурированная модель, широко использовавшаяся в СУБД эпохи мэйнфреймов. В иерархических СУБД данные

организованы в соответствии с древовидной логической схемой, то есть все объекты логической модели данных в обязательном порядке связаны отношениями «родитель — потомок», причем родительский объект может иметь много потомков, но каждый дочерний объект имеет строго одного родителя (то есть в иерархии устанавливаются исключительно связи типа «один-ко-многим»). Пример иерархической структуры данных — дерево каталогов. XML-документы также используют иерархическую модель. Хотя их и можно представить в виде реляционной БД, фактически их структура соответствует пути обхода дерева.

Рисунок 59. Спектр вариантов организации баз данных

1.3.8.2 РЕЛЯЦИОННЫЕ БД

Многие ошибочно полагают, что реляционные базы данных получили свое название из-за наличия связей (или отношений — relation) между таблицами. Это не так. Модель основана на теории множеств и реляционной алгебре, где элементы данных или атрибуты (столбцы) связываются в виде кортежей (строк) (см. главу 5). Таблицы представляют собой наборы кортежей идентичной структуры (отношения). Операции над множествами (объединение, пересечение, вычитание и т. д.) используются для упорядочения и извлечения данных с помощью структурированных запросов на языке SQL. Чтобы записать данные, нужно заранее знать их структуру (схему при записи — schema on write). Реляционные базы данных являются строчно-ориентированными (row-oriented).

Система управления базой данных в случае реляционной модели называется RDBMS (Relational Database Management System). Для хранения динамично меняющейся информации используются в основном именно реляционные базы данных. Вариациями на тему реляционных баз данных являются многомерные и темпоральные БД.

1.3.8.2.1 Многомерные БД

Технологии многомерных БД обеспечивают такую структуру хранения данных, которая позволяет осуществлять поиск с использованием одновременно нескольких фильтров по различным элементам данных. Чаще всего многомерная структура используется в хранилищах данных

и бизнес-аналитике. Некоторые модели баз данных этого типа являются интеллектуальной собственностью, хотя большинство крупных СУБД имеют встроенную в виде объектов технологию многомерного («кубического») представления данных. Доступ к данным осуществляется посредством запросов на языке многомерных выражений MDX (MultiDimensional eXpression), который является усложненным вариантом SQL.

1.3.8.2.2 Темпоральные БД

Темпоральная база данных (temporal database) представляет собой реляционную базу данных со встроенной поддержкой обработки данных, включающих элементы, связанные со временем. Обычно учитываются такие характеристики, как время действия и время транзакции. Эти атрибуты могут быть, в частности, объединены в виде битемпоральных (bi-temporal) данных.

- **Действительное время (valid time)** — это временны́е рамки, в которых данные о фактах соответствуют действительности (отражают действительную ситуацию в отношении описываемой ими сущности).
- **Транзакционное время (transaction time)** — это период времени, в течение которого данные о фактах считаются истинными с точки зрения логики базы данных.

Помимо двух вышеописанных в темпоральной БД возможны и другие временны́е шкалы — например, Время принятия решения (decision time). В таком случае БД называется мультитемпоральной (multi-temporal) в противопоставление битемпоральным. Темпоральные базы данных позволяют разработчикам и администраторам БД управлять текущей, прогнозируемой и исторической версиями данных в одной и той же базе данных.

1.3.8.3 НЕРЕЛЯЦИОННЫЕ БД

В нереляционных (non-relational) базах данные могут храниться в виде строк или целых файлов. Данные могут считываться из них по-разному, в зависимости от потребностей (эта характеристика нереляционных баз данных называется «схема при чтении» — schema on read). Нереляционные БД могут быть строчно-ориентированными, однако это требование не обязательно.

Нереляционная БД удобна для быстрого доступа к базам данных, использующим менее строгие в отношении согласованности данных модели по сравнению с традиционной реляционной моделью. Основными мотивами применения такого подхода являются простота проектных решений, горизонтальная масштабируемость и более эффективный контроль доступности данных.

Нереляционные базы данных часто обозначают акронимом NoSQL (который означает «Not Only SQL» — «не только SQL»). Основным отличительным признаком нереляционных БД является сама структура хранилища, где данные более не привязаны к соотнесенным между собой реляционным таблицам. Это может быть структура в виде дерева, графа, сети или пар «ключ-значение». Обозначение NoSQL в данном случае не совсем подходит, поскольку некоторые версии

вполне поддерживают все стандартные команды языка SQL. Нереляционные БД часто представляют собой склады данных (data stores), в высокой степени оптимизированные для выполнения простых операций извлечения и добавления. Основной целью таких баз данных является повышение производительности, прежде всего за счет уменьшения времени ожидания и повышения пропускной способности. Базы данных NoSQL находят всё более широкое применение в области больших данных и при создании веб-приложений, работающих в режиме реального времени (см. главу 5).

1.3.8.3.1 Колоночные БД

Колоночные базы данных используются преимущественно в приложениях для бизнес-аналитики (BI), поскольку позволяют существенно сжимать избыточные данные. Например, колонка (столбец) с идентификаторами статуса чего-либо содержит только уникальные значения строго по одному разу, которые не дублируются в миллионах строк, как это бывает в таблицах реляционных БД.

При выборе между колоночной (нереляционной) и строчно-ориентированной (реляционной) схемой организации данных следует учитывать следующие факторы.

- Колоночная организация более эффективна, когда требуется суммирование или усреднение данных по множеству строк лишь в незначительной части столбцов, поскольку в таком случае выборочное считывание столбцов заметно ускоряет расчеты по сравнению с построчным считыванием всей таблицы.
- Колоночная организация более эффективна, когда новыми значениями заполняется сразу весь столбец (соответствующее столбцу поле обновляется сразу во всех строках), поскольку старые данные замещаются новыми только в отдельном столбце, а остальные поля во всех строках не обрабатываются.
- Строчно-ориентированная организация более эффективна, когда необходим одновременный доступ к данным во многих полях одной и той же строки, а сама строка достаточно коротка для того, чтобы быть считанной за одно обращение к жесткому диску.
- Строчно-ориентированная организация более эффективна при записи новой строки, если одновременно доступны все поля строки: вся строка записывается за одно обращение к диску.
- Практика показывает, что строчно-ориентированная схема хорошо справляется с рабочими нагрузками, характерными для задач типа оперативной обработки транзакций (OnLine Transaction Processing, OLTP), предполагающих высокую интенсивность выполнения интерактивных операций. Колоночная схема, в свою очередь, лучше подходит для решения задач типа оперативной аналитической обработки (OnLine Analytical Processing, OLAP), в частности для организации хранилищ данных, где запросы относительно немногочисленны, но крайне сложны по структуре и требуют просмотра всего массива данных (который может занимать терабайты памяти).

1.3.8.3.2 Пространственные БД

Пространственная база данных (spatial database) оптимизирована для хранения и выдачи данных об объектах, определенных в геометрическом пространстве. Пространственные БД поддерживают несколько простейших типов геометрических фигур (квадрат, прямоугольник, куб, цилиндр и т. д.) и геометрические композиции из наборов точек, прямых, кривых и фигур.

Системы управления пространственными базами данных используют индексы для ускорения поиска запрошенных значений. Обычные алгоритмы индексирования для обработки пространственных запросов не подходят, и вместо них используется особое пространственное индексирование, позволяющее ускорить операции с базой данных.

Пространственные БД могут выполнять широкий спектр разнообразных операций. Согласно стандартам Open Geospatial Consortium[1], пространственная БД может выполнять следующие операции (или некоторые из них).

- **Пространственные измерения.** Расчет длины отрезков, площади многоугольников, расстояния между объектами и т. д.
- **Пространственные функции.** Модификация существующих объектов (например, окружение буферной зоной) и создание новых объектов из существующих (слияние, пересечение и т. п.).
- **Пространственные предикаты.** Обработка логических запросов (истина/ложь) о геометрических соотношениях. Примеры запросов: «Пересекаются ли две указанные геометрические фигуры?»; «Имеется ли жилая застройка в радиусе километра от планируемого полигона для захоронения отходов?».
- **Геометрические построения.** Создание новых геометрических фигур — как правило, с помощью указания вершин (точек или узлов).
- **Функции обозрения.** Обработка запросов на предоставление информации о характеристиках объекта (например, координат центра окружности).

1.3.8.3.3 Мультимедийные БД

Мультимедийные базы данных включают систему управления иерархическим хранилищем мультимедийных объектов на магнитных и оптических носителях. В состав системы входит набор классов объектов, образующих ее основу.

1.3.8.3.4 БД на основе плоского файла

База данных на основе плоского файла (flat file database) представляет данные в виде единственного файла. Файл может быть текстовым или двоичным. Строго говоря, БД в виде файла не содержит ничего кроме полей данных, которые могут отличаться по длине и иметь разные разделители. В более широкой трактовке к этой же категории относят и базы данных, сохраненные

[1] Open Geospatial Consortium (OGC, *досл.* «Открытый геопространственный консорциум») — существующая с 1994 г. международная некоммерческая организация, занимающаяся разработкой и согласованием единых стандартов геопространственных данных и сервисов. — *Примеч. пер.*

в единственном файле в формате таблицы, но в таком файле отсутствует какая-либо информация об отношениях или связях между записями и полями, за исключением самой структуры таблицы. Текстовый файл обычно содержит по одной записи в строке. Примером БД на основе плоского файла является список фамилий, имен, адресов и телефонов, написанный от руки на листе бумаги. Плоские файлы могут использоваться не только в качестве способа хранения данных в СУБД, но и для обмена данными между системами. Базы данных Hadoop используют плоские файлы в качестве средства хранения данных.

1.3.8.3.5 БД «ключ-значение»

В базах данных типа «ключ-значение» (key-value pair databases) содержатся исключительно попарно организованные элементы: ключ-идентификатор и значение. Вариантов применения такого рода баз данных немного. К ним относятся следующие.

- **Документальные БД.** Документоориентированные базы данных содержат наборы файлов (документов), включающих и структуру, и данные. Каждому документу присваивается ключ. Более развитые документоориентированные БД могут хранить также и некоторые атрибуты, касающиеся содержания документов, — например, даты или теги. В таких базах данных допускается хранение незавершенных документов наряду с завершенными. Документоориентированные базы данных могут использовать структуры языков XML или JSON (объектная нотация JavaScript — Java Script Object Notation).
- **Графовые БД.** В графовых базах данных сохраняемые пары «ключ-значение» описывают связи между узлами (вершинами графов), а не сами узлы.

1.3.8.3.6 Хранилища триплетов

Трехэлементная связка данных «субъект-предикат-объект» называется триплетом. В модели консорциума W3C для описания ресурсов (Resource Description Framework, RDF) триплет скомпонован из субъекта, который обозначает ресурс, предиката, который отражает связь между субъектом и объектом, и объекта. Хранилища триплетов представляют собой узкоспециализированные базы данных, предназначенные для хранения и извлечения триплетов в форме выражений «субъект-предикат-объект».

Хранилища триплетов можно разделить на три общие категории: специализированные; реализованные на основе реляционных СУБД; нереляционные хранилища.

- **Специализированные хранилища** создаются с нуля с изначальной ориентацией на модель RDF в целях обеспечения эффективного хранения и доступа к данным.
- **Хранилища на основе реляционных СУБД** строятся путем добавления специального функционального уровня RDF к функциональности существующей реляционной СУБД.
- **Хранилища NoSQL** в настоящее время исследуются на предмет возможности их использования в качестве средства управления хранением данных в рамках RDF-модели.

Хранилища триплетов являются лучшим решением для управления таксономиями и тезаурусами, интеграции взаимосвязанных данных и создания всевозможных порталов знаний.

1.3.9 Специализированные базы данных

В некоторых случаях требуется использование специализированных баз данных. Управление ими отличается от управления традиционными реляционными базами данных. Вот лишь некоторые примеры.

- **Системы автоматизированного проектирования и подготовки производства (CAD/CAM)** используют объектные базы данных; то же самое касается большинства встроенных приложений, работающих в режиме реального времени.
- **Географические информационные системы (ГИС**) используют специализированные геопространственные базы данных, справочные данные в которых обновляются как минимум ежегодно. Узкоспециализированные ГИС используются в управлении коммунальным хозяйством (электросетями, системами газоснабжения и т. п.) и телекоммуникационными сетями, в морских навигационных системах и т. п.
- **Приложения, обслуживающие корзины покупок** на сайтах большинства компаний розничной торговли, используют базы данных XML для сохранения данных о заказах. Такие базы также могут использоваться в режиме реального времени системами социальных медиа для размещения информации на веб-сайтах.

Некоторые данные из специализированных систем затем копируются в более традиционные базы данных систем оперативной обработки транзакций (OLTP) или хранилища данных. Кроме того, многие коммерческие приложения используют собственные проприетарные решения, схемы которых являются коммерческой тайной и тщательно скрываются даже в тех случаях, когда построены они на использовании традиционных реляционных СУБД.

1.3.10 Общие процессы обработки данных в базах данных

В базах данных любого типа в том или ином виде реализованы следующие процессы.

1.3.10.1 АРХИВИРОВАНИЕ

Архивированием называют процесс перемещения данных из операционной среды на носители, не поддерживающие прямого доступа к данным. При необходимости оперативного использования архивных данных их можно извлекать обратно в исходную систему. Данные, которые активно не используются приложениями и процессами, следует перемещать в архивы, хранящиеся на недорогих дисках, магнитной ленте или CD/DVD. При этом процедура восстановления данных из архива должна быть предельно простой и сводиться к копированию данных из архива обратно в систему.

Процедуры архивирования должны планироваться в привязке к стратегии создания разделов хранилищ, чтобы обеспечить удобный доступ к архивным данным и их сохранность. Для повышения надежности рекомендуется использовать комплексный подход к архивированию, включающий следующие операции:

- создание вторичного хранилища, желательно на резервном сервере БД;
- разбивка существующих таблиц БД на архивные блоки;
- перенос нечасто используемых данных в отдельную БД;
- создание резервных копий всех данных на магнитной ленте или дисках;
- создание автоматически выполняемых заданий (jobs), которые осуществляли бы периодическую проверку актуальности и удаление устаревших и ненужных данных.

Настоятельно рекомендуется регулярно тестировать работоспособность процедуры восстановления данных из архивных копий во избежание неприятных сюрпризов в случае реальных аварийных сбоев в работе основного хранилища.

При внесении технологических или структурных изменений в эксплуатируемую систему нужно убедиться в совместимости обновленной версии системы с архивными данными. С архивами, утратившими согласованность с текущей версией системы, следует поступать по обстоятельствам.

- Определите, какая часть архивных данных по-настоящему нужна и обязательна для сохранения. Всё остальное можно из архива просто удалить.
- Перед проведением крупных технологических изменений восстановите архивные данные в действующую систему (до изменения технологии), обновите систему до технологически новой версии или проведите миграцию на новую версию, после чего повторно заархивируйте данные с использованием новой технологии.
- Если речь идет о ценных архивных данных, то в случае изменения структуры БД восстановите данные из архива, внесите изменения в их структуру в соответствии с новой моделью и повторно заархивируйте реструктурированные данные.
- Если речь идет о редко используемых архивах, то в случае изменения исходной технологии или структуры БД сохраните упрощенную версию старой системы с ограниченным доступом — и используйте ее исключительно для извлечения данных из архивов в целях переноса в новую систему по мере надобности.

Архивы, полностью несовместимые с новыми технологиями или невосстановимые с помощью имеющихся программно-аппаратных средств, бесполезны и подлежат удалению, поскольку содержать на балансе устаревшее оборудование исключительно ради считывания архивов экономически нецелесообразно и даже контрпродуктивно.

1.3.10.2 ПРОГНОЗИРОВАНИЕ РОСТА ТРЕБУЕМОЙ ЕМКОСТИ БД

Представьте себе базу данных в образе коробки для фруктов, данные — в образе фруктов, а надстройку над БД (индексы и т. п.) — в роли обертки для фруктов. Коробка внутри разделена перегородками на ячейки, куда укладываются фрукты в обертке. А теперь задайтесь следующими вопросами.

- Прежде всего, какого размера коробка нужна, чтобы в нее гарантированно поместились все фрукты со всеми обертками? Ответ на этот вопрос дает оценку емкости.
- Сколько фруктов добавляется в коробку и с какой скоростью?
- Сколько фруктов изымается из коробки и с какой скоростью?

Теперь определитесь, нужно или не нужно со временем увеличивать размеры и вместимость коробки. То есть нужно спрогнозировать темпы увеличения емкости коробки, которые обеспечат возможность размещать все поступающие фрукты плюс обертку, исходя из прогноза роста объемов поступления. Если коробка расширению не поддается, значит, нужно запланировать изъятие из нее старых фруктов теми же темпами, которыми поступают новые, и в таком случае прогноз роста будет нулевым.

Каким должен быть срок хранения фруктов в ячейках коробки? Если фрукт в какой-то ячейке засох или по иным причинам сделался не столь полезен, как прежде, то как с ним поступать? Переложить в дальнюю коробку (то есть заархивировать)? А потребуется ли когда-либо возвращать его в основную коробку? Перекладывание фруктов в другую коробку с возможностью всегда вернуть их на прежнее место в главной коробке — в этом и заключается важнейшая роль архивирования. Оно избавляет от нужды слишком часто расширять основную коробку или заменять ее более вместительной.

Но если фрукт сгнил и вовсе ни на что не годен — отправляйте его на помойку (то есть стирайте данные без возможности восстановления).

1.3.10.3 РЕГИСТРАЦИЯ ИЗМЕНЕНИЙ ДАННЫХ

Регистрацией изменений данных (Change Data Capture, CDC) называют процесс выявления факта изменений и сохранения исчерпывающей информации о них. Процедура DCD, которую также часто называют репликацией на основе журнала (log-based replication), дает возможность воспроизводить в целевой системе (с минимальным вмешательством в ее работу) изменения в данных, произошедшие в системе-источнике (без какого-либо влияния на работу последней). В предельно упрощенном контексте процедура CDC сводится к следующему: предположим, что в одной компьютерной системе данные только что изменились, и теперь нужно отразить ровно те же изменения в другой компьютерной системе; и тогда вместо отправки по сети обновленной версии всей базы данных ради внесения незначительных изменений в отдельные элементы данных в целевую систему отправляются только минимально необходимые для надлежащего обновления данных характеристики изменений (так называемые «дельты»).

Существуют два основных метода выявления изменений и сбора информации об изменениях. Первый из них заключается в контроле версий, что позволяет выявлять строки с изменившимися данными по какому-либо столбцу (например, указывающему время последнего обновления, номер версии или статус обновления строки); второй подразумевает считывание журналов, документирующих изменения, а затем — их воспроизведение во вторичных системах.

1.3.10.4 СТИРАНИЕ ДАННЫХ

Наивно полагать, что все накопленные данные могут храниться в основном хранилище вечно. Рано или поздно оно переполнится, а производительность СУБД упадет. И тогда так или иначе придется решать вопрос об архивации и/или стирании избыточных данных. Не менее важно и то, что часть данных со временем обесценивается — и затраты на их хранение перестают окупаться. Стиранием (purging) называется процедура безвозвратного и необратимого удаления данных с носителей. Принципиальная задача управления данными — следить за рентабельностью базы данных, то есть за тем, чтобы затраты на их сопровождение не превышали финансовую отдачу, получаемую организацией. Стирание данных избавляет от дальнейших издержек и рисков. В общем случае стиранию подлежат все данные, которые оцениваются как устаревшие или ненужные и не требующиеся для формальной отчетности в силу действующих законов и регламентов. Кроме того, за хранение некоторых данных дольше установленного законами или регламентами срока организацию могут еще и привлечь к ответственности. Наконец, безвозвратное стирание данных сводит к нулю и риск злоупотребления ими.

1.3.10.5 РЕПЛИКАЦИЯ ДАННЫХ

Репликация (replication) данных означает, что одни и те же данные хранятся на многих запоминающих устройствах. В некоторых ситуациях наличие дублирующих друг друга баз данных (реплик) просто необходимо: например, в операционных средах, требующих быстрого и бесперебойного доступа к данным, где распределение рабочей нагрузки между многими серверами или даже вычислительными центрами с идентичными базами данных позволяет сохранить функциональность системы в период пиковых нагрузок или в случае аварий.

Репликация может быть активной или пассивной.

- **Активная репликация** подразумевает, что вслед за проведением каких-либо операций по обновлению и сохранению данных в любой из реплик, в каждой из остальных реплик воспроизводятся аналогичные операции по обновлению и сохранению.
- **Пассивная репликация** подразумевает проведение обновления и сохранения данных в единственной первичной реплике с последующим переносом ее конечного видоизмененного состояния во вторичные реплики.

Репликация данных обеспечивает масштабирование (scaling) по двум направлениям (в двух измерениях — dimensions).

- Горизонтальное масштабирование данных заключается в создании дополнительного количества реплик.
- Вертикальное масштабирование данных заключается в создании реплик в местах, расположенных на всё большем географическом удалении от первичной реплики.

Наиболее предпочтительный вариант репликации подразумевает, что изменения могут быть внесены в любой из сетевых узлов БД, после чего эти изменения распространяются по другим серверам как круги по воде (multi-master replication). Одновременно с удобством этот вариант также несет множество технических сложностей и финансовых затрат.

Прозрачность репликации (replication transparency) возникает, когда обеспечивается такой уровень согласованности данных во всех репликах, при котором пользователи не могут сказать или даже не знают, с какой копией базы данных они работают.

Два основных метода репликации данных — зеркальное отображение (mirroring) и доставка журналов (log shipping) (см. рис. 60).

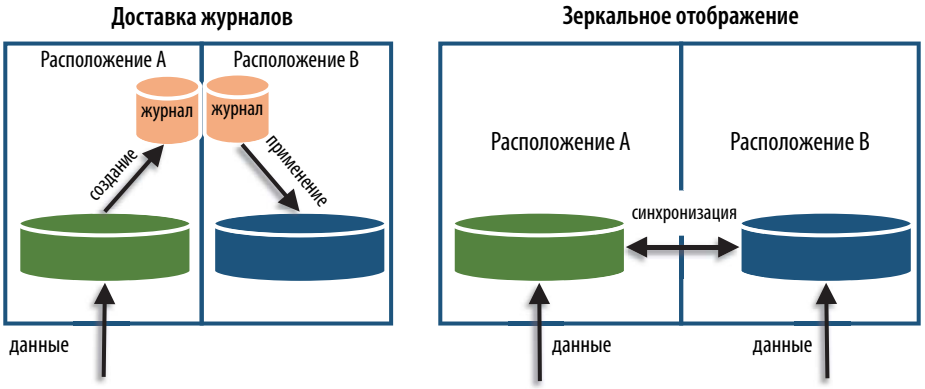


Рисунок 60. Альтернативные методы репликации данных

- При зеркальном отображении обновления в первичной базе данных мгновенно (в смысле «максимально оперативно») воспроизводятся во всех вторичных БД в рамках процесса выполнения протокола двухфазного подтверждения транзакции (Two-phase Commit Protocol, 2PC).
- При доставке журналов вторичный сервер с установленной периодичностью получает от первичного сервера БД и применяет копии журналов транзакций.

Выбор метода репликации зависит от критичности данных и требуемой срочности приведения данных на вторичных серверах в соответствие с данными на первичном сервере БД с точки зрения обеспечения отказоустойчивости системы в целом. Зеркальное отображение, как правило, обходится значительно дороже доставки журналов. Для одного вторичного сервера использование зеркального отображения — вполне оправданное и эффективное решение; проводить обновления на дополнительных вторичных серверах можно с помощью доставки журналов.

1.3.10.6 ОТКАЗОУСТОЙЧИВОСТЬ И ВОССТАНОВЛЕНИЕ РАБОТОСПОСОБНОСТИ

Применительно к базам данных под отказоустойчивостью понимается способность системы функционировать в условиях возникновения ошибок. Если система продолжает функционировать и выдавать ожидаемые результаты даже при большом количестве ошибок обработки данных, она отказоустойчива. Если же приложение прекращает функционировать при первой же непредвиденной ситуации, такой системе отказоустойчивости явно недостает. Если СУБД умеет выявлять распространенные ошибки обработки (например, некорректные запросы) и либо останавливать задачу, либо автоматически восстанавливаться до работоспособного состояния, она отказоустойчива. Но всегда имеются и внешние причины отказов, которые никакая система не способна ни спрогнозировать, ни избежать (например, отключение электропитания), однако в этом случае речь идет уже об аварийных условиях.

При выработке рекомендаций относительно ускорения процесса восстановления и приоритетности выполняемых шагов обычно выделяют три типа задач по восстановлению.

- **Немедленное восстановление (immediate recovery)** после некоторых ожидаемых сбоев, возможности по обеспечению которого иногда предусматриваются еще на стадии проектирования: например, прогнозирование и автоматическое решение проблем — в частности, путем переключения на резервную систему.
- **Критическое восстановление (critical recovery)** предусматривает наличие плана максимально быстрого восстановления работоспособности системы с целью минимизировать задержку или время остановки бизнес-процессов.
- **Некритическое восстановление (non-critical recovery)** означает, что восстановление функций может быть отложено до того момента, когда будет завершено критическое восстановление системы.

Ошибки обработки данных включают сбои при чтении/записи данных, выполнении запросов, операциях извлечения, преобразования и загрузки данных и т. п. Стандартные способы повышения отказоустойчивости систем обработки данных включают прерывание обработки и перенаправление данных, вызывающих ошибки; выявление типов данных, вызывающих ошибки, и разработку инструкций по исключению таких данных из процесса обработки; реализация меток завершения этапов обработки данных во избежание зацикливания или повторения ранее проделанных операций при перезапуске процесса.

Отказоустойчивость в той или иной мере требуется любой системе; весь вопрос в том, насколько она должна быть высокой или низкой. Некоторые приложения просто по характеру реализованных в них процессов вовсе не допускают ошибок обработки данных и останавливаются при первой же ошибке (низкая отказоустойчивость), в то время как менее чувствительные к ошибкам приложения могут отлавливать ошибки и перенаправлять сбойные данные на анализ или же просто игнорировать их.

В случае критически важных данных администраторам БД нужно предусмотреть их обязательную репликацию с размещением копии БД на удаленном сервере. Такая схема при отказе первичной БД приложения позволит переключиться на удаленную (резервную) БД и продолжить работу.

1.3.10.7 СОХРАНЕНИЕ ДАННЫХ

Сохранение данных (data retention) связано с выработкой решений относительно того, как долго данные остаются доступными. Планирование сохранения данных является частью проектирования физической БД. При этом требования по сохранению влияют на планирование емкости БД.

Соображения информационной безопасности также оказывают воздействие на планы сохранения данных, поскольку некоторые данные требуется хранить не меньше или не дольше установленного законом срока, а несоблюдение этих требований может повлечь юридические последствия. То же самое касается и обязательных нормативных требований, касающихся стирания данных. За превышение допустимого законом срока хранения данных организация также рискует быть привлеченной к ответственности. Потому организация должна разрабатывать четкие политики сохранения данных, основанные на требованиях регулирующих органов и рекомендациях по управлению рисками. Эти же политики должны определять и основные подходы в отношении практики стирания и архивирования устаревших данных.

1.3.10.8 СЕГМЕНТИРОВАНИЕ (ШАРДИНГ)

Сегментирование (или шардинг — sharding) — это разбиение базы данных на независимые друг от друга полностью изолированные небольшие блоки (мелкие части — shards), каждый из которых может обновляться независимо от других блоков. При такой сегментированной архитектуре репликация сводится к простому копированию файлов. Поскольку размеры блоков невелики, оптимальным может являться обновление путем обычной перезаписи.

2. ПРОВОДИМЫЕ РАБОТЫ

Два основных направления работ в области хранения и операций с данными — технологическая поддержка и операционное сопровождение баз данных. Технологическая поддержка БД начинается с выбора программного обеспечения, используемого для хранения и управления данными, и в дальнейшем сводится к деятельности по обеспечению его работоспособности. Содержание работ по операционному сопровождению баз данных зависит от характера данных и процессов, находящихся под управлением этого программного обеспечения.

2.1 Управление технологиями баз данных

При управлении технологиями баз данных следует руководствоваться теми же принципами и стандартами, которые распространяются на любые другие информационные технологии.

Главной эталонной моделью в области управления информационными технологиями на сегодняшний день остается библиотека инфраструктуры информационных технологий (Information Technology Infrastructure Library, ITIL) — процессная модель, созданная в Великобритании. Принципы ITIL полностью применимы и к управлению технологиями баз данных[1].

2.1.1 Изучение и углубление понимания характеристик технологий баз данных

Важно понимать, во-первых, как работает используемая технология, а во-вторых, как извлечь из нее максимум выгоды в контексте конкретного бизнеса. Администраторы БД совместно с другими командами по обслуживанию данных должны в тесном сотрудничестве с бизнес-пользователями и менеджерами точно определять потребности организации в данных и информации. АБД следует объединять усилия с архитекторами баз данных, чтобы в полной мере использовать имеющиеся у тех и у других знания и как можно более эффективно применять технологии для удовлетворения потребностей бизнеса.

Профессионалы в области управления данными должны хорошо изучить характеристики выбираемых технологий баз данных, прежде чем рекомендовать конкретные решения. Например, технологии, не поддерживающие функциональность управления транзакциями (в частности, операции подтверждения (commit) и отката (rollback) транзакций), не подходят для использования в системах продаж.

Недопустимо считать какую-либо единственную архитектуру или СУБД универсально применимой для удовлетворения всех потребностей. В большинстве случаев в одной и той же организации установлено довольно много программных средств управления базами данных, поддерживающих исполнение широкого спектра функций — от настройки БД на максимальную производительность до резервного копирования, включая собственно управление базой данных. И лишь в единичных случаях такие программные комплексы удовлетворяют каким-либо обязательным стандартам.

2.1.2 Комплексная оценка технологии баз данных

Выбор СУБД для использования в стратегически значимых целях особенно важен. От этого зависят и схемы интеграции данных, и производительность приложений, и эффективность ведения бизнеса. Ниже перечислена лишь часть факторов, которые следует принимать во внимание при выборе СУБД:

- архитектура и уровень сложности продукта;
- ограничения по объему и скорости получения, обработки и передачи данных;
- профиль приложения (обработка транзакций, бизнес-аналитика, управление личными профилями и т. п.);
- поддержка специфической функциональности (например, расчета показателей, зависящих от времени);

[1] http://bit.ly/1gA4mpr

- требования к аппаратной платформе и операционной системе;
- доступность вспомогательного программного обеспечения;
- производительность, демонстрируемая на эталонных тестах, в том числе в режиме реального времени;
- масштабируемость;
- требования к программному обеспечению, оперативной памяти и объему внешних накопителей;
- отказоустойчивость, включая обработку ошибок и формирование отчетов о них.

Кроме того, следует учитывать и некоторые факторы, относящиеся не столько к технологии как таковой, сколько к различным организационным и финансовым аспектам приобретения тех или иных программных продуктов. Например:

- готовность организации к техническому риску;
- наличие кадров с надлежащим уровнем профессиональной технической подготовки;
- стоимость владения (стоимость лицензии, технического обслуживания, вычислительных ресурсов и т. п.);
- репутация поставщика;
- политика поставщика в отношении поддержки и частота выпуска обновлений;
- отзывы пользователей.

Расходы на приобретение и сопровождение программного продукта, включая администрирование, лицензирование и техническую поддержку, не должны превышать ожидаемую отдачу от его использования в бизнесе. В идеале технология должна быть как можно более удобной в использовании и оснащенной средствами самостоятельного контроля и администрирования. Если эти требования не соблюдены, то, возможно, целесообразно пригласить на работу специалиста, обладающего опытом работы с выбранной системой.

Хорошим началом может стать небольшой пилотный проект или опытная проверка концепции (Proof-of-Concept, POC), чтобы лучше оценить истинное соотношение затрат и выгод, прежде чем переходить к полномасштабному внедрению системы в эксплуатационную среду.

2.1.3 Управление и мониторинг технологий баз данных

Администраторы БД часто выполняют функции по технической поддержке второго уровня, наряду со службами технической поддержки (help desks) и командами поддержки от производителя/поставщика ПО, работая над изучением, анализом и решением проблем, с которыми сталкиваются пользователи. Ключом к полноценному пониманию и использованию любой технологии является обучение. Организациям следует планировать и финансировать переподготовку всех, кто будет участвовать во внедрении, техническом сопровождении и использовании технологий баз данных. Планы подготовки должны обеспечивать должный уровень знаний в смежных областях,

чтобы наилучшим образом способствовать разработке прикладных решений, прежде всего с использованием методик гибкой разработки (agile development). АБД должны обладать практическими навыками разработки приложений, включая моделирование данных, анализ сценариев использования и управление доступом приложений к данным.

АБД персонально отвечает за обеспечение регулярного резервного копирования БД и тестирование системы восстановления данных из резервных копий. Однако если потребуется слияние данных в имеющихся базах с другими базами, то у администраторов могут возникнуть проблемы с интеграцией. Поэтому АБД следует не просто заниматься непосредственно объединением (слиянием) данных, а детально прорабатывать с другими заинтересованными сторонами все детали предстоящей интеграции данных, чтобы обеспечить ее корректность и эффективность.

Когда бизнесу требуется новая технология, АБД должны совместно с бизнес-пользователями и разработчиками приложений определить наиболее эффективные пути использования данной технологии, исследовать новые прикладные возможности ее применения и разобраться с вероятными проблемами, которые могут возникнуть при ее использовании. Затем АБД развертывают продукты, реализующие новые технологические решения в предэксплуатационной среде, а после — в среде эксплуатации. При этом им нужно разработать (и документально оформить) процессы и процедуры администрирования продукта, ориентированные на минимальные затраты времени и финансовых ресурсов.

2.2 Управление базами данных

Сопровождение баз данных можно назвать «сердцем» управления данными. База данных размещается в управляемой среде хранения данных (managed storage). Управляемая среда хранения может быть небольшой, как дисковод на персональном компьютере (управляемый операционной системой), или очень объемной, как RAID-массивы в сети хранения данных (SAN). Накопители с резервными копиями баз данных также относятся к управляемым средам хранения.

АБД управляют различными программными приложениями, работающими со средами хранения, посредством подготовки структур данных, осуществления технического сопровождения физических баз данных (включая физические модели и физические представления данных — например, назначение файлов или областей дисков), установки и конфигурирования СУБД на серверах.

2.2.1 Изучение и углубление понимания требований

2.2.1.1 ОПРЕДЕЛЕНИЕ ТРЕБОВАНИЙ К ЕМКОСТИ УСТРОЙСТВ ХРАНЕНИЯ ДАННЫХ

АБД устанавливают требования к системам хранения данных для СУБД и файловым хранилищам, поддерживающим базы данных NoSQL. Также АБД (совместно с администраторами сетей хранения данных) играют ключевую роль в организации систем хранения данных и файловых хранилищ. Данные заносятся на носители устройств хранения в процессе выполнения бизнес-операций, а затем, в зависимости от установленных требований, могут оставаться там постоянно или временно. Крайне важно заранее планировать потребности в хранении дополнительных

объемов данных и расширять емкости устройств хранения, прежде чем дадут о себе знать первые признаки их дефицита. Любые технические работы в авральном режиме сопряжены с риском системных сбоев и утери данных.

Все проекты должны предусматривать проведение оценки требуемой емкости устройств хранения на первый год эксплуатации, а также прогнозирование потребностей в дальнейшем расширении емкости еще на несколько лет. Емкость устройств хранения и потребности по ее наращиванию должны оцениваться исходя из предполагаемого объема не только данных как таковых, но и пространства, которое потребуется для хранения индексов, журналов и всевозможных дополнительных копий данных, таких как зеркала.

При планировании емкости устройств хранения должны учитываться требования регулирующих органов по сохранению данных (data retention). Законами и регламентами может быть предписано обязательное сохранение данных того или иного вида в течение определенного периода времени (см. главу 9). В некоторых случаях организации, напротив, обязаны стирать данные не позднее установленного законами срока. Рекомендуется обсуждать потребности в сохранении и допустимые сроки хранения данных с их владельцами, а также согласовывать с ними порядок обращения с данными на протяжении всего их жизненного цикла.

Кроме того, АБД, совместно с разработчиками приложений и другим техническим персоналом, включая администраторов систем хранения и серверов, должны проводить работу по согласованию и внедрению плана сохранения данных.

2.2.1.2 ОПРЕДЕЛЕНИЕ ШАБЛОНОВ ИСПОЛЬЗОВАНИЯ

Базы данных обычно вполне предсказуемы в плане шаблонов их использования (usage patterns). Типичные шаблоны использования (схемы распределения нагрузки на базы данных):

- пропорционально потоку обрабатываемых транзакций;
- пропорционально объемам записываемых или извлекаемых массивов данных;
- привязанные к определенным периодам времени пики/спады (повышение нагрузки в конце месяца, спад по выходным и т. п.);
- географическое распределение (в густонаселенных местностях нагрузка выше, чем в малонаселенных, и т. п.);
- в зависимости от приоритетности задач (некоторым отделам или пакетным запросам может присваиваться повышенный приоритет).

В некоторых СУБД могут наблюдаться различные комбинации вышеперечисленных шаблонов использования. Администраторам БД нужно уметь прогнозировать всплески и спады спроса на данные — и предусмотреть ограничения на обработку запросов в периоды пикового спроса (например, правила распределения запросов по очередям обработки или управления обработкой запросов в порядке их приоритетности), а также в полной мере задействовать ресурсы системы в периоды спада нагрузки (то есть откладывать обработку ресурсоемких запросов и задач до

периода устойчиво низкой загруженности системы). Собираемая информация о распределении нагрузки может также использоваться при планировании мер по повышению производительности баз данных.

2.2.1.3 ОПРЕДЕЛЕНИЕ ТРЕБОВАНИЙ ПО ДОСТУПУ К ДАННЫМ

Деятельность по сохранению, извлечению и изменению данных в БД или ином информационном хранилище предполагает наличие доступа к этим данным. Проще говоря, доступ к данным подразумевает авторизацию выполнения действий с различными наборами данных.

Доступ к данным, записанным в БД или иное хранилище, может быть реализован с помощью различных стандартных языков программирования, методов и форматов. В системах, ориентированных на обработку данных в соответствии с принципами ACID, используются языки SQL, ODBC, JDBC, XQJ, ADO.NET, XML, X Query, X Path и веб-сервисы. Методы доступа к системам, организованным в соответствии с принципами BASE, реализуются на языках C, C++, REST, XML и Java[1]. Некоторые стандарты поддерживают перевод данных из неструктурированных форматов (например, HTML или тестовых файлов) в структурированные (например, XML или SQL).

Архитекторы данных и администраторы БД могут и должны оказывать помощь организациям в выборе подходящих методов и средств доступа к данным.

2.2.2 Планирование непрерывности бизнеса

Организациям нужно планировать мероприятия по обеспечению непрерывности бизнеса (business continuity) в случае аварий или чрезвычайных ситуаций, приводящих к выходу из строя каких-либо систем и нарушению доступа к данным. АБД должны иметь надежный план восстановления работоспособности всех баз данных и серверов БД на случай любых событий, способных привести к утере или повреждению данных, включая, в частности:

- выход из строя сервера базы данных;
- выход из строя дисковых запоминающих устройств;
- выход из строя базы данных, хранилищ временных данных, журналов транзакций и т. п.;
- повреждение индекса или страниц данных;
- повреждение файловых систем, поддерживающих работу с базами данных и журналами;
- потерю файлов с резервными копиями баз данных и журналов транзакций.

Каждая база данных подлежит тщательной оценке с точки зрения ее критичности для бизнеса и мер по обеспечению максимально быстрого восстановления. Некоторые базы данных настолько значимы для бизнес-операций, что сбои в их работе вовсе недопустимы, а доступ к данным должен восстанавливаться путем мгновенного переключения на резервные системы. Восстановление доступа к не столь критичным для обработки текущих операций базам данных может быть

[1] Полный список методов доступа к нереляционным базам данных см.: http://bit.ly/1rWAUxS

отложен до восстановления работоспособности основных систем. Впрочем, системный сбой иногда бывает и неожиданно полезным средством выявления данных, которые можно и вовсе не восстанавливать, — к примеру, множественных резервных копий, созданных перед установкой обновлений.

Ответственные менеджеры и группа обеспечения непрерывности бизнеса (если такая группа в организации существует) должны рассмотреть и утвердить план аварийного восстановления данных. Группа АБД в идеале должна отвечать за регулярный пересмотр плана с целью обеспечения его корректности и полноты. Обязательно должны быть в наличии копия плана, программное обеспечение, необходимое для инсталляции и конфигурирования СУБД, инструкции, а также коды доступа (пароли администратора и т. п.) к безопасной удаленной среде, предусмотренной на случай аварии.

Никакая система не может быть восстановлена после аварийного сбоя, если отсутствуют или не считываются резервные копии. Регулярное резервное копирование данных — обязательное условие обеспечения их восстановления, но если копии не считываются, то утрачивается последняя возможность привести базу в работоспособное состояние. В таком случае время, ушедшее на резервное копирование, потрачено впустую, а определить причину нарушения целостности копии очень сложно. Поэтому хранить резервные копии следует в надежном удаленном месте.

2.2.2.1 СОЗДАНИЕ РЕЗЕРВНЫХ КОПИЙ

Делайте резервные копии баз данных (и, если нужно, журналов транзакций) с частотой не реже указанной в действующем соглашении об уровне обслуживания системы. Соизмеряйте важность данных с затратами на их защиту. В случае больших баз данных регулярное полное резервное копирование может оказаться слишком ресурсоемким решением (как с точки зрения объемов требуемой емкости хранилищ, так и с точки зрения загрузки мощностей серверов). Поэтому регулярно делайте инкрементные резервные копии (incremental backups) для каждой БД, и периодически — полные. Кроме того, базы данных должны храниться в управляемых средах хранения, в идеале — на RAID-массивах дисков в сети хранения данных (SAN), откуда резервные копии раз в сутки переносятся в отдельное хранилище. В случае баз данных OLTP частота создания резервных копий журналов транзакций будет зависеть от частоты обновлений и объема данных. Чем чаще и чем в бóльших объемах обновляются базы данных, тем чаще нужно сохранять журналы обновлений в резервных копиях, — тем самым вы не только обеспечите более надежную защиту данных, но и не допустите перегрузки серверов и приложений, которая возникла бы при резервном копировании слишком объемных журналов.

Резервные копии должны храниться в отдельной от баз данных файловой системе на устройстве хранения, предусмотренном соглашением об уровне обслуживания. Ежедневные резервные копии БД храните в безопасном и удаленном месте. Большинство СУБД поддерживают создание резервных копий без приостановки доступа приложений к рабочим базам данных. Если в процессе подобного «горячего» резервирования какие-то данные обновляются в результате обработки текущей транзакции, система либо дожидается ее завершения и затем создает резервную копию,

либо откатывает транзакцию, создает копию и затем повторяет обработку. Альтернативный вариант — «холодное» резервирование БД в офлайновом режиме. При всей надежности такого метода он не подходит для систем, в которых приложениям безоговорочно требуется непрерывный доступ к данным

2.2.2.2 ВОССТАНОВЛЕНИЕ ДАННЫХ ИЗ РЕЗЕРВНЫХ КОПИЙ

В большинстве СУБД предусмотрена возможность автоматического восстановления баз данных из резервных копий. АБД следует проработать с техническим персоналом, отвечающим за аппаратную инфраструктуру, порядок подключения носителей с резервной копией и выполнение процедуры восстановления. Утилиты, предназначенные для восстановления, и процедуры восстановления БД различаются в зависимости от используемой СУБД.

В базах данных на основе файлов процедуры восстановления, в целом, проще, чем в реляционных СУБД, поскольку в первом случае достаточно сверки каталогов и восстановления лишь измененных файлов, а реляционные БД приходится восстанавливать целиком.

Критическое значение имеет периодическое тестирование работоспособности процедуры восстановления БД. Никогда не будет лишним удостовериться, что всё под контролем, с целью избежать неприятных сюрпризов при реальной аварии или чрезвычайной ситуации. Отрабатывать порядок действий при восстановлении можно на копиях системы, не относящейся к среде эксплуатации, но рассчитанной на идентичную инфраструктуру и с такой же конфигурацией, или на резервной копии (если система предусматривает аварийное переключение на резервную копию при отказе основной).

2.2.3 Создание экземпляров БД

Администраторы БД отвечают за создание экземпляров БД. Связанные с выполнением этой задачи работы включают следующее.

- **Установка и обновление программного обеспечения СУБД.** АБД устанавливают новые версии программного обеспечения СУБД и отвечают за установку обновлений, предоставляемых поставщиком СУБД, во всех средах — от среды разработки до эксплуатационной — в строгом соответствии с инструкциями поставщика. Состав работ и приоритеты подлежат согласованию между АБД и специалистами по информационной безопасности и ответственными менеджерами. Это критически важная работа с точки зрения защиты данных от возможных атак и обеспечения постоянной целостности данных как в централизованных, так и в распределенных конфигурациях баз данных.
- **Поддержка экземпляров, установленных в различных средах, в том числе с различными версиями СУБД.** АБД могут устанавливать и сопровождать экземпляры БД в различных средах, созданные в СУБД различных версий, включая «песочницу», среды разработки, тестирования, пользовательского приемочного тестирования, проверки качества, тестирования перед вводом в эксплуатацию, проверки после исправлений, восстановления после аварий

и собственно эксплуатационную. При этом АБД отвечают за управление миграциями на новые версии СУБД и обеспечение согласованности всех данных и изменений в них в различных версиях СУБД с точки зрения внешних систем и приложений.

- **Установка и сопровождение сопутствующего программного обеспечения.** АБД могут привлекаться к установке программного обеспечения интеграции данных и инструментов для администрирования от сторонних поставщиков.

2.2.3.1 УПРАВЛЕНИЕ ФИЗИЧЕСКОЙ СРЕДОЙ ХРАНЕНИЯ ДАННЫХ

Управление физической средой хранения данных должно осуществляться в соответствии с традиционным процессом управления конфигурацией программного обеспечения (Software Configuration Management, SCM) или методами библиотеки инфраструктуры информационных технологий. В любом случае должны документироваться любые изменения конфигурации баз данных, структур, ограничений, прав, пороговых значений и т. д. Администраторам БД нужно своевременно обновлять физическую модель данных, приводя ее в соответствие с изменениями в объектах систем хранения данных (что является частью процесса управления конфигурацией). При использовании методик гибкой разработки и экстремального программирования своевременность обновлений физической модели данных играет особо важную роль в предотвращении ошибок проектирования и разработки.

Администраторы БД должны организовать процесс SCM для отслеживания изменений и проверки установки всех последних обновлений программного обеспечения управления базами данных, развернутыми в средах разработки, тестирования и эксплуатации, даже если изменения носят чисто косметический характер или затрагивают лишь виртуализированный уровень представления данных.

Четыре обязательные процедуры SCM — идентификация конфигурации, контроль ее изменений, учет состояния конфигурации и ее аудит.

- **Идентификация конфигурации** проводится администратором БД совместно с распорядителями данных, архитекторами данных и разработчиками моделей данных с целью выявления всех атрибутов, определяющих каждый аспект конфигурации БД с точки зрения конечного пользователя. Все эти атрибуты фиксируются в документации в качестве базовой версии конфигурации. Для последующего изменения любого из задокументированных атрибутов необходимо строго придерживаться формальной процедуры контроля изменений конфигурации.
- **Контроль изменений конфигурации** — комплекс процедур и стадий согласования изменений атрибутов конфигурации и ее закрепления в качестве новой базовой версии.
- **Учет состояния конфигурации** позволяет АБД фиксировать и сообщать сведения о базовой конфигурации по всем ее компонентам по состоянию на любой запрошенный момент времени.
- **Аудит конфигурации** производится при сдаче системы в эксплуатацию и при каждом применении любых последующих изменений. Предусмотрено два типа аудита: физический аудит

конфигурации гарантирует, что компонент конфигурации установлен в соответствии с требованиями его детальной проектной документации; функциональный аудит конфигурации гарантирует, что компонент конфигурации демонстрирует требуемые функциональные характеристики.

В целях обеспечения целостности и прослеживаемости данных по всему их жизненному циклу АБД уведомляют специалистов по моделированию, разработчиков и ответственных за управление метаданными обо всех изменениях атрибутов физической базы данных.

АБД также должны проводить оценки в соответствии с метриками фактических объемов данных, прогнозируемых потребностей в дополнительной емкости устройств хранения, производительности в части обработки запросов, а также вести раздельную статистику по физическим объектам с целью выявления потребностей в репликации данных, переносе данных и создании контрольных точек восстановления данных. Для больших баз данных к функциям АБД добавляется разбивка базы на разделы с последующим мониторингом и сопровождением разделов в целях обеспечения оптимального распределения данных между ними.

2.2.3.2 УПРАВЛЕНИЕ МЕХАНИЗМАМИ КОНТРОЛЯ ДОСТУПА К БД

АБД отвечают за управление механизмами контроля доступа к данным. В целях защиты информационных активов и обеспечения целостности данных АБД обеспечивают выполнение следующих функций.

- **Контроль операционной среды.** АБД совместно с администраторами сетей хранения прорабатывают вопросы обеспечения полной контролируемости внешнего доступа к данным, включая: управление сетевыми ролями и допусками; круглосуточный мониторинг сетевого трафика и состояния сети; управление настройками межсетевого экрана; управление обновлениями безопасности; интеграцию с анализатором безопасности Microsoft Baseline Security Analyzer (MBSA).
- **Физическая защита данных.** Обеспечивается посредством мониторинга сетевого трафика с помощью простого протокола управления сетью (Simple Network Management Protocol, SNMP), ведения контрольных журналов изменений данных, применения мер по обеспечению отказоустойчивости систем и планового (по графику) создания резервных копий данных. АБД конфигурируют все вышеперечисленные протоколы и средства, а главное — производят регулярный мониторинг их состояния. Особого внимания требует контроль выполнения протоколов безопасности.
- **Мониторинг программно-аппаратного комплекса серверов баз данных** необходим для обеспечения бесперебойного доступа пользователей к данным.
- **Механизмы контроля.** АБД обеспечивают информационную безопасность за счет использования средств контроля доступа к данным, регулярного аудита баз данных, выявления случаев несанкционированного доступа, взлома или вторжения и выявления уязвимостей в защите.

Основные понятия и концепции, а также работы, проводимые в области обеспечения информационной безопасности, обсуждаются в главе 7.

2.2.3.3 СОЗДАНИЕ И КОНФИГУРИРОВАНИЕ ВЛОЖЕННЫХ СТРУКТУР ХРАНЕНИЯ ДАННЫХ

Все данные должны храниться на физических накопителях и организовываться таким образом, чтобы их конфигурация обеспечивала максимальную простоту и высокую скорость загрузки, поиска и извлечения. Структура хранения данных состоит из объектов среды хранения, которые, в свою очередь, могут включать в себя другие объекты и играют роль своего рода контейнеров (containers). Управление объектами на каждом уровне вложенности осуществляется с помощью соответствующих этому уровню методов. Например, в реляционных базах данных создаются схемы, которые содержат таблицы; а в нереляционных — структуры, содержащие файлы.

2.2.3.4 РЕАЛИЗАЦИЯ ФИЗИЧЕСКИХ МОДЕЛЕЙ ДАННЫХ

АБД обычно отвечают за создание и сопровождение всей совокупности физической среды хранения данных, включая оперативное управление ею, а структура хранилищ зависит от физических моделей данных. Физическая модель данных включает объекты хранения, индексные объекты и объекты с инкапсулированными кодами, необходимые для решения всевозможных технических задач, включая проверку соблюдения бизнес-правил и контроль качества данных, обеспечение связности объектов БД и оптимизацию производительности СУБД.

В зависимости от практики, установленной в организации, специалисты по моделированию данных могут предоставлять АБД разработанную ими логическую модель данных, а АБД — заниматься ее физической реализацией. В качестве альтернативного варианта АБД могут сами проектировать каркас физической модели, а затем по согласованию с руководством обогащать и дополнять ее на стадии реализации специфическими деталями, включая индексы, разделы и кластеры, оценки потребностей в емкостях и схемы конфигурации хранилищ.

Для случаев использования баз данных сторонних разработчиков (в которых реализованы собственные модели данных), большинство инструментов моделирования предлагают средства реверс-инжиниринга. Они позволяют получить архитектуру и модели данных, используемые в коммерческих программных продуктах и системах планирования ресурсов предприятия (ERP), — главное, чтобы программное средство моделирования могло прочитать каталог коммерческой системы хранения. На основе полученной информации можно разрабатывать собственные физические модели данных. При этом АБД или специалисты по моделированию данных должны следить за актуальностью используемых физических моделей и по мере необходимости обновлять ограничения для приложений или связи между таблицами, поскольку далеко не все они бывают учтены в каталогах баз данных, поставляемых в пакетах с коммерческими СУБД. Особенно последнее соображение касается приложений, созданных в те времена, когда желательным считалось максимальное абстрагирование баз данных от контекста их прикладного использования.

В тех же случаях, когда АБД отвечают за эксплуатацию баз данных, предлагаемых в качестве услуги, обеспечение актуальности физических моделей становится их прямой обязанностью.

2.2.3.5 ЗАГРУЗКА ДАННЫХ

Построив физическую модель, мы получаем пустую базу данных. Заполнить ее актуальными данными — задача АБД. Если данные, которые нужно загрузить на сервер, экспортированы или выгружены откуда-то еще с помощью специальных утилит, то, скорее всего, они получены в формате, не требующем каких-либо дальнейших действий по обеспечению их совместимости с новой базой данных. Большинство СУБД поддерживают пакетное считывание больших массивов данных, и в этом случае главное — удостовериться в совместимости формата исходной базы данных с объектами целевой БД или простой функции преобразования исходных данных в целевой формат.

Многие организации практикуют также скачивание бесплатных или приобретение коммерческих баз данных (например, списков потенциальных клиентов у информационных брокеров, адресных справочников или справочников почтовых индексов, каталогов продукции поставщиков и т. п.). Какие-то данные подобного рода предоставляются по лицензионным соглашениям, какие-то находятся в открытом доступе; форматы исходных данных могут варьироваться (CD, DVD, EDI, XML, RSS-каналы, тестовые файлы и т. п.); доступ к обновленным данным может предоставляться по запросу, подписке или в режиме регулярных обновлений. Иногда требуется заключение соглашений, подтверждающих законность приобретения данных. АБД должны это понимать и следить за тем, чтобы загрузка данных не могла быть квалифицирована как противоправное действие.

АБД может быть поручено самостоятельно производить загрузку данных или создать карту источников и схему загрузки. В любом случае загрузки данных в ручном режиме должны ограничиваться ситуациями первоначальной установки (инсталляции) и иными экстраординарными ситуациями, а во всех остальных случаях следует обеспечивать плановую дозагрузку и обновление данных по графику, заданному при определении настроек системы.

Управляемый подход к получению данных позволяет обеспечивать централизованное управление службами подписки на данные (data subscription services), используемые аналитиками данных. Аналитики данных должны документировать информацию о необходимых им для работы источниках внешних данных в логических моделях и словарях данных. Разработчики на ее основе смогут создавать сценарии или программы считывания и загрузки данных в БД, а администраторы баз данных будут отвечать за реализацию всех процессов, необходимых для загрузки данных, и/или обеспечение доступа к ним приложений.

2.2.3.6 УПРАВЛЕНИЕ РЕПЛИКАЦИЕЙ ДАННЫХ

АБД могут влиять на принятие руководящих решений, касающихся правил тиражирования и процедур размножения копий баз данных, как минимум посредством выработки и представления рекомендаций по следующим вопросам:

- выбор активной или пассивной схемы репликации;
- определение схемы управления параллельными задачами в распределенных БД;
- выбор метода выявления потребности в обновлении данных (по отметкам времени или номерам версий) в рамках процесса контроля изменений данных (Change Data Control process).

В случае небольших систем или объектов оптимальным методом согласования и обновления данных можно считать полную перезапись. Если же речь идет о крупных, а тем более особо крупных объектах, где данные меняются настолько избирательно, что основная их масса вовсе не изменяется, логичнее и целесообразнее с точки зрения обеспечения производительности отслеживать лишь накапливающиеся изменения по каждому объекту, а их запись откладывать до планового обновления основной БД. Однако если изменения затрагивают значительную часть крупных объектов БД, то лучше запланировать полное обновление БД: оно пройдет быстрее, чем покомпонентное.

2.2.4 Управление производительностью базы данных

Производительность базы данных определяется двумя взаимосвязанными факторами — доступностью и скоростью обработки. Высокая производительность обеспечивается путем обеспечения гарантированного наличия свободного места, оптимизации маршрутизации и порядка обработки запросов, а также учетом иных факторов влияния на эффективность обработки данных. При отсутствии доступа о производительности речи не идет: ее можно считать нулевой. Администраторы БД и сетей хранения обеспечивают их доступность и оптимальную производительность следующими средствами.

- Настройка параметров ОС и приложений.
- Управление подключениями базы данных: администраторы БД и сетей хранения осуществляют техническое руководство и поддержку доступа сотрудников блока ИТ и бизнес-пользователей к базам данных по тем каналам и с использованием тех протоколов, которые установлены стандартами организации.
- Настройка операционных систем, сетевого ПО и промежуточного ПО обработки транзакций (совместно с системными программистами и сетевыми администраторами).
- Выделение достаточных объемов памяти устройств хранения и организация работы баз данных с устройствами хранения и поддерживающим эти устройства программным обеспечением. Поддерживающее ПО позволяет оптимизировать распределение данных по устройствам хранения различных технологических типов эффективным с точки зрения затрат образом: относительно старые и редко запрашиваемые данные переносятся на дешевые устройства, высвобождая место на дорогих устройствах, обеспечивающих высокую скорость доступа, для актуальных и часто запрашиваемых данных (эффективность распределения данных по устройствам хранения различных типов отслеживается АБД совместно с администраторами устройств).
- Изучение динамики роста объемов данных в целях своевременного приобретения дополнительных накопителей и планирования деятельности в рамках жизненного цикла данных, включая хранение, архивирование, резервное копирование, стирание и аварийное восстановление.
- Определение (совместно с системными администраторами) показателей уровня рабочих нагрузок и производительности развернутых баз данных в целях управления соглашением

об уровне обслуживания, расчета процента отклоненных транзакций, оценки требуемой производительности серверов и темпов повторения жизненного цикла данных в пределах принятого в организации горизонта планирования.

2.2.4.1 ОПРЕДЕЛЕНИЕ УРОВНЕЙ ОБСЛУЖИВАНИЯ НА ОСНОВЕ ЭКСПЛУАТАЦИОННЫХ ХАРАКТЕРИСТИК

Производительность системы, ожидаемые параметры доступности, сроки восстановления и обязательства провайдера по устранению проблем обычно регулируются соглашением об уровне обслуживания (SLA) между организацией-провайдером ИТ-услуг по управлению данными и владельцами данных (см. рис. 61).

Рисунок 61. Соглашения об уровне обслуживания в контексте эксплуатационных характеристик базы данных

Обычно SLA устанавливает временны́е рамки, в пределах которых провайдером гарантируется доступность базы данных для пользователей. Часто в SLA оговаривается также предельно допустимое время обработки некоторых стандартных транзакций, реализуемых приложениями (комбинация сложных запросов и обновлений). В случае отсутствия доступа к БД дольше предельно допустимого времени или превышения сроков обработки запросов, предусмотренных SLA, владельцы данных обращаются к АБД с просьбой разобраться с причинами проблем и устранить их.

2.2.4.2 УПРАВЛЕНИЕ ДОСТУПНОСТЬЮ БД

Доступность (availability) — это процент времени, когда база данных или система могут использоваться для продуктивной работы. По мере наращивания объемов данных, используемых организациями в повседневной работе, и расширения спектра вариантов их практического применения, растут как требования к провайдерам услуг, так и риски убытков или издержек, обусловленных недоступностью данных. В целях удовлетворения растущего спроса провайдеры услуг

вынуждены неуклонно сокращать технологические перерывы на проведение регламентных работ по обслуживанию баз данных. На доступность БД оказывают влияние четыре взаимосвязанных фактора, а именно:

- **управляемость** — способность создавать и поддерживать операционную среду;
- **восстанавливаемость** — способность оперативно возобновлять обслуживание после сбоя или прерывания связи, а также исправлять ошибки, обусловленные непредвиденными обстоятельствами или отказом каких-либо компонентов систем;
- **надежность** — способность предоставлять услуги на оговоренных уровнях обслуживания в оговоренные периоды времени;
- **обслуживаемость** — способность выявлять имеющиеся проблемы, диагностировать их причины и решать проблемы и/или устранять их причины.

Имеется множество причин, по которым базы данных становятся недоступными, в том числе:

- плановые отключения;
- перерывы на профилактику и техническое обслуживание;
- ограничение доступа на период проведения обновлений;
- внеплановые или экстренные отключения;
- выход из строя серверного оборудования;
- отказы жестких дисков;
- сбои в работе операционной системы (ОС);
- сбои в работе ПО СУБД;
- потеря связи с ЦОД;
- отказы сетевого оборудования;
- проблемы с приложениями;
- проблемы с системами безопасности и авторизации;
- резкое снижение производительности;
- проблемы с восстановлением;
- проблемы с данными;
- повреждение или нарушение целостности данных (из-за программных или аппаратных сбоев, некорректной структуры данных или ошибок пользователей);
- потеря объектов БД;
- потеря данных;
- ошибки репликации данных;
- человеческий фактор.

АБД обязаны делать всё возможное для сохранения работоспособности баз данных в режиме онлайнового доступа, включая следующее:

- регулярное резервное копирование БД с использованием специальных утилит;
- своевременный запуск утилит реорганизации БД;
- регулярное обновление текущей статистики БД с помощью специальных утилит;
- регулярная проверка целостности данных с помощью специальных утилит;
- автоматизация запуска на исполнение всех вышеперечисленных утилит;
- использование кластеризации и разбиения таблиц данных;
- регулярная репликация данных на всех зеркалах с целью обеспечения высокого уровня их доступности.

2.2.4.3 УПРАВЛЕНИЕ ФУНКЦИОНИРОВАНИЕМ БД

АБД также устанавливает и отслеживает порядок функционирования БД, ведения журналов изменения данных и синхронизации дублируемых сред. Объемы и количество ведущихся на различных участках журналов бывают таковы, что в некоторых случаях для их организации приходится использовать отдельную (например, файловую) базу данных. Управлять следует также и приложениями, имеющими доступ к журналам и использующими их в собственных целях. Нужно следить, чтобы такие приложения использовали соответствующие журналы и реализовывали требуемый уровень протоколирования. Чем больше деталей протоколируется, тем больше пространства на устройствах хранения и процессорных ресурсов отбирает ведение журналов, что чревато снижением производительности БД.

2.2.4.4 ОБЕСПЕЧЕНИЕ ПРОИЗВОДИТЕЛЬНОСТИ, ПРЕДУСМОТРЕННОЙ СОГЛАШЕНИЕМ ОБ УРОВНЕ ОБСЛУЖИВАНИЯ

АБД отвечают за оптимизацию производительности путем применения как проактивного, так и реактивного подходов, осуществляя мониторинг производительности и немедленно реагируя на любые проблемы с целью их быстрого и эффективного устранения. В большинстве коммерческих СУБД реализованы функции мониторинга производительности, которые позволяют АБД создавать аналитические отчеты. Большинство серверных ОС также предоставляют аналогичную функциональность. АБД должны регулярно формировать отчеты о производительности как СУБД, так и серверов, в том числе и в периоды пиковых нагрузок. Новые отчеты должны сравниваться с предыдущими в целях выявления негативных тенденций и сохраняться в качестве вспомогательного средства при анализе возникающих проблем.

2.2.4.4.1 Сравнение показателей производительности обработки транзакций и пакетной обработки данных

Перемещение данных может происходить как в режиме реального времени в процессе обработки онлайновых транзакций, так и при выполнении программ пакетной обработки, которые могут передавать данные из системы в систему или проводить с ними внутренние операции. Пакетные задачи должны выполняться строго в отведенные для этого интервалы времени, выделенные в графике работ по обслуживанию баз данных. АБД совместно со специалистами по интеграции данных отвечают за мониторинг производительности пакетной обработки данных, выявление

ошибок и случаев слишком долгого выполнения пакетных задач с последующим определением причин ошибок/задержек и их устранением.

2.2.4.4.2 Разрешение проблемных вопросов

При появлении проблем с производительностью администраторы баз данных, сетей хранения и серверов должны использовать средства мониторинга и администрирования СУБД для выявления источника проблемы. Распространенные причины низкой производительности СУБД включают следующее.

- **Нехватка оперативной памяти или конфликты при одновременном обращении к одним и тем же ее участкам.** Проблемы решаются с помощью буферизации или кэширования данных.
- **Блокировка доступа к ресурсам БД.** Случается, что какой-либо процесс захватывает ресурсы БД (например, таблицы или страницы с данными) и блокирует доступ к ним других процессов, нуждающихся в тех же данных. Если проблема повторяется раз за разом, АБД лучше запретить вызывающему блокировку процессу доступ к запираемому им ресурсу. В некоторых случаях два процесса, конкурируя за нужные обоим ресурсы, входят в «тупик» (deadlock) и блокируют друг друга. В большинстве СУБД на такие случаи предусмотрено автоматическое снятие одной из конфликтующих задач через установленное время после выявления конфликта. Часто такого рода проблемы становятся следствием некачественного программного кода либо в СУБД, либо в приложении.
- **Неверно указываемые значения статистических характеристик БД.** В большинстве реляционных СУБД имеется встроенный оптимизатор запросов, который полагается на сохраненную статистику данных и индексов при решении о выборе наиболее эффективной процедуры обработки каждого запроса. Эти статистические показатели нуждаются в частом обновлении, особенно в активно используемых базах данных. Если этого не делать, результатом станет падение скорости обработки запросов.
- **Некачественное кодирование запросов SQL.** Пожалуй, самой распространенной причиной низкой производительности СУБД является неудачное кодирование SQL-запросов. Кодировщики должны понимать, как именно запросы оптимизируются и обрабатываются СУБД, и составлять их таким образом, чтобы преимущества оптимизатора обработки были использованы максимально. Некоторые системы позволяют инкапсулировать семантически сложные SQL-запросы в хранимые процедуры, которые можно заранее скомпилировать и оптимизировать, а не внедрять их в коды приложений или файлы скриптов.
- **Неэффективные сложные объединения таблиц.** Используйте представления (views) для предварительного сложного объединения (join) таблиц. Кроме того, избегайте использования сложных SQL-конструкций (в частности, задающих объединение таблиц) в функциях базы данных; в отличие от хранимых процедур функции оптимизатором запросов не обрабатываются.
- **Неправильное использование индексов.** Сложные запросы и запросы к большим таблицам должны обрабатываться с помощью индексов. Если нужно, создайте индексы, необходимые

для поддержки таких запросов. С осторожностью подходите к созданию большого количества индексов для часто и интенсивно обновляемых таблиц, поскольку это замедлит обработку обновлений.

- **Работа приложений.** В идеале приложения должны запускаться на отдельном от СУБД сервере, чтобы исключить конкуренцию за системные ресурсы. Сконфигурируйте и настройте серверы БД для обеспечения максимальной производительности. Новые СУБД позволяют инкапсулировать объекты-приложения (например, классы Java и .NET) в объекты БД с последующим их выполнением в среде СУБД. Однако пользоваться этой функциональностью нужно осмотрительно. В некоторых случаях она бывает очень полезна, но выполнение кодов приложений на сервере БД может негативно сказаться на интероперабельности, архитектуре приложений и скорости обработки данных.
- **Перегрузка серверов.** Для СУБД, поддерживающих работу многих баз данных и приложений, может наступить момент, когда после добавления очередной базы данных производительность ранее созданных баз заметно падает. В таком случае придется создавать новый сервер БД. Кроме того, рекомендуется перемещать на отдельный сервер слишком увеличившиеся в размерах или чаще и интенсивнее, чем прежде, использующиеся базы данных. В некоторых случаях проблемы с большими БД можно решить путем архивирования редко используемых данных и перемещения их в другое хранилище или же путем удаления устаревших данных.
- **Нестабильная производительность базы данных.** В отдельных случаях массовые добавления и удаления табличных записей за короткий промежуток времени приводят к неверной статистике распределения данных и, как следствие, дестабилизации производительности БД. Поэтому перед массовым обновлением данных временно отключайте функции ведения статистики у таблиц, которые затрагиваются при обновлении, поскольку некорректная статистика негативно влияет на работу оптимизатора запросов.
- **Неуправляемые запросы.** Пользователи иногда непреднамеренно формируют настолько сложные запросы, что они отбирают большую часть системных ресурсов совместного доступа. Используйте средства управления запросами или процессами для снятия или приостановки задач по обработке подобных запросов, после чего можно заняться их оценкой и оптимизацией.

Установив причину проблемы, АБД принимает все необходимые меры по ее решению, включая проработку (совместно с разработчиками приложений) требований по совершенствованию приложений и оптимизации кодов работы с БД, архивированию или удалению устаревших, избыточных и более не требующихся приложениям для активного использования данных. В исключительных случаях, когда речь идет о базах данных OLTP, АБД может совместно со специалистом по моделированию данных проработать возможности реструктуризации неудачно спроектированной и негативно сказывающейся на работе БД части модели данных. Используйте этот вариант лишь в крайнем случае, перепробовав предварительно все альтернативные меры (в частности, создание представлений и индексов, переписывание SQL-кодов и т. п.) и тщательно взвесив все

возможные последствия изменения модели, такие как нарушение целостности данных или излишнее усложнение структуры SQL-запросов, обращенных к денормализованным таблицам.

Однако в случае баз данных для отчетности и аналитических БД, доступных только для чтения, денормализация ради ускорения и упрощения доступа — скорее норма, чем исключение, поскольку никакого риска для данных не представляет.

2.2.4.5 ПОДДЕРЖКА АЛЬТЕРНАТИВНЫХ СРЕД

Базы данных находятся в состоянии непрерывного развития и видоизменения. Меняются бизнес-правила и бизнес-процессы, меняются и технологии. Среды разработки и тестирования позволяют апробировать изменения до их внедрения в среду эксплуатации. АБД могут создавать полные или частичные копии структур БД в других средах с целью разработки и испытания системных изменений. Имеется несколько основных типов альтернативных сред.

- **Среды разработки** используются для реализации и апробирования изменений, которые предполагается реализовать в эксплуатационной среде. Среда разработки должна быть максимально приближенной к реальной эксплуатационной среде, пусть и с пропорционально урезанными ресурсами.
- **Среды тестирования** могут быть предназначены для проверки качества, интеграционного тестирования, пользовательского приемочного тестирования, тестирования производительности и т. п. В идеале в тестовой среде должно использоваться то же программное и аппаратное обеспечение, что и в эксплуатационной. В частности, не допускается снижение уровня обеспечения ресурсами среды тестирования производительности по сравнению с эксплуатационной средой.
- **Экспериментальные среды («песочницы»)** используются для оперативной проверки гипотез, поиска и проработки новых применений данным. АБД обычно настраивают такие среды, открывают доступ к ним разработчикам и осуществляют мониторинг целевого использования этих сред. Кроме того, АБД должны следить за тем, чтобы «песочницы» были надежно изолированы от баз данных, используемых в эксплуатационной среде.
- **Альтернативные эксплуатационные среды** требуются для систем поддержки резервного копирования данных в режиме офлайн, переключения на запасные версии при авариях и обеспечения отказоустойчивости. Эти системы должны быть идентичны эксплуатационным, хотя вычислительная мощность системы резервного копирования (и восстановления) может быть уменьшена по сравнению с эксплуатационной, поскольку она используется в основном для ввода/вывода, а не обработки данных.

2.2.5 Управление наборами тестовых данных

Тестирование программного обеспечения — занятие трудоемкое и ресурсоемкое: на него приходится около половины от общей суммы затрат на разработку систем. Для эффективного тестирования необходимо иметь высококачественные наборы контрольных данных, и этими данными

также нужно управлять. Генерирование тестовых данных — критически важный компонент испытаний программного обеспечения.

Данные для проверки работоспособности системы подбираются особым образом. Тестирование может включать верификацию соответствия данных на выходе техническим заданиям при строго фиксированном наборе исходных данных или испытание способности программного обеспечения адекватно реагировать на необычные, экстремальные, недопустимые или неожиданные входные данные. Тестовые данные могут быть полностью выдуманными или сгенерированными с использованием бессмысленных значений, а могут являться образцом или выборкой вполне реалистичных данных. В частности, выборка данных для тестирования может быть получена из реальных производственных данных (по содержанию и/или структуре) или сгенерирована на их основе. Данные из производственной системы могут фильтроваться или компоноваться в различные профильные наборы тестовых данных, предназначенные для различных нужд. В тех случаях, когда в производственных данных содержатся защищенные или конфиденциальные данные, такие поля в выборке маскируются.

Наборы тестовых данных могут изготавливаться с использованием подходов, ориентированных на определенные цели или систематический подбор, предполагающих использование статистических данных или фильтров (обычно в случае функциональных испытаний), или иных, менее целенаправленных подходов (обычно в случае автоматизированного тестирования на случайных выборках большого объема). Тестовые данные могут подготавливаться вручную или с помощью вспомогательной программы/функции, помогающей тестировщику сформировать набор данных, или посредством копирования и целевой фильтрации данных из производственной среды. Наборы тестовых данных могут быть либо сохранены с целью повторного использования в ближайшей перспективе или поддержки регрессионного тестирования, либо удалены после однократного использования. В большинстве организаций процесс освобождения места по завершении проекта не включает этапа удаления тестовых данных, поэтому АБД должны следить за своевременной очисткой устройств хранения от устаревших тестовых наборов.

Не всегда представляется возможным наработать достаточный объем данных для проведения некоторых тестов, особенно тестов производительности. Допустимые ограничения по объему данных для тестирования в таких случаях определяются исходя из затрат времени и средств в соотнесении с качеством. Также могут сказываться и нормативно-правовые ограничения на возможность использования данных среды эксплуатации в среде тестирования (см. главу 7).

2.2.6 Управление миграцией данных

Под миграцией понимают перенос данных из одних устройств хранения, форматов или систем в другие с минимально возможными изменениями. Вопросы модификации данных при миграции обсуждаются в главе 8.

Миграция данных — ключевой вопрос при внедрении, обновлении версий или консолидации систем. Обычно она осуществляется программными средствами в автоматизированном режиме на основе правил. Однако людям всё равно нужно контролировать корректность исполнения

программ и соблюдения правил. Миграция данных может потребоваться по самым разнообразным причинам, включая замену или модернизацию серверного оборудования, консолидацию интернет-ресурсов, регламентные работы на серверах или смену расположения ЦОД. Большинство реализаций процесса миграции данных позволяют производить ее, не прерывая текущих операций, — например, параллельно с продолжением работы сервера хост-системы с вводом-выводом данных на логическое устройство (или LUN).

От степени фрагментации данных при переносе зависит, насколько быстро удастся обновить метаданные, сколько хранилищ дополнительной емкости потребуется при миграции и насколько быстро удастся обозначить предыдущее место хранения данных как свободное. Чем меньше степень фрагментации, тем меньше времени занимает перенос, меньше потребность в дополнительном пространстве на устройствах хранения и быстрее освобождаются старые устройства.

Многие повседневные задачи администратора устройства хранения могут решаться просто и в параллельном режиме с помощью миграции, включая:

- перенос данных с перегруженного устройства хранения в отдельную среду;
- перенос данных на устройство хранения, обеспечивающее более быстрый доступ к данным;
- реализацию политики управления данными на протяжении жизненного цикла;
- миграцию данных с устаревших или выводимых из эксплуатации устройств хранения в новые автономные или облачные хранилища.

Часто при миграции производится также автоматизированное и ручное исправление данных с целью повышения их качества, отбраковки избыточной или устаревшей информации и приведения данных в соответствие с требованиями новой системы. Фазы миграции данных (проектирование, извлечение, исправление, загрузка и верификация) для приложений средней и высокой степени сложности часто повторяются неоднократно, прежде чем новая система оказывается развернутой в полном объеме.

3. ИНСТРУМЕНТЫ

Помимо собственно СУБД, в распоряжении АБД имеется множество иных инструментов и средств управления базами данных. К ним относятся, например, программные средства моделирования и разработки, приложения, обеспечивающие интерфейс для формирования и отправки на обработку запросов, средства оценки и повышения качества данных путем их модификации, а также средства мониторинга рабочих нагрузок и производительности.

3.1 Инструменты моделирования данных

Программные средства моделирования данных позволяют автоматизировать многие задачи, стоящие перед разработчиками моделей. Некоторые инструменты подобного рода поддерживают

также генерирование языка определения данных (Data Definition Language, DDL), который будет использоваться в проектируемой БД. Большинство программных продуктов такого типа поддерживают также реверс-инжиниринг готовых баз данных до моделей. Продвинутые средства моделирования включают функционалы проверки соблюдения стандартов наименований, спелчекеры, хранилища метаданных, включая определения и генеалогию, а некоторые поддерживают даже и веб-публикацию (см. главу 5).

3.2 Инструменты мониторинга баз данных

Средства мониторинга баз данных позволяют автоматизировать отслеживание ключевых измеримых показателей — производительности, доступности, кэширования, статистики обращений и т. п. — и уведомлять администраторов БД и сетей хранения о проблемах. Многие программные средства подобного рода позволяют осуществлять параллельный мониторинг множества баз данных различных типов.

3.3 Инструменты управления конфигурацией баз данных

СУБД обычно включают всю необходимую инструментальную оснастку для управления конфигурационными настройками. Однако дополнительно АБД могут использовать и программные продукты сторонних производителей, поддерживающие согласованное управление множественными базами данных. Такие приложения включают функции конфигурирования, установки исправлений и обновлений, резервного копирования и восстановления, клонирования БД, управления тестированием и очистки памяти и данных.

3.4 Инструменты разработки приложений

Инструментальные средства поддержки разработки приложений обычно включают графический интерфейс для подключения к базам данных и исполнения команд в их среде. Инструменты программиста могут как входить в пакет программного обеспечения СУБД, так и предлагаться в качестве отдельных приложений сторонними разработчиками.

4. МЕТОДЫ

4.1 Тестирование в средах более низкого уровня

Возьмите за правило не устанавливать никаких обновлений или исправлений операционных систем, СУБД, структурной модели данных и приложений для БД в эксплуатационной среде, не протестировав их предварительно в самой низкоуровневой среде (обычно это среда разработки). Убедившись в корректной работе обновленной версии на самом низком уровне, переходите к установке обновления на следующем уровне (например, в среде тестирования) — и так до тех пор, пока не дойдет очередь до установки обновления в среде эксплуатации, где всякие исправления и обновления устанавливаются в последнюю очередь. Помимо проверки корректности работы

обновленной версии такой подход позволяет полностью разобраться и освоиться с процедурой обновления к тому моменту, когда дело дойдет до эксплуатационной версии, что поможет избежать сюрпризов и сведет к минимуму риск сбоя в работе систем, находящихся в эксплуатации.

4.2 Стандарты именования для физической модели данных

Последовательность и согласованность наименований способствует взаимопониманию. Архитекторы данных, разработчики и администраторы БД могут согласовывать стандартные наименования либо через определения метаданных, либо через выработку правил обмена документами между организациями.

Стандарт ISO/IEC 11179[1] призван обеспечивать упорядоченность семантики данных, представления данных и регистрации описаний данных. Только с помощью этих описаний и можно точно понять доподлинное значение и смысловое содержание тех или иных данных. Важнейшей с точки имен, допустимых в физических реализациях БД, является «Часть 5. Принципы наименования и идентификация», которая описывает порядок выработки соглашений о наименованиях элементов данных и компонентов систем.

4.3 Использование сценариев для внесения любых изменений

Всякое прямое внесение изменений в данные крайне рискованно с точки зрения целостности БД. Однако бывают ситуации, когда без этого никак не обойтись (например, при перегруппировке бухгалтерских счетов в новом финансовом году; реорганизации, переименовании или слиянии компаний; в экстренных случаях, когда поступает «разовый» запрос на внесение массированных идентичных изменений в однотипные данные и отсутствуют иные средства их внесения, кроме как пакетной или контекстной заменой). В таких случаях полезно бывает записать планируемые изменения в файл сценария и тщательно протестировать процедуру и результаты замены в предэксплуатационных средах, прежде чем вносить эти изменения в среду эксплуатации.

5. РЕКОМЕНДАЦИИ ПО ВНЕДРЕНИЮ

5.1 Оценка готовности / Оценка рисков

Оценка готовности к изменениям и рисков, сопряженных с их внедрением, применительно к базам данных сводится ко всестороннему анализу двух важнейших факторов — риска потери данных и рисков, связанных с вопросами технологической готовности.

- **Риск потери данных.** Данные могут быть утрачены как из-за технических или процедурных ошибок, так и вследствие чьих-либо злонамеренных действий. Организациям нужно иметь

[1] ГОСТ Р ИСО/МЭК 11179-5-2012 Информационная технология (ИТ). Регистры метаданных (РМД).

проработанные стратегии защиты от обеих угроз. Общие требования по обеспечению информационной безопасности и защите данных обычно сформулированы в соглашении об уровне обслуживания. Однако общие требования SLA должны быть подкреплены тщательно регламентированными процедурными правилами. Также на регулярной основе должна производиться текущая оценка адекватности технических средств защиты данных от злонамеренного уничтожения или хищения, поскольку киберугрозы в наши дни множатся как никогда раньше. Оценку и планирование мер по минимизации рисков рекомендуется включать в плановые проверки выполнения SLA и контрольные проверки данных.

- **Технологическая готовность.** Новейшие технологии (NoSQL, большие данные, хранилища триплетов, федеративные СУБД и т. п.) требуют от ИТ-специалистов особых знаний, навыков и опыта. У многих организаций просто не хватит возможностей, чтобы собрать у себя всех узкопрофильных экспертов, которые требуются для грамотной реализации новейших решений подобного рода. Также и АБД, и инженерам-системотехникам, и программистам, и бизнес-пользователям нужно тщательно подготовиться к полноценному использованию этих новинок в бизнес-аналитике и других полезных для организации областях.

5.2 Организационные и культурные изменения

Администраторы БД зачастую не умеют эффективно доносить до организации понимание ценности их работы. Чтобы изменить эту ситуацию, АБД нужно для начала самим научиться понимать обоснованные тревоги владельцев и потребителей данных, сбалансированно учитывать их текущие и долгосрочные нужды, разъяснять сотрудникам организации важность высококачественного управления данными и оптимизировать практику обработки данных, с тем чтобы гарантировать их максимальную полезность для организации и безвредность для потребителей. Рассматривая работу с данными как соблюдение абстрактного набора принципов и практических рекомендаций, то есть без учета человеческого фактора, АБД рискуют противопоставить себя всем остальным, привить в организации менталитет «они и мы» и прослыть догматиками, демагогами и обструкционистами.

Взаимное непонимание может возникнуть по многим причинам, но прежде всего — из-за несопоставимости систем координат. В организациях обычно рассматривают информационные технологии с точки зрения конкретных прикладных компьютерных программ, а вовсе не данных, на сами же данные смотрят как на придаток к приложениям. Долгосрочная ценность защищенных, многократно используемых, высококачественных данных как важного корпоративного ресурса обычно ускользает от всеобщего понимания, тем более что признать за данными статус ценного актива не так-то просто в силу их эфемерности.

Разработчики приложений часто видят в администраторах БД досадную помеху скорейшему завершению проектов, которая привносит дополнительные расходы без всякой видимой пользы. АБД медленно принимают изменения в технологиях (XML, объектно- и сервис-ориентированные архитектуры и т. п.) и новые методы разработки приложений (например, методологии Agile, Extreme Programming (XP) или Scrum). При этом разработчики, как правило, не осознают,

насколько большую помощь может оказать качественное управление данными в достижении ими долгосрочных целей повторного использования своих решений и создания по-настоящему сервис-ориентированной архитектуры приложений.

Администраторы БД совместно с другими специалистами по управлению данными могут и должны способствовать преодолению этих организационно-культурных барьеров. Они вполне способны и сами посодействовать налаживанию конструктивного сотрудничества со всеми, кто заинтересован в удовлетворении информационных потребностей организации, если будут придерживаться следующих руководящих принципов: выявление и использование возможностей для автоматизации; проектирование и разработка с учетом многократного использования результатов; применение лучших практик; согласование стандартов баз данных с требованиями по их поддержке; определение роли АБД в проектных работах. Кроме того, им следует помнить и соблюдать описанные ниже правила.

- **Упреждающее обсуждение возможных проблем.** Администраторам баз данных необходимо наладить тесное взаимодействие с разработчиками еще на стадии проектирования приложений и продолжать его вплоть до выявления и оперативного устранения всех проблем на стадии реализации и после внедрения. Они должны проводить анализ программных кодов, реализующих доступ к данным, хранимых процедур, представлений и функций работы с базой данных, созданных командами разработчиков, а также оказывать помощь в обнаружении возможных проблем, связанных с проектными решениями в части БД.
- **Общение с другими людьми на уровне их знаний и на понятном им языке.** С менеджерами лучше обсуждать их бизнес-потребности, ожидаемые результаты и экономический эффект, а с разработчиками приложений — объектно-ориентированный подход, степень жесткости связей, предельно допустимую сложность запросов к базе данных и т. п.
- **Постоянная сфокусированность на потребностях бизнеса.** В равной мере нужна и АБД, и разработчикам приложений, поскольку всё, что они делают, должно наилучшим образом соответствовать бизнес-требованиям и получению максимальной ценностной отдачи от проекта.
- **Готовность помогать другим.** Получая раз за разом категорический отказ в выполнении их просьб или пожеланий, люди перестают на вас рассчитывать и начинают действовать по собственному усмотрению, а это чревато нарушением ими стандартов и поиском вредных обходных решений. Нужно понимать: если людям что-то действительно нужно, они найдут тот или иной способ это получить — и отказ им в помощи чреват неприятностями для обеих сторон.
- **Постоянное обучение и учет ошибок.** Учет ошибок помогает избегать их повторения при реализации будущих проектов. Если кто-нибудь что-то сделал неверно, найдите время и объясните, как это делается по правилам и по каким причинам такие правила установлены, — и в будущем те же люди подобных ошибок уже не допустят.

Подводя итоги, отметим, что нужно понимать и учитывать потребности всех заинтересованных сторон. Разрабатывайте ориентированные на требования бизнеса четкие и ясные стандарты, с тем чтобы обеспечить наилучшее выполнение всеми и каждым своей части работы. Более того, разъясняйте и внедряйте эти стандарты таким образом, чтобы их соблюдение приносило бизнесу и всем заинтересованным лицам максимум пользы, — тем и заслужите всеобщее уважение.

6. РУКОВОДСТВО ХРАНЕНИЕМ И ОПЕРАЦИЯМИ С ДАННЫМИ

6.1 Метрики

Метрики в отношении хранения данных могут включать следующие параметры:

- количество баз данных по типам;
- сводную статистику транзакций;
- метрики для оценки емкости и вычислительной мощности, например:
  - объем и процент занятого данными пространства на устройствах хранения;
  - количество физических устройств хранения данных;
  - количество объектов данных (в терминах занятых и свободных блоков или страниц);
  - объем данных в очереди;
- интенсивность использования сервисов хранения данных;
- распределение запросов по сервисам хранения данных;
- рост производительности приложений, использующих сервис.

Метрики производительности могут включать:

- частоту и количество транзакций;
- скорость обработки запросов;
- производительность интерфейса прикладного программирования (API).

Операционные метрики могут включать:

- сводную статистику по затратам времени на поиск запрашиваемых данных;
- размер резервной копии;
- измеримые показатели качества данных;
- показатели доступности системы.

Метрики качества обслуживания могут включать:

- количество выявленных, решенных и переданных в вышестоящие инстанции проблем с разбивкой по типам;
- среднее время устранения одной проблемы.

Состав необходимых метрик подлежит согласованию между АБД, архитекторами данных и службами обеспечения качества данных.

6.2 Отслеживание и учет информационных активов

Один из аспектов руководства хранением данных — проверка соблюдения всех лицензионных соглашений и требований нормативно-правового регулирования. Проводите ежегодный учет установленного программного обеспечения и следите за тем, чтобы не истекли сроки действия лицензий или подписок на техническое сопровождение программных продуктов, а также договоров аренды серверных мощностей и других услуг. Помните, что нарушение лицензионных соглашений подвергает организацию серьезному финансовому и юридическому риску.

Результаты подобного аудита помогают точно определять совокупную стоимость владения (Total Cost-of-Ownership, TCO) каждой технологией и каждым ИТ-продуктом. Регулярно оценивайте все технологии и продукты на предмет целесообразности их сохранения и отказывайтесь от использования устаревших, не поддерживаемых, малополезных, не окупающих себя или просто непомерно дорогих продуктов и услуг.

6.3 Аудит и проверка корректности данных

Аудит данных заключается в оценке набора данных на предмет соответствия определенным критериям. Как правило, аудит проводится в тех случаях, когда возникают какие-либо сомнения относительно пригодности конкретного набора данных для эффективного использования, а также при необходимости установить, не нарушает ли хранение проверяемых данных условий каких-либо договорных, нормативно-правовых или методологических требований. Подход к проведению аудита может быть основан на применении чек-листов (ориентированных на специфику проекта или всесторонних), экспертных заключений, критериев контроля качества.

Проверка корректности данных заключается в оценке хранимых данных на предмет их соответствия установленным критериям качества и пригодности для использования. Процедуры проверки корректности данных зависят от набора критериев, используемых командой контроля качества данных (если таковая имеется) или определяемых исходя из требований потребителей данных. АБД должны в обязательном порядке оказывать всестороннюю поддержку как при проведении аудита, так и при проверках корректности данных, включая:

- участие в выработке и согласовании общего подхода к проверке;
- проведение предварительного анализа и отбраковки данных;
- разработку методов мониторинга данных;

- применение статистических методов оптимизации подлежащих анализу данных по географическим и биометрическим параметрам;
- обеспечение технической поддержки формирования выборки и анализа данных;
- рецензирование данных;
- оказание содействия в деятельности по раскрытию данных;
- оказание экспертной поддержки по вопросам администрирования баз данных.

7. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Amir, Obaid. *Storage Data Migration Guide*. 2012. Kindle.

Armistead, Leigh. *Information Operations Matters: Best Practices*. Potomac Books Inc., 2010. Print.

Axelos Global Best Practice (ITIL website), http://bit.ly/1H6SwxC

Bittman, Tom. «Virtualization with VMWare or HyperV: What you need to know». Gartner Webinar, 25 November, 2009, http://gtnr.it/2rRl2aP, Web.

Brewer, Eric. «Toward Robust Distributed Systems». PODC Keynote 2000, http://bit.ly/2sVsYYv Web.

Dunham, Jeff. *Database Performance Tuning Handbook*. McGraw-Hill, 1998. Print.

Dwivedi, Himanshu. *Securing Storage: A Practical Guide to SAN and NAS Security*. Addison-Wesley Professional, 2005. Print.

EMC Education Services, ed. *Information Storage and Management: Storing, Managing, and Protecting Digital Information in Classic, Virtualized, and Cloud Environments*. 2nd ed. Wiley, 2012. Print.

Finn, Aidan, et al. *Microsoft Private Cloud Computing*. Sybex, 2013. Print.

Finn, Aidan. *Mastering Hyper-V Deployment*. Sybex. 2010. Print.

Fitzsimmons, James A. and Mona J. Fitzsimmons. *Service Management: Operations, Strategy, Information Technology*. 6th ed. Irwin/McGraw-Hill, 2007. Print with CDROM.

Gallagher, Simon, et al. *VMware Private Cloud Computing with vCloud Director*. Sybex. 2013. Print.

Haerder, T. and A. Reuter. «Principles of transaction-oriented database recovery». *ACM Computing Surveys* 15 (4) (1983), https://web.stanford.edu/class/cs340v/papers/recovery.pdf Web.

Hitachi Data Systems Academy, *Storage Concepts: Storing and Managing Digital Data*. Volume 1. HDS Academy, Hitachi Data Systems, 2012. Print.

Hoffer, Jeffrey, Mary Prescott, and Fred McFadden. *Modern Database Management.* 7th Edition. Prentice Hall, 2004. Print.

Khalil, Mostafa. *Storage Implementation in vSphere 5.0.* VMware Press, 2012. Print.

Kotwal, Nitin. *Data Storage Backup and Replication: Effective Data Management to Ensure Optimum Performance and Business Continuity*. Nitin Kotwal, 2015. Amazon Digital Services LLC.

Kroenke, D. M. *Database Processing: Fundamentals, Design, and Implementation*. 10th Edition. Pearson Prentice Hall, 2005. Print.

Liebowitz, Matt et al. *VMware vSphere Performance: Designing CPU, Memory, Storage, and Networking for Performance-Intensive Workloads*. Sybex, 2014. Print.

Matthews, Jeanna N. et al. *Running Xen: A Hands-On Guide to the Art of Virtualization*. Prentice Hall, 2008. Print.

Mattison, Rob. *Understanding Database Management Systems*. 2nd Edition. McGraw-Hill, 1998. Print.

McNamara, Michael J. *Scale-Out Storage: The Next Frontier in Enterprise Data Management*. FriesenPress, 2014. Kindle.

Mullins, Craig S. *Database Administration: The Complete Guide to Practices and Procedures*. Addison-Wesley, 2002. Print.

Parsaye, Kamran and Mark Chignell. *Intelligent Database Tools and Applications: Hyperinformation Access, Data Quality, Visualization, Automatic Discovery*. John Wiley and Sons, 1993. Print.

Pascal, Fabian. *Practical Issues in Database Management: A Reference for The Thinking Practitioner*. Addison-Wesley, 2000. Print.

Paulsen, Karl. *Moving Media Storage Technologies: Applications and Workflows for Video and Media Server Platforms*. Focal Press, 2011. Print.

Piedad, Floyd, and Michael Hawkins. *High Availability: Design, Techniques and Processes*. Prentice Hall, 2001. Print.

Rob, Peter, and Carlos Coronel. *Database Systems: Design, Implementation, and Management*. 7th Edition. Course Technology, 2006. Print.

Sadalage, Pramod J., and Martin Fowler. *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Addison-Wesley, 2012. Print. Addison-Wesley Professional.

Santana, Gustavo A. *Data Center Virtualization Fundamentals: Understanding Techniques and Designs for Highly Efficient Data Centers with Cisco Nexus, UCS, MDS, and Beyond*. Cisco Press, 2013. Print. Fundamentals.

Schulz, Greg. *Cloud and Virtual Data Storage Networking*. Auerbach Publications, 2011. Print.

Simitci, Huseyin. *Storage Network Performance Analysis*. Wiley, 2003. Print.

Tran, Duc A. *Data Storage for Social Networks: A Socially Aware Approach*. 2013 ed. Springer, 2012. Print. Springer Briefs in Optimization.

Troppens, Ulf, et al. *Storage Networks Explained: Basics and Application of Fibre Channel SAN, NAS, iSCSI, InfiniBand and FCoE*. Wiley, 2009. Print.

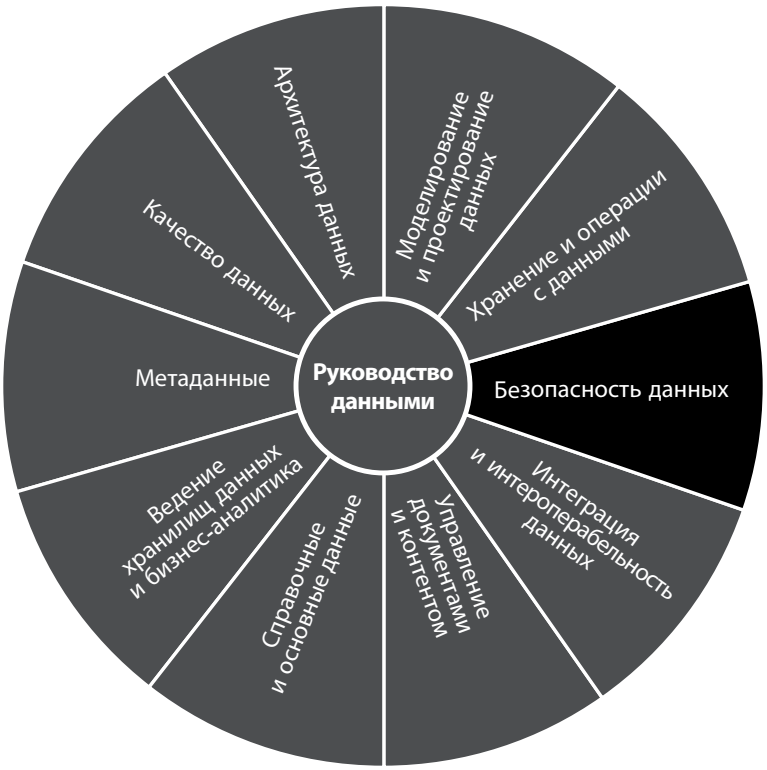
US Department of Defense. *Information Operations: Doctrine, Tactics, Techniques, and Procedures*. 2011. Kindle.

VMware. *VMware vCloud Architecture Toolkit (vCAT): Technical and Operational Guidance for Cloud Success*. VMware Press, 2013. Print.

Wicker, Stephen B. *Error Control Systems for Digital Communication and Storage*. US ed. Prentice-Hall, 1994. Print.

Zarra, Marcus S. *Core Data: Data Storage and Management for iOS, OS X, and iCloud*. 2nd ed. Pragmatic Bookshelf, 2013. Print. Pragmatic Programmers.

Безопасность данных



DAMA-DMBOK2 Data Management Framework

1. ВВЕДЕНИЕ

Деятельность в области безопасности данных включает в себя планирование, разработку и осуществление политик и процедур, обеспечивающих надлежащую аутентификацию, авторизацию и доступ пользователей, а также аудит данных и информационных ресурсов. Специфические требования по информационной безопасности (ИБ), в частности перечни данных, подлежащих защите, варьируются по отраслям и странам, однако общая цель остается неизменной повсеместно и заключается в защите информационных активов в соответствии с действующими законодательными нормами защиты информации о частной жизни, персональных и конфиденциальных данных, условиями действующих договоров и соглашений, а также потребностями бизнеса. Требования по обеспечению ИБ обусловлены следующими факторами.

- **Заинтересованные лица.** Организации должны выявлять и учитывать потребности в защите информации о частной жизни, персональных данных и конфиденциальных данных всех заинтересованных лиц, к которым могут относиться (в зависимости от типа и характера организации), например, клиенты, пациенты, студенты, граждане, поставщики, деловые партнеры и т. п. Все сотрудники организации несут ответственность за соблюдение требований, касающихся безопасности данных.
- **Нормативно-правовое регулирование.** Нормативно-правовое регулирование различных аспектов безопасности данных и интересов определенных заинтересованных сторон. Законы и правительственные постановления могут преследовать различные цели: одни ограничивают доступ к определенным данным, другие, напротив, призваны обеспечить открытость, прозрачность и подотчетность.
- **Охрана интеллектуальной собственности и коммерческой тайны.** В каждой организации имеются данные, которые можно расценивать в качестве предмета ее интеллектуальной собственности или коммерческой тайны. Такие данные нуждаются в защите. К примеру, клиентские базы данных помогают организации эффективно вести бизнес и получать преимущество перед конкурентами. В случае кражи, взлома системы хранения или уничтожения данных такое преимущество будет сразу же утеряно.

Рисунок 62. Источники требований по защите данных (Ray, 2012)

Рисунок 63.
Контекстная диаграмма: безопасность данных

- **Потребности в санкционированном доступе к данным.** Защита данных не должна ограничивать законного доступа к данным тех лиц, которые имеют на это право согласно действующему законодательству, равно как и санкционированного доступа к ним с целью использования, обслуживания, сопровождения, упорядочения и обработки в рамках бизнес-процессов.
- **Договорные обязательства.** Договорные обязательства и условия соглашений о неразглашении данных также сказываются на требованиях по обеспечению ИБ. Например, стандарт безопасности индустрии платежных карт (PCI-DSS) требует от платежных систем, банков-эмитентов и коммерческих предприятий строго определенных мер по защите данных (например, обязательного шифрования паролей).

Эффективные политики и процедуры в области безопасности данных гарантируют доступ к данным и возможность их целевого использования всем, кому это положено по праву или обязанности, и исключают возможность несанкционированного доступа или изменения данных (см. рис. 62). Обеспечение понимания и соблюдения потребностей и интересов всех сторон в отношении безопасности информации о частной жизни, персональных и конфиденциальных данных — важный аспект деятельности любой организации, поскольку ее отношения с клиентами, поставщиками и прочими заинтересованными сторонами строятся всецело на доверии, ключевой составляющей которого является ответственное обращение с данными.

1.1 Бизнес-драйверы

Основным драйвером деятельности в области обеспечения ИБ выступает стремление минимизировать риски и обеспечить устойчивый рост бизнеса. Эффективное обеспечение безопасности данных снижает риски и дает дополнительные конкурентные преимущества. Безопасность данных саму по себе можно рассматривать в качестве ценного актива организации.

Риски в области безопасности данных связаны с нормативно-правовыми требованиями, ответственностью руководства и/или владельцев перед компаниями, репутацией, исполнением юридических и моральных обязательств перед сотрудниками, партнерами и клиентами в части неразглашения их личных сведений, конфиденциальных данных и прочей нежелательной для раскрытия (чувствительной — sensitive) информации. С организаций могут взыскиваться крупные штрафы за несоблюдение установленных законом норм или неустойки за нарушение договорных обязательств. Утечки данных могут повлечь непоправимый репутационный ущерб и утрату доверия клиентов (см. главу 2).

Рост бизнеса включает достижение поставленных и формулировку новых целей. Проблемные вопросы безопасности, утечки данных, равно как и необоснованные ограничения доступа к ним сотрудников могут напрямую помешать успешному решению текущих задач.

Цели минимизации рисков и роста бизнеса могут быть согласованы и взаимно дополнять друг друга, если они объединены в комплексную стратегию управления информацией и обеспечения информационной безопасности.

1.1.1 Снижение рисков

По мере появления всё новых нормативно-правовых актов, регламентирующих всевозможные аспекты хранения и циркуляции данных, — а появляются они, как правило, в ответ на очередные эпизоды хищения или утечки данных, — возникают и новые требования по обеспечению нормативно-правового соответствия. Структурам организации, отвечающим за безопасность, всё чаще поручают управление не только выполнением соответствующих требований в отношении ИТ, но также политиками, практиками, классификацией данных и правилами авторизации доступа в масштабах организации.

Как и другие аспекты управления данными, обеспечение безопасности данных лучше всего рассматривать как корпоративную инициативу. Без должной координации каждое подразделение будет изыскивать собственные решения в области ИБ, что приведет к неоправданному росту затрат и даже возможному снижению общего уровня безопасности из-за несогласованности или несовместимости средств защиты данных различных подразделений. Неэффективная архитектура или рассогласованные процессы обеспечения безопасности данных могут дорого обойтись организации, поскольку чреваты утечками и падением производительности информационных систем. Снизить эти риски поможет полноценно финансируемая и системно-ориентированная операционная стратегия безопасности данных, действующая в масштабах организации.

Информационная безопасность начинается с классификации данных, которыми располагает организация, и выявления тех их категорий, которые нуждаются в защите. В общих чертах процесс включает следующие основные этапы.

- **Выявление и классификация информационных активов с чувствительными данными.** В зависимости от отрасли и специфики организации она может обладать различными по объемам и структуре массивами чувствительных данных (включая персональные, медицинские, финансовые и многие другие).
- **Определение мест расположения чувствительных данных в организации.** Требования по ИБ могут варьироваться в зависимости от фактического места хранения данных. Особую озабоченность вызывают хранилища с высокой концентрацией чувствительных данных, поскольку в таких случаях лишь одного взлома базы данных достаточно для причинения огромного ущерба.
- **Определение мер по защите каждого из выявленных активов с чувствительными данными.** Необходимые меры могут варьироваться в зависимости от содержания данных и технологий управления данными.
- **Определение характера взаимодействия чувствительных данных с бизнес-процессами.** Анализ бизнес-процессов необходим для определения условий доступа к данным.

Помимо классификации данных, необходимо провести оценку внешних угроз (хакерских атак, криминальных взломов и т. п.) и внутренних рисков (обусловленных ошибками и злонамеренными действиями сотрудников или недостатками процессов). Самой распространенной причиной утери или утечки данных являются некомпетентность персонала, непонимание сотрудниками

степени чувствительности или значимости данных или пренебрежение правилами безопасности[1]. Данные о продажах, оставшиеся на веб-сервере, который подвергся взлому; отправка списка сотрудников с личными данными на электронную почту подрядчика, у которого после этого похитили ноутбук; хранение данных, составляющих коммерческую тайну, в незашифрованном виде на переносном жестком диске, который был утерян, — все подобные случаи утечки данных происходят от отсутствия правил обеспечения информационной безопасности или из-за пренебрежения ими на фоне отсутствия надлежащего контроля.

Несоблюдение правил информационной безопасности в последние годы привело ко множеству скандалов и колоссальным убыткам вследствие репутационных потерь и утраты доверия потребителей к самым, казалось бы, солидным брендам. Помимо нарастания интенсивности, изощренности и целенаправленности внешних атак со стороны криминального хакерского сообщества, год от года стабильно растет и объем ущерба, обусловленного внешними и внутренними рисками намеренных или непреднамеренных нарушений (Kark, 2009).

В мире практически всецело электронной инфраструктуры бизнеса достоверность данных, содержащихся в информационных системах, становится существенным козырем в конкурентной борьбе.

1.1.2 Рост бизнеса

Электронные технологии в глобальных масштабах проникли во все офисы и производства, торговые площади и жилые дома. Десктопы и ноутбуки, смартфоны и планшеты, а также всевозможные другие устройства стали неотъемлемыми и крайне важными элементами в деятельности большинства коммерческих и правительственных структур. Взрывной рост электронной коммерции кардинально изменил подход организаций к предложению своих товаров и услуг потребителям. В частной жизни люди привыкли улаживать все дела в режиме онлайн: делать покупки, оплачивать коммунальные счета, записываться на прием к врачам, обращаться в госучреждения, управлять банковскими счетами. Доверие к электронной коммерции — залог прибыльности и роста бизнеса, тогда как качество продуктов и услуг напрямую зависит от информационной безопасности, ведь без надежной защиты персональных данных и онлайновых транзакций ни о каком доверии потребителей не может быть и речи.

1.1.3 Безопасность как актив

Особо следует отметить подход к управлению чувствительными данными с помощью метаданных. Классификацию в области безопасности и степень чувствительности к требованиям регулирующих органов можно фиксировать на уровне элементов и наборов данных. Существуют технологии

[1] По данным корпорации Cisco, «70% специалистов в области информационных технологий считают использование неразрешенных программ причиной не менее половины случаев утери данных их компанией; чаще всего такие встречались в США (74%), Бразилии (75%) и Индии (79%)», а из совместного отчета Ponemon Institute и Symantec следует, что «в 2012 году две трети утечек данных стали следствием человеческих ошибок и системных проблем». *Источники*: http://bit.ly/1dGChAz, http://symc.ly/1FzNo5l, http://bit.ly/2sQ68Ba, http://bit.ly/2tNEkKY

присвоения данным меток (метаданных), которые сопровождают данные в процессе их перемещения в информационных потоках организации. Создав основной репозиторий (master repository) характеристик данных, вы обеспечиваете ситуацию, при которой во всех частях организации точно известно, какой уровень защиты требуется для той или иной чувствительной информации.

При внедрении единого стандарта такой подход позволяет всем департаментам, бизнес-единицам и поставщикам использовать общий для всех набор метаданных. Стандартные метаданные, относящиеся к безопасности, позволяют не только оптимизировать защиту данных, но и снизить издержки за счет прямого указания на правила обращения с данными в ходе различных технологических и бизнес-процессов. Этот же подход к обеспечению безопасности позволяет предотвращать несанкционированный доступ и снижать риск злоупотребления информационными активами. Когда чувствительные данные корректно идентифицированы, это способствует также и укреплению доверия к организации со стороны клиентов и партнеров. Метаданные, относящиеся к информационной безопасности, сами по себе становятся стратегическим активом, повышающим качество транзакций, отчетности и бизнес-анализа при одновременном снижении затрат на защиту данных и рисков утери или хищения ценной информации.

1.2 Цели и принципы

1.2.1 Цели

Основные цели работ, проводимых в области безопасности данных, следующие.

- Обеспечение санкционированного доступа и исключение возможности несанкционированного доступа к информационным активам организации.
- Обеспечение соблюдения нормативно-правовых требований и политик в отношении защиты информации о частной жизни, персональных и конфиденциальных данных.
- Обеспечение соблюдения требований всех заинтересованных сторон в отношении защиты информации о частной жизни, персональных и конфиденциальных данных.

1.2.2 Принципы

Обеспечение безопасности данных организации должно быть основано на следующих принципах.

- **Сотрудничество.** Деятельность в области безопасности данных требует слаженных усилий администраторов по безопасности ИТ, распорядителей данных, ответственных за руководство данными, команд, проводящих внутренний и внешний аудит, а также юридической службы организации.
- **Корпоративный подход.** Стандарты и политики в области безопасности данных должны применяться согласованно в масштабах всей организации.
- **Проактивное управление.** Успех управления защитой данных зависит от динамичного и своевременного, а по возможности упреждающего выявления и устранения силами всех

заинтересованных сторон узких мест в деятельности организации или организационной культуре, которые мешают эффективному обеспечению ИБ, включая преодоление барьеров традиционного строгого разграничения ответственности между специалистами по ИБ, ИТ, управлению данными и бизнес-администрированию.

- **Четкое распределение ответственности.** Функциональные роли и обязанности должны определяться предельно четко, включая цепочки передачи данных «с рук на руки» под ответственность различных должностных лиц и подразделений.
- **Управление на основе метаданных (metadata-driven).** Классификация элементов данных с точки зрения безопасности — неотъемлемая часть определения данных.
- **Снижение риска за счет уменьшения объема распространяемых данных.** Минимизируйте объемы распространяемой чувствительной и/или конфиденциальной информации, особенно в средах, не относящихся к рабочей эксплуатации.

1.3 Основные понятия и концепции

В области ИБ используется специфическая профессиональная терминология. Знание ключевых терминов позволяет ответственным за руководство данными четко формулировать требования по безопасности данных.

1.3.1 Уязвимость

Уязвимостью в ИБ принято называть слабое место или дефект в системе, которое может создать условия для спешной атаки извне и раскрытия или потери информации. По существу, это «дыра» (hole) в защите организации. Некоторые уязвимости называют эксплойтами (exploits)[1].

Примеры уязвимостей: сетевые компьютеры с неустановленными обновлениями безопасности; веб-страницы, не защищенные надежными паролями; пользователи, не обученные игнорировать активные ссылки и вложения, содержащиеся в электронных письмах от неизвестных отправителей; корпоративное ПО, не защищенное от выполнения служебных команд, позволяющих злоумышленнику получать контроль над системой.

Во многих случаях предэксплуатационные среды отличаются большей уязвимостью, чем эксплуатационные. Следовательно, важнейшим аспектом ИБ является надежная изоляция данных в эксплуатационной среде, используемых в операционной деятельности организации, от всякой возможности доступа со стороны предэксплуатационных сред.

1.3.2 Угроза

Угроза (*threat*) — потенциальное атакующее воздействие, которое может быть предпринято против организации. Угрозы бывают внешние и внутренние. Злой умысел — не обязательный атрибут угрозы. Постоянный сотрудник вполне может по неопытности или неведению, сам того не

[1] Под термином «эксплойт» (от *англ.* exploit — эксплуатировать) обычно понимается компьютерная программа (а также фрагмент программного кода или последовательность команд), использующая уязвимости в программном обеспечении и применяемая для проведения атаки на вычислительную систему. — *Примеч. науч. ред.*

сознавая, выполнить операции, приносящие организации ущерб. Угрозы могут быть связаны с теми или иными уязвимостями, которым должны быть назначены приоритеты в отношении их устранения. Для каждой угрозы следует предусмотреть ту или иную возможность ее предотвращения или ослабления ущерба в случае ее реализации. Совокупность ресурсов, которые подвержены угрозе, называют *поверхностью атаки* (*attack surface*).

Примерами угроз являются: получение представителями организации электронных писем с вирусами во вложениях; запуск процессов, вызывающих перегрузку сетевых серверов и блокирующих тем самым бизнес-транзакции, — так называемые DoS-атаки (атаки, вызывающие отказ в обслуживании, — denial-of-service attacks); выявление и использование злоумышленниками известных уязвимостей.

1.3.3 Риск

Риском (*risk*) называют одновременно возможность ущерба и источник (предмет или объект) или причину (условие) возможного ущерба. Риск, связанный с любой угрозой, может быть оценен на основе учета следующих факторов.

- Вероятность реализации угрозы и возможная частота ее реализации.
- Характер и размер ущерба, включая репутационный, причиняемого каждым случаем реализации угрозы.
- Влияние понесенного ущерба на размер дохода и бизнес-операции.
- Затраты на устранение последствий понесенного ущерба.
- Затраты на предотвращение угрозы, включая исправление уязвимостей.
- Цели или намерения потенциального злоумышленника.

Риски можно классифицировать по их приоритетности с точки зрения потенциальной тяжести вероятного ущерба или вероятности реализации угроз, которая тем выше, чем легче отыскиваются уязвимости. Часто приоритеты расставляются с учетом обоих этих факторов. Оценка приоритетности рисков должна представлять собой формализованный процесс, в который должны быть вовлечены все заинтересованные стороны.

1.3.4 Классификация рисков

Классификация рисков описывает степень чувствительности данных и вероятность того, что они могут заинтересовать злоумышленников. Классы рисков используются для определения круга лиц (то есть ролей), имеющих право доступа к данным. Элемент данных с наивысшим классом риска в пользовательском разделе данных (к которому открыт доступ тем или иным пользователям) определяет класс риска всего раздела. Примеры классов риска:

- **Данные критического риска (Critical Risk Data, CRD).** Включают, в частности, персональные данные, несанкционированный доступ к которым агрессивно стремятся получить

злоумышленники как внутри, так и вне организации, ввиду их высокой и непосредственной финансовой ценности. Раскрытие CRD причинит ущерб не только отдельным лицам — владельцам данных, но и компании, которая может понести существенные убытки вследствие крупных штрафов и затрат на удержание клиентов и сотрудников, а также получить серьезный удар по своему имиджу и репутации.

- **Данные высокого риска (High Risk Data, HRD).** Являются объектом активных попыток несанкционированного использования из-за их потенциально высокой финансовой ценности. HRD дают компании преимущество перед конкурентами. Их утечка чревата недополученной прибылью и упущенными возможностями, а утеря — утратой доверия к бизнесу и, как следствие, упущенной выгодой, прямыми убытками вследствие возможных судебных преследований, штрафов и иных санкций со стороны надзорных органов, а также имиджевым и репутационным ущербом.
- **Данные умеренного риска (Moderate Risk Data, MRD).** Не представляют ощутимого интереса для взломщиков; однако несанкционированное использование любой не предназначенной для открытого доступа информации чревато негативными последствиями для компании.

1.3.5 Организация обеспечения безопасности

В зависимости от размера организации функция обеспечения информационной безопасности может находиться в ведении специальной группы обеспечения ИБ, обычно входящей в блок ИТ. В крупных компаниях часто имеется директор по информационной безопасности (Chief Information Security Officer, CISO), подчиняющийся либо CIO, либо CEO. В относительно небольших организациях, не имеющих штатных специалистов по ИБ, вся ответственность за обеспечение безопасности данных возлагается на сотрудников, занимающихся управлением данными. Последние, впрочем, должны участвовать в работе по обеспечению безопасности данных в любом случае.

В крупных организациях штатные сотрудники, занимающиеся вопросами ИБ, могут делегировать часть функций по руководству данными и авторизации пользователей бизнес-менеджерам. Например, назначение прав доступа и обеспечение нормативно-правового соответствия. Деятельность штатных выделенных специалистов по ИБ часто связана в основном с техническими аспектами защиты данных, включая борьбу с вредоносными программами и отражение хакерских атак. Тем не менее в ходе проектов по разработке и развертыванию приложений имеется обширное поле для сотрудничества специалистов по ИБ, управлению данными и бизнес-менеджеров.

Эта возможность синергетического взаимодействия часто не используется по причине разобщенной работы двух организационных систем — ИТ и управления данными. Для них, как правило, не предусмотрен процесс совместной работы с требованиями по обеспечению нормативно-правового соответствия и ИБ. Поэтому необходимо проработать стандартную процедуру взаимного информирования о требованиях регулирующих органов в области данных, угрозах потери или утечки информации, требованиях по защите данных, — и согласовываться всё это

должно с первых же шагов планирования любого проекта по разработке или развертыванию нового программного обеспечения.

Первым шагом в рамочной структуре управления рисками Национального института стандартов и технологий США (National Institute of Standards and Technology, NIST) является классификация всех данных организации[1]. Ключевым условием достижения этой цели является создание корпоративной модели данных. Без четкого и наглядного представления о расположении всей чувствительной информации невозможно приступить к созданию всеобъемлющей и эффективной программы защиты данных.

Специалисты по управлению данными должны активно привлекаться к сотрудничеству с разработчиками ИТ-решений и специалистами по ИБ при проведении работ по идентификации регламентируемых данных, реализации соответствующей защиты чувствительных систем и разработке и внедрению необходимых механизмов контроля доступа пользователей, с тем чтобы обеспечить надлежащий уровень конфиденциальности, целостности и соответствия требованиям регулирующих органов. Чем крупнее организация, тем важнее слаженность действий различных функциональных групп, подкрепляемая корректной и регулярно обновляемой корпоративной моделью данных.

1.3.6 Процессы обеспечения безопасности данных

Требования и процедуры, связанные с обеспечением безопасности, разбиваются на четыре группы, известные как «четыре A» (four A's): доступ (Access), аудит (Audit), аутентификация (Authentication) и авторизация (Authorization). С недавних пор к ним стали добавлять еще и «E» (пятую группу) — набор прав (Entitlement), — связанную с обеспечением нормативно-правового соответствия. Классификация информации, права доступа, ролевые группы, пользователи и пароли являются средствами реализации политики ИБ и требований «четырех A». Мониторинг безопасности не менее важен и служит средством проверки и подтверждения успешного функционирования других процессов. Мониторинг и аудит могут осуществляться непрерывно или периодически. Формальный аудит (чтобы его результаты были официально признаны) должен проводиться незаинтересованной стороной. Незаинтересованная сторона может быть как внутренней, так и внешней.

1.3.6.1 ТРЕБОВАНИЯ И ПРОЦЕДУРЫ «ЧЕТЫРЕ A»

- **Доступ.** Авторизованным лицам предоставляется своевременный доступ к информационным системам. В зависимости от контекста слово «доступ» (access) может указывать либо на действие — активное подключение к информационной системе и осуществление работы с данными, либо на факт — у субъекта есть действующее разрешение на работу с данными.
- **Аудит.** Проверка выполняемых операций по обеспечению безопасности и действий пользователей, которая проводится с целью подтверждения соответствия требованиям регулирующих

[1] См.: National Institute of Standards and Technology (US) (http://bit.ly/1eQYolG).

органов и соблюдения политики и стандартов организации. Специалисты по ИБ также периодически проверяют журналы и документы, чтобы удостовериться в выполнении всех вышеназванных требований. Результаты таких аудитов регулярно публикуются.

- **Аутентификация.** Подтверждает действительность доступа. Когда кто-то пытается войти в систему, она должна проверить, что это именно тот человек, реквизиты которого указаны. Один из способов проверки — с помощью паролей. Более строгие методы аутентификации включают использование токенов безопасности (security tokens), контрольных вопросов, отпечатков пальцев. Все данные в процессе аутентификации передаются в зашифрованном виде во избежание хищения.
- **Авторизация.** Заключается в выдаче различным лицам разрешений на доступ к определенным представлениям данных в соответствии с их ролями. После того как разрешение зафиксировано (принято решение о его выдаче), система контроля доступа при входе пользователя проверяет наличие у него действительного токена авторизации (authorization token). С технической точки зрения токен представляет собой запись в поле данных корпоративной службы каталогов Active Directory, указывающую на то, что пользователю выдано разрешение кем-то из ответственных лиц, имеющих доступ к соответствующим данным. Токен также показывает, что ответственное лицо приняло решение о выдаче разрешения, поскольку пользователю оно требуется для работы или полагается в силу должностного статуса.
- **Набор прав.** Определяет совокупность элементов данных, которые становятся доступными для пользователя после его авторизации. Ответственное лицо должно определить набор прав, предоставляемых пользователю, прежде чем выдать разрешение на доступ. Для того чтобы обеспечить выполнение требований нормативно-правового соответствия и конфиденциальности при принятии решений о назначении наборов прав, по каждому набору необходимо иметь перечень данных, доступ к которым открывается.

1.3.6.2 МОНИТОРИНГ

Системы должны включать средства мониторинга, позволяющие выявлять непредвиденные события, включая потенциальные взломы защиты и нарушения правил безопасности. В системах, содержащих конфиденциальную информацию, например бухгалтерские данные, часто реализованы механизмы активного мониторинга в режиме реального времени, мгновенно уведомляющие администратора безопасности о подозрительной активности или несанкционированном доступе.

Некоторые системы оснащены активной защитой, которая срабатывает автоматически и прерывает любые действия, не соответствующие профилю доступа. В этом случае учетная запись или процесс остаются в заблокированном состоянии, пока персонал службы ИБ не проведет анализ события.

Пассивный мониторинг, напротив, лишь отслеживает изменения, делая моментальные снимки состояния системы через установленные промежутки времени, и сравнивает фактические тенденции с контрольными эталонами или иными критериями. Система отправляет отчеты

распорядителям данных или администраторам, отвечающим за безопасность. Таким образом, если активный мониторинг является механизмом обнаружения угроз, то пассивный мониторинг относится к средствам оценки состояния систем и процессов.

1.3.7 Целостность данных

В сфере информационной безопасности под *целостностью данных* (*data integrity*) понимается их невредимое цельное состояние, предполагающее защиту от несанкционированного изменения, удаления или добавления данных. Например, в США законом Сарбейнса — Оксли установлены строгие требования по обеспечению целостности финансовой информации, предписывающие придерживаться строго определенных правил создания и редактирования финансовых данных.

1.3.8 Шифрование данных

Шифрование (*encryption*) — это процесс преобразования открытого текста в сложные коды с целью сокрытия информации ограниченного доступа, подтверждения передачи передаваемой информации или удостоверения личности отправителя. Зашифрованные данные невозможно прочесть без ключа или алгоритма для дешифрования, который обычно хранится отдельно и не может быть вычислен на основе других элементов данных в том же наборе. Четыре основных метода шифрования — хеширование, симметричное шифрование и асимметричное шифрование частным (закрытым) ключом и публичным (открытым) ключом — могут реализовываться с различными уровнями сложности и структурами ключей.

1.3.8.1 ХЕШИРОВАНИЕ

При хешировании используются различные алгоритмы математического преобразования данных. Для дешифрования данных нужно знать точные алгоритмы шифрования и порядок их применения. Иногда хеширование используют в качестве средства проверки целостности или идентичности данных после передачи. Наиболее распространенные алгоритмы хеширования — Message Digest (дайджест сообщения) 5 (MD5) и Secure Hashing Algorithm (безопасный алгоритм хеширования) (SHA).

1.3.8.2 ЧАСТНЫЙ КЛЮЧ

При шифровании с использованием частного (private) ключа (такие ключи еще называют закрытыми) один и тот же ключ применяется для шифрования и дешифрования данных, то есть идентичные экземпляры ключа должны иметься и у отправителя, и у получателя. Данные могут шифроваться посимвольно (поточным методом) или блоками. Распространенные алгоритмы шифрования данных с использованием частного ключа включают Data Encryption Standard (стандарт шифрования данных) (DES), Triple (тройной) DES (3DES), Advanced Encryption Standard (усовершенствованный стандарт шифрования) (AES), International Data Encryption Algorithm (международный алгоритм шифрования данных) (IDEA), Cyphers Twofish и Serpent. Простой алгоритм DES использовать не рекомендуется по причине относительной легкости взлома зашифрованных данных.

1.3.8.3 ПУБЛИЧНЫЙ КЛЮЧ

При шифровании с использованием публичного (public) ключа (такие ключи еще называют открытыми) отправитель и получатель применяют для шифрования и дешифрования данных различные ключи, причем отправитель часто использует ключ, находящийся в открытом публичном доступе, а вот получателю необходимо иметь частный ключ, чтобы раскрыть исходные данные. Этот вид шифрования полезен для отправки защищенных данных из множества источников единичным получателям (например, при централизованном сборе данных). К методам шифрования с использованием публичного ключа относятся алгоритм Ривеста — Шамира — Адельмана (Rivest — Shamir — Adelman, RSA) и протокол обмена ключами Диффи — Хеллмана (Diffie — Hellman Key Agreement). В свободном доступе имеется программное обеспечение PGP (Pretty Good Privacy — «достаточно хорошая секретность») для реализации всех функций шифрования с использованием публичного ключа.

1.3.9 Обфускация или маскировка данных

Данные можно сделать менее доступными посредством обфускации (представления в неясном или запутанном виде) или маскировки: в частности, путем подмены, искажения или перемешивания элементов в представлениях данных или внесения в них иных явных изменений без утери лежащего в основе смысла и без нарушения их связей с другими наборами данных (например, связей с помощью внешних ключей с другими объектами или системами). Значения атрибутов могут изменяться произвольным образом в пределах области допустимых значений этих атрибутов. Обфускацию полезно использовать, в частности, при выводе на экран чувствительной информации в качестве примера или при создании тестовых выборок данных из эксплуатационной БД, которые соответствовали бы логике приложения.

Маскировка является одним из видов обеспечения датацентричной (data-centric) безопасности. Различают два типа маскировки данных — постоянную (persistent) и динамическую. Постоянная маскировка, в свою очередь, может выполняться в процессе переноса данных (на лету — in-flight) или по месту их хранения (in-place).

1.3.9.1 ПОСТОЯННАЯ МАСКИРОВКА ДАННЫХ

Постоянная маскировка изменяет данные необратимым образом. В эксплуатационных средах такой тип маскировки обычно не используется, но находит достаточно широкое применение при переносе данных из эксплуатационной среды в среду разработки или тестирования. Постоянная маскировка приводит к необратимому изменению данных, но при этом они остаются пригодными для использования в целях тестирования процессов, приложений, отчетов и т. п.

- **Постоянная маскировка при переносе данных (на лету)** осуществляется посредством внесения в них изменений или искажений в процессе копирования из среды-источника (как правило, эксплуатационной среды) в другую (целевую) среду (обычно предэксплуатационную). Такая маскировка при правильной реализации обеспечивает очень высокую степень

безопасности, поскольку не оставляет никаких следов в виде промежуточных файлов или массивов незамаскированных данных. Другой плюс метода — возможность перезапуска процедуры переноса данных с маскированием в случае возникновения каких-либо проблем.

- **Постоянная маскировка по месту хранения данных** используется для перезаписи данных в одной и той же среде. Незамаскированные данные считываются из источника, маскируются, а затем записываются поверх исходных. Такая маскировка используется, если чувствительные данные хранятся в местах, где они присутствовать не должны (и связанный с этим риск необходимо снизить), или если есть хранящаяся в безопасном месте копия данных, которую нужно перенести в открытое хранилище (а перед этим замаскировать). Однако этот метод сопряжен с рисками. В частности, в случае сбоя в процессе маскировки данных восстановить весь обрабатываемый массив в пригодном для использования формате будет проблематично. Узкоспециализированные ниши для применения маскировки по месту хранения существуют, но в целом метод маскировки данных при переносе более приемлем для обеспечения проектных потребностей в части безопасности.

1.3.9.2 ДИНАМИЧЕСКАЯ МАСКИРОВКА ДАННЫХ

Динамическая маскировка изменяет представление данных для конечного пользователя или системы, оставляя исходные данные неизменными. Это крайне полезный метод, позволяющий приоткрывать пользователям некоторую часть чувствительной информации, хранящейся в эксплуатационной среде, не раскрывая ее целиком. Например, в базе данных хранится полный номер полиса социального страхования — 123456789, но оператору колл-центра для проверки личности обращающегося достаточно последних цифр, и на экране высвечивается ***-**-6789.

1.3.9.3 МЕТОДЫ МАСКИРОВКИ

Для маскировки или обфускации данных используются следующие методы и приемы.

- **Подстановка.** Символы или значения в элементах данных замещаются случайным образом или по определенной стандартной схеме. Например, все имена заменяются случайно выбранными из списка.
- **Перемешивание.** Перестановка элементов данных одного типа в записи или данных в столбце между строками таблицы. Например, перетасовав названия фирм-поставщиков в счетах-фактурах, мы получаем набор корректно оформленных счетов-фактур, не отражающий реальных источников поставок.
- **Временны́е отклонения.** Даты случайным образом изменяются в пределах ± *n* суток от фактических дат, где *n* достаточно мало, чтобы сохранились общие тенденции, но отлично от нуля, чтобы было невозможно идентифицировать по записям реальные события.
- **Отклонения значений.** Значения случайным образом изменяются в пределах ± *x* процентов от реальных, опять же не столь значительно, чтобы исказить тенденции, но на достаточную величину для того, чтобы данные нельзя было идентифицировать.

- **Обнуление или удаление.** Удаление данных, которые не должны присутствовать в среде тестирования.
- **Рандомизация.** Замена элементов данных или их частей случайными символами или группами случайных символов.
- **Шифрование.** Преобразование осмысленного и распознаваемого потока символов в нераспознаваемый поток посредством кодирования. Является наиболее радикальным вариантом маскировки данных по месту хранения.
- **Маскировка с пояснением.** Замена всех значений текстовым комментарием. Например, реальные текстовые строки в текстовом поле (потенциально с конфиденциальными данными) во всех записях БД заменяются на пояснение «Поле для комментария».
- **Маскировка ключа.** Маскировка ключевого поля БД помогает удостовериться в уникальности и воспроизводимости алгоритма или процедуры маскировки, а потому является важным средством проверки целостности данных в масштабах организации.

1.3.10 Термины сетевой безопасности

Защита данных должна обеспечиваться не только в местах хранения (data-at-rest — «данные в покое»), но и при их передаче из системы в систему (data-in-motion — «данные в движении»). Передача данных требует наличия сети, которая должна быть надежно защищена. Полагаться на одни лишь корпоративные межсетевые экраны недостаточно, поскольку они не вполне надежно защищают организацию от вредоносных программ и зараженных электронных писем. В случаях атак с использованием средств социально-психологического воздействия они и вовсе бессильны. Каждый компьютер в сети должен иметь собственную линию обороны, но в особо изощренной защите нуждаются веб-серверы, подверженные неиссякаемым угрозам со стороны всего мира, подключенного к интернету.

1.3.10.1 БЭКДОРЫ

Бэкдор[1] — оставленный по недосмотру или умышленно потайной вход в компьютерную систему или приложение. Он позволяет неавторизованным пользователям получить доступ к системе в обход парольной защиты. Часто тайные входы специально создаются разработчиками в целях, например, технического обслуживания. Бэкдоры другой распространенной категории предусматриваются создателями коммерческих программных продуктов.

Пароли по умолчанию, оставленные без изменения после установки любых систем или приложений, включая загружаемые с веб-страниц, представляют собой бэкдоры. Они, вне всякого сомнения, станут известны хакерам. Любой бэкдор влечет за собой риск, связанный с безопасностью.

1.3.10.2 БОТЫ ИЛИ ЗОМБИ

Сетевой *бот* (от «робот») или *зомби* — рабочая станция под дистанционным контролем злоумышленника, получившего доступ к функциям управления ОС и/или приложениями с помощью

[1] Бэкдор (от *англ.* back door — «черный ход», *букв.* «задняя дверь») переводится как «тайный вход». — *Примеч. пер.*

трояна, вируса, фишинга или в результате загрузки инфицированного файла. Боты используются для выполнения всевозможных вредоносных задач наподобие массовой рассылки спама, организации перегрузки серверов массовыми потоками бессмысленных запросов, выполнения незаконных денежных переводов, хостинга мошеннических веб-сайтов и т. п. *Ботнет* — это сеть ботов (инфицированных компьютеров)[1].

В 2012 году 17% персональных компьютеров (около 187 млн из 1,1 млрд) не имели антивирусной защиты[2]. В США в том же году этот показатель был даже хуже общемирового — 19,32% незащищенных ПК с выходом в интернет. Значительный процент таких ПК «зомбирован». По состоянию на 2016 год общее число активно работающих ПК с интернет-подключением составляло около 2 млрд единиц[3]. Учитывая тот факт, что в последние годы число ПК (настольных и портативных) было стремительно превзойдено числом всевозможных мелких гаджетов с интернет-подключением (смартфонов, планшетов и всевозможных «умных» устройств с веб-интерфейсами), многие из которых также используются для бизнес-транзакций, число всевозможных рисков для данных год от года будет только возрастать[4].

1.3.10.3 COOKIE-ФАЙЛЫ

Небольшие файлы *cookie* устанавливаются веб-сайтами на жесткий диск компьютера с целью его идентификации и сохранения профилей пользовательских настроек и предпочтений, с тем чтобы использовать их при последующих посещениях сайта тем же пользователем. Cookie-файлы широко используются в целях интернет-коммерции. Однако практика использования cookie весьма противоречива, поскольку вызывает немало вопросов относительно вторжения в личную жизнь благодаря возможности их использования в качестве компонентов шпионских программ.

1.3.10.4 МЕЖСЕТЕВОЙ ЭКРАН

Межсетевой экран (*firewall*) — программное и/или аппаратное средство фильтрации сетевого трафика, призванное защитить отдельный компьютер или локальную сеть от попыток несанкционированного доступа или взлома. Экран может сканировать как входящие, так и исходящие сообщения на предмет наличия в них информации ограниченного доступа и предотвращать несанкционированную передачу таких данных (предотвращение утечки данных — Data Loss Prevention). Некоторые межсетевые экраны могут также использоваться для ограничения доступа к отдельным внешним веб-сайтам.

---

[1] http://bit.ly/1FrKWR8, http://bit.ly/2rQQuWJ

[2] См.: http://tcrn.ch/2rRnsGr (17% компьютеров без антивирусного ПО), http://bit.ly/2rUE2R4, http://bit.ly/2sPLBN4, http://ubm.io/1157kyO (статистика числа компьютеров с ОС Windows 8 без антивирусной защиты).

[3] http://bit.ly/2tNLO0i, http://bit.ly/2rCzDCV, http://bit.ly/2tNpwfg

[4] По прогнозу Cisco Corporation, «к 2018 году в мире будет 8,2 млрд всевозможных портативных персональных и мобильных устройств и 2 млрд устройств с межмашинным интерфейсом (GPS-навигаторов, систем отслеживания на транспорте и производстве, медицинских приложений для быстрого получения электронных историй болезни и диагностики состояния здоровья пациентов и т. п.)»; см.: http://bit.ly/Msevdw

1.3.10.5 ПЕРИМЕТР

Периметр (*perimeter*) — граница между средой организации и внешними системами. Обычно между внутренней и внешней средой устанавливается межсетевой экран.

1.3.10.6 ДЕМИЛИТАРИЗОВАННАЯ ЗОНА (DMZ)

Демилитаризованная зона (*De-Militarized Zone, DMZ*) — буферная область между средой организации и сетью интернет. Среда организации отделяется от DMZ межсетевым экраном. Межсетевой экран также устанавливается на границе между DMZ и интернетом (см. рис. 64). Среды DMZ могут использоваться для временного хранения и обмена данными между организациями.

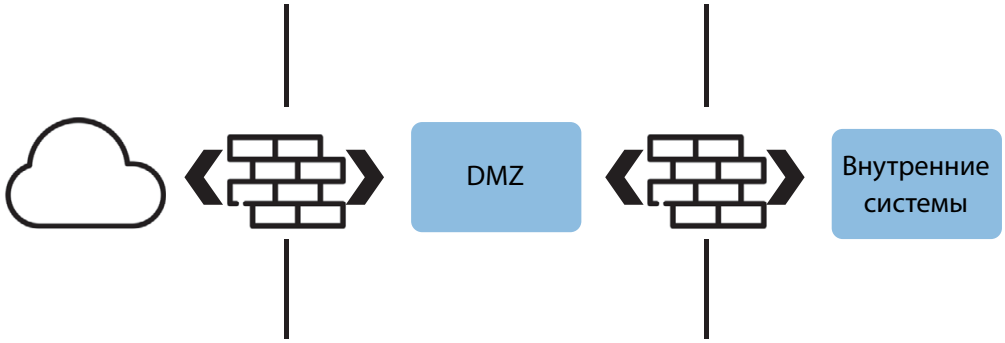


Рисунок 64. Пример DMZ

1.3.10.7 УЧЕТНАЯ ЗАПИСЬ СУПЕРПОЛЬЗОВАТЕЛЯ

Учетная запись суперпользователя (*Super User Account*) — учетная запись, которая имеет доступ с правами администратора (или root-доступ) и предназначается для использования в экстренных случаях. Такие учетные записи следует максимально защищать и выдавать лишь в крайних случаях при наличии документального подтверждения полномочий, а сроки их действия должны быть непродолжительными. Например, сотруднику, отвечающему за управление каким-либо производственным процессом, может потребоваться доступ к ряду систем, имеющих высокую значимость. Доступ должен быть предоставлен лишь на условиях строгого контроля по времени, ID пользователя, месту входа и прочим критериям, которые позволят избежать злоупотреблений.

1.3.10.8 КЛАВИАТУРНЫЕ ШПИОНЫ

Кейлоггеры (*Key Loggers*), или *клавиатурные шпионы*, — хакерские программы, регистрирующие последовательность нажатий клавиш пользователем компьютера с дальнейшей передачей этих сведений кому-либо через интернет. Так злоумышленники могут завладеть паролями, важными заметками, формулами, документами, веб-адресами и т. п. Часто такие программы устанавливаются вместе с ПО, загруженным с инфицированных веб-сайтов. Также это может произойти после загрузки с непроверенных ресурсов документов некоторых типов.

1.3.10.9 ТЕСТИРОВАНИЕ НА ПРОНИКНОВЕНИЕ

Установка защищенной сети или веб-сайта не является завершенной без тестирования средств защиты. Для этого используют тестирование на проникновение (Penetration Testing), иногда называемое просто «пентест» (penn test). При таком тестировании этичный хакер (ethical hacker) из числа собственных или приглашенных специалистов по ИБ предпринимает всяческие попытки несанкционированного проникновения в систему извне, воспроизводя действия настоящих хакеров, с тем чтобы выявить возможные уязвимости в ее защите. Ни одно приложение не должно выпускаться без предварительной проверки его защиты от взлома и/или злонамеренного использования с помощью тестирования на проникновение.

У части специалистов проверки защиты методом этичного хакинга вызывают неприятие или опасение, что ни к чему хорошему они не приведут, а лишь спровоцируют поиск крайних и сваливание ответственности друг на друга в случае выявления реальных недоработок. На самом же деле в динамичном противостоянии безопасности бизнеса криминальному хакерству любое программное обеспечение — как коммерческое, так и собственной разработки — содержит потенциальные уязвимости, о существовании которых на момент завершения разработки этих программ ничего известно не было. Следовательно, все программные реализации должны периодически проходить проверку на уязвимость. Поиск уязвимостей — непрекращающийся процесс, а при их обнаружении нужно искать не виновных, а средства устранения найденных слабых мест.

В качестве доказательства необходимости непрерывного совершенствования средств защиты программного обеспечения от угроз ИБ достаточно понаблюдать за потоками обновлений и исправлений систем безопасности, поступающих от ведущих поставщиков программного обеспечения. И свидетельствуют эти постоянные доработки средств защиты отнюдь не о том, что в программных продуктах изначально было не всё продумано, а о добросовестности и профессионализме служб технической поддержки клиентов этих компаний. Кстати, многие патчи («заплатки» — patches), предназначенные для устранения уязвимостей, и появляются по результатам этичного взлома, производимого самими поставщиками или по их заказу.

1.3.10.10 ВИРТУАЛЬНЫЕ ЧАСТНЫЕ СЕТИ (VPN)

VPN-подключения[1] позволяют создавать в небезопасном интернете выделенные защищенные каналы (или «туннели») обмена данными между клиентами корпоративной сети. Высокая надежность шифрования данных в «туннеле» дополнена контролем доступа пользователей к среде организации с использованием многофакторной аутентификации и защитного межсетевого экрана. После подключения все данные передаются в зашифрованном виде.

1.3.11 Виды безопасности данных

Обеспечение безопасности данных включает не только предотвращение несанкционированного доступа, но и поддержку надлежащего порядка доступа, который санкционирован. Доступ

[1] *сокр.* от *англ.* Virtual Private Network. — *Примеч. пер.*

к чувствительным данным должен контролироваться посредством предоставления разрешений («включения» — opt-in). Без разрешения пользователь не должен иметь возможности просматривать данные, не говоря уже о внесении в них изменений или иных действиях в системе. «Минимум привилегий» (Least Privilege) — важнейший принцип ИБ. Пользователю, процессу или программе должен быть разрешен доступ лишь к той информации, которая им требуется в законных целях.

1.3.11.1 БЕЗОПАСНОСТЬ ОБЪЕКТОВ

Безопасность объекта — первая линия защиты от посягательств на данные. Как минимум, объекты должны иметь защищенные от физического проникновения центры обработки данных, с доступом, предоставляемым только авторизованным сотрудникам. Социальные угрозы (см. раздел 1.3.15) являются самыми опасными, поскольку человеческий фактор традиционно считается слабейшим звеном безопасности любого объекта. Следите за наличием у сотрудников всех необходимых средств защиты данных на объектах и навыков их использования.

1.3.11.2 БЕЗОПАСНОСТЬ УСТРОЙСТВ

Мобильные устройства, включая ноутбуки, планшеты и смартфоны, уязвимы по своей природе, поскольку могут быть утеряны или похищены, подвергнуты физическому или электронному взлому со стороны злоумышленников. К сожалению, на них часто хранятся служебные данные, включая электронные и физические адреса, таблицы и документы, утечка которых чревата ущербом для организации и ее сотрудников и/или клиентов.

Взрывной рост числа портативных устройств и носителей данных требует наличия плана управления безопасностью всех этих средств хранения информации (включая корпоративные и находящиеся в личной собственности сотрудников), который должен рассматриваться как составная часть общей стратегической архитектуры безопасности компании. План должен включать использование как программных, так и аппаратных средств защиты данных. Стандарты безопасности устройств должны предусматривать:

- политики предоставления доступа к данным с мобильных устройств;
- правила хранения данных на мобильных устройствах и съемных носителях (DVD, CD, USB-накопителях и т. п.);
- порядок стирания данных и ликвидации носителей в соответствии с политиками управления записями;
- установку антивирусных программ и средств шифрования данных;
- информирование сотрудников в отношении имеющихся и потенциальных уязвимостей.

1.3.11.3 БЕЗОПАСНОСТЬ РЕКВИЗИТОВ ПОЛЬЗОВАТЕЛЯ

Каждый пользователь имеет реквизиты, используемые для входа в систему. Чаще всего это сочетания идентификатора пользователя (User ID) и пароля. Имеется достаточно широкий спектр вариантов использования реквизитов пользователей в системах, в зависимости от чувствительности

обрабатываемых в них данных и наличия возможностей подключения к репозиториям пользовательских реквизитов.

1.3.11.3.1 Системы управления идентификацией

Традиционно сложилось, что у каждого пользователя имеются отдельные учетные записи с паролями для каждого ресурса, платформы, прикладной системы или рабочей станции. Соответственно, всем пользователям приходится управлять множественными парами логин/пароль. При наличии в организации корпоративного каталога пользователей она может иметь механизм синхронизации между разнородными ресурсами в отношении управления паролями. В таких случаях задача пользователя упрощается: достаточно единожды ввести пароль (как правило, при входе в систему на рабочей станции), после чего все последующие аутентификации и авторизации происходят через обращение к каталогу пользователей. Системы идентификации, в которых реализована такая возможность, известны как системы, использующие технологию единого входа (single-sign-on). С точки зрения пользователей они являются оптимальными.

1.3.11.3.2 Стандарты по созданию идентификаторов пользователей в системах электронной почты

Все ID пользователей в пределах одного почтового домена должны быть уникальными. В большинстве компаний принято использовать имена или инициалы плюс фамилии (с добавлением цифр в случае коллизий) при формировании ID как для электронной почты, так и для корпоративной сети. Такой подход удобен для бизнес-контактов, поскольку имена и фамилии обычно известны.

Настоятельно не рекомендуется применять при формировании ID пользователей электронной почты или корпоративных сетей идентификационные номера сотрудников, поскольку обычно такие данные не предназначены для использования за пределами организации и должны быть защищены.

1.3.11.3.3 Стандарты по созданию паролей

Пароль — первая линия обороны от несанкционированного доступа. Каждая учетная запись должна быть защищена надежным паролем, устанавливаемым пользователем (владельцем учетной записи); при этом структура пароля должна соответствовать по своей сложности требованиям стандартов ИБ. Чем сложнее структура, тем надежней пароль в плане стойкости против взлома.

При создании новой учетной записи пользователя ей присваивается временный пароль для разового первичного входа в систему, подлежащий немедленной замене пользователем на собственноручно созданный и подтвержденный новый пароль, по которому и будут осуществляться последующие вхождения. Незаполненные поля паролей недопустимы.

Большинство экспертов по ИБ рекомендуют требовать от пользователей регулярной смены паролей не реже чем раз в 45–180 дней, в зависимости от характера системы, типа данных и режима секретности предприятия. Однако такой подход чреват тем, что из-за слишком частой смены

паролей сотрудники начинают в них путаться и, не полагаясь на память, берут в привычку записывать новые пароли (и хорошо, если только на бумаге).

1.3.11.3.4 Многофакторная идентификация

Некоторые системы требуют дополнительных идентификационных процедур: например, ввода отправленного на мобильное устройство пользователя кода подтверждения, использования для входа аппаратных компонентов, определения биометрических параметров (дактилоскопии, распознавания лица или радужной оболочки). Даже двух факторов обычно бывает достаточно для практически полного исключения возможности взлома защиты системы или входа на пользовательское устройство. Двухфакторная идентификация обязательна для всех пользователей, входящих в сеть, открывающую доступ к чувствительным данным.

1.3.11.4 Безопасность электронных коммуникаций

Пользователям нужно разъяснять недопустимость отправки персональных и личных данных, равно как и данных ограниченного доступа или конфиденциальной корпоративной информации по электронной почте или передачи их с помощью программ для общения. Эти общедоступные методы коммуникации не защищены от перлюстрации, перехвата, прослушивания, считывания и т. п. Отправитель электронного письма не в состоянии контролировать дальнейшие маршруты распространения содержащейся в нем информации. Письмо могут как перехватить по пути к получателю, так и перенаправить по любым посторонним адресам без ведома и согласия отправителя.

Вышесказанное также касается социальных медиа. Блоги, соцсети, порталы, вики, форумы и прочие интернет- и интранет-площадки следует априори считать незащищенными, не допускающими размещения или передачи через них любой конфиденциальной или чувствительной информации.

1.3.12 Типы ограничений при обеспечении безопасности данных

Два основных аспекта оказывают влияние на ограничения безопасности данных — уровень их конфиденциальности и нормативно-правовые требования, относящиеся к данным.

- **Уровни конфиденциальности.** *Конфиденциальный* означает секретный или частный. Организация самостоятельно определяет категории данных, которые предназначены для использования только в пределах организации или даже в пределах одной из ее частей. Общий принцип организации доступа к конфиденциальной информации — «необходимо знать» (need-to-know). Уровень конфиденциальности сведений определяется тем, насколько широк круг людей, которым необходимо их знать.
- **Нормативно-правовые требования.** Определяются действующими внешними нормами и правилами — законами, международными пактами, соглашениями с клиентами, отраслевыми регламентами и т. п. Общий принцип организации доступа к информации на основе

нормативно-правовых требований — «разрешено знать» (allowed-to-know). Порядок распространения подобных данных также может детально определяться регламентирующими нормативно-правовыми документами.

Основная разница между ограничениями, связанными с конфиденциальными данными, и данными, доступ к которым определяется нормативно-правовыми требованиями, состоит в том, откуда эти ограничения исходят. В первом случае ограничения определяются внутри организации, а во втором — накладываются извне.

Еще одно различие заключается в том, что любой набор данных, будь то документ или представление базы данных, может иметь лишь один уровень конфиденциальности. Этот уровень должен определяться по самому чувствительному (и относящемуся к самой высокой категории) элементу данных из числа имеющихся в наборе. В то же время подходы к категоризации на основе различных нормативно-правовых требований могут дополнять друг друга. В одном и том же наборе могут содержаться данные, на которые накладываются различные по источнику и характеру внешние ограничения. Для обеспечения нормативно-правового соответствия следует последовательно проверять соблюдение внешних ограничений по каждой категории данных (при одновременном соблюдении требований конфиденциальности).

При назначении пользователю набора прав доступа (определении совокупности отдельных элементов данных, к которым открывается доступ) должны соблюдаться все политики безопасности, вне зависимости от того, определяются они внутренними или внешними требованиями.

1.3.12.1 КОНФИДЕНЦИАЛЬНЫЕ ДАННЫЕ

Степень конфиденциальности данных может варьироваться от самой высокой (к примеру, к данным о размерах заработной платы сотрудников имеют доступ единицы) до низкой (каталоги продукции в открытом доступе). Типичная схема классификации может включать два и более уровней (классов) конфиденциальности информации, включая (в достаточно общем случае) следующие пять уровней.

- **Для всеобщего пользования (for general audiences).** Общедоступная, в том числе и для широкой публики, информация.
- **Для внутреннего пользования (internal use only).** Информация, сообщаемая или доступная только сотрудникам или членам организации, разглашение которой, однако, не влечет особого риска. Обсуждать или показывать такую информацию любым лицам не запрещено, однако снятие копий для их использования за пределами организации не допускается.
- **Конфиденциальная (confidential).** Сведения, не подлежащие распространению за пределами организации без подписания их получателем соглашения о конфиденциальности и неразглашении информации (non-disclosure agreement) или другого аналогичного документа. Конфиденциальная информация о клиенте не подлежит передаче другим клиентам ни при каких условиях.

- **Конфиденциальная, ограниченного доступа (restricted confidential).** К информации допускается ограниченный круг лиц, которым ее «необходимо знать» для исполнения своих обязанностей. Допуск может требовать проверки сотрудника службой безопасности.
- **Строго конфиденциальная, под подписку о неразглашении (registered confidential).** Информация настолько конфиденциального свойства, что для получения доступа к ней человек обязан подписать особое юридическое соглашение о принятии на себя ответственности за сохранение данных в тайне.

Уровень конфиденциальности никак не связан с внешними ограничениями нормативно-правового характера. В частности, знание того, что информация имеет определенный уровень конфиденциальности, не дает специалисту по управлению данными возможности определить, распространяется ли на нее, например, запрет на трансграничное перемещение или ограничение доступа строго определенными категориями сотрудников наподобие предусмотренного американским законом HIPAA (Закон о преемственности и учете данных в медицинском страховании).

1.3.12.2 РЕГЛАМЕНТИРУЕМЫЕ ДАННЫЕ

Обращение с некоторыми категориями информации регламентируется извне[1] — законами, подзаконными актами, отраслевыми стандартами или действующими договорами и соглашениями, предписывающими порядок использования данных, доступа к ним, а также определяющими круг лиц, имеющих право доступа, и допустимые цели использования данных. Поскольку всевозможных регламентирующих документов имеется множество, а их требования часто накладываются на одни и те же данные, проще бывает разделить существующие нормы и правила по предметным областям на несколько категорий или «семейств», чтобы специалистам по управлению данными проще было разбираться с действующими требованиями и обеспечивать их соблюдение.

Категории регламентирующих норм и правил каждой организации приходится определять самостоятельно, с тем чтобы они наилучшим образом отражали именно ее ситуацию и отвечали ее собственным потребностям в отношении нормативно-правового соответствия. Важно максимально упростить структуру этих категорий и процедуры выполнения требований, чтобы обеспечить эффективную практическую реализацию мер по защите. Если защитные меры в различных категориях схожи, их следует объединять в «семейство». В каждой регламентируемой категории меры по защите данных должны быть проверяемыми, поскольку это не инструмент управления организацией, а метод обеспечения соблюдения обязательных требований.

Поскольку нормативные документы зависят от отрасли, организациям нужно определить группы норм и правил, которые отвечают их операционным потребностям. Например, если компания работает исключительно на отечественных рынках, ей нет необходимости учитывать действующие правила, связанные с экспортными операциями.

[1] Далее в разделе 2.3.2 для обозначения подобного рода информации вводится термин «регламентируемая информация» (regulated information). — *Примеч. науч. ред.*

Однако, поскольку в каждой стране имеется собственный свод законов и норм, регламентирующих защиту персональных данных и данных о частной жизни, а среди клиентов организации могут оказаться граждане самых разных стран, правильным решением будет обобщить все нормативно-методические документы в области защиты данных о частной жизни в единое семейство регламентирующих норм и правил и принять меры по соблюдению требований любых государств. Подобное решение обеспечивает нормативно-правовое соответствие в большинстве юрисдикций и предлагает единый стандарт, которому нужно следовать.

Примером возможного требования нормативно-правового соответствия может служить запрет на перемещение из страны происхождения за рубеж массивов данных, содержащих элементы определенной категории. Некоторые регламентирующие документы, как национальные, так и международные, предусматривают подобные ограничения.

Оптимальное число категорий регламентирующих норм и правил — не более девяти. Ниже описаны примеры подобных категорий.

1.3.12.2.1 Примеры семейств регламентирующих норм и правил

Некоторые регламентирующие нормативно-правовые документы точно определяют по именам элементы данных, подлежащие защите, а также устанавливают требования в отношении порядка их защиты. Не следует относить каждый элемент данных подобного рода к отдельной категории; вместо этого используйте единое семейство мер по защите всех элементов. В число подобных категорий могут быть включены, например, данные о держателях платежных карт, хотя порядок их хранения и обработки обычно регулируется не законодательством, а контрактными обязательствами. В индустрии платежных карт (Payment Card Industry, PCI) контрактные обязательства во всем мире в основном идентичны.

- **Персональная идентификационная информация.** Персональная идентификационная информация (Personal Identification Information, PII), называемая также персональной частной информацией (Personally Private Information, PPI), включает любую информацию, по которой можно идентифицировать отдельного человека. К такой информации можно отнести следующие сведения: ФИО, адреса, телефоны, данные паспортов и иных официально выданных документов, номера счетов, возраст, расовая, этническая и религиозная принадлежность, дата рождения, ФИО членов семьи, родственников или друзей, данные о работе (HR data) и во многих случаях оклады. Очень похожие меры по защите таких сведений обеспечивают выполнение требований директив ЕС о неприкосновенности частной жизни, канадского Закона о защите личных сведений и электронных документов (PIPEDA), аналогичного японского закона 2003 года (PIP Act 2003), стандартов PCI, документов Федеральной торговой комиссии США (US Federal Trade Commission, FTC), закона Грэмма — Лича — Блайли о финансовой модернизации (Gramm — Leach — Bliley Financial Services Modernization Act, GLBA), равно как и множества других законов, регламентов и стандартов, регулирующих порядок хранения и использования персональной идентификационной информации.

- **Чувствительные данные финансового характера.** Включают любые данные об акционерах, интересах и прочую так называемую «инсайдерскую» информацию, включая показатели еще только готовящейся к публикации финансовой отчетности. Сюда же относятся не афишируемые бизнес-планы, включая планируемые слияния и поглощения или, напротив, разукрупнения, отчеты о серьезных текущих проблемах, планируемых изменениях в составе высшего руководства, детализированные и сводные отчеты о продажах, заказах, текущем платежном балансе и т. п. Все данные, подобные перечисленным выше, могут быть объединены в одну категорию, и их защита может осуществляться с использованием единого набора политик. В США подобные вопросы регулируются целым рядом актов, часть которых в явном виде описывает круг лиц, имеющих право доступа к финансовым данным и ответственных за их целостность[1].
- **Чувствительные медицинские данные / Персональная информация о состоянии здоровья.** В США обращение с чувствительными медицинскими данными и персональной информацией о состоянии здоровья (Personal Health Information, PHI) регулируется Законом о преемственности и подотчетности медицинского страхования (Health Information Portability and Accountability Act, HIPAA). Ограничения на обращение с информацией подобного рода законодательно установлены также в других странах и становятся всё строже. Поэтому корпоративные юристы должны принимать во внимание необходимость соблюдения всех установленных на уровне национальных законодательств требований в данной области во всех странах, где организация ведет свой бизнес или имеет клиентов.
- **Записи об образовании.** К записям об образовании (Educational Records) относится вся информация об образовании отдельных людей. В США обращение с данными об образовании регулируется Законом о правах семьи на образование и неприкосновенность частной жизни (Family Educational Rights and Privacy Act, FERPA).

1.3.12.2.2 Отраслевые стандарты и договорные обязательства

В некоторых отраслях действуют специфические стандарты, регламентирующие запись, хранение и шифрование данных. В некоторых отраслях также запрещены удаление, редактирование или передача данных в определенные места. Например, в фармацевтической, пищевой и косметической промышленности, а также в высокотехнологичных отраслях введены ограничения на передачу или хранение определенных видов информации за пределами страны происхождения или действуют требования по шифрованию данных при их передаче.

- **Стандарт безопасности данных индустрии платежных карт (PCI DSS).** PCI DSS (Payment Card Industry Data Security Standard) — один из наиболее известных отраслевых стандартов в области ИБ. Он определяет порядок и механизмы защиты информации, по которой можно

[1] В частности, речь идет о законодательных запретах на использование закрытой внутренней информации в операциях на фондовых рынках, а именно о законах Сарбейнса — Оксли (SOX) и Грэмма — Лича — Блайли (GLBA).

идентифицировать владельцев счетов в кредитно-финансовых организациях, включая фамилии/имена, номера и сроки действия кредитных карт и банковских счетов. Обращение с большинством данных подобного рода регулируется законами и политиками. Любые данные, отнесенные к этой категории (обозначенные соответствующим образом в метаданных), должны тщательно отслеживаться распорядителями данных и анализироваться на предмет правомочности их использования в тех или иных базах данных, приложениях, отчетах, информационных панелях или пользовательских представлениях.

- **Нормативно-правовые документы в области защиты конкурентных преимуществ и коммерческой тайны.** Собственные методы, химические составы, формулы, исходные тексты, конструкции, инструменты, рецепты или технические приемы, используемые компаниями для получения преимуществ над конкурентами, могут быть защищены отраслевыми нормативными документами и/или законами о защите интеллектуальной собственности.
- **Ограничения, накладываемые договорами.** Ограничения, накладываемые договорами с поставщиками, клиентами или партнерами организации, могут включать прямые указания на допустимые и недопустимые способы использования конкретных видов данных. Иногда в договорах также содержатся перечни информации, которая может быть передана третьим лицам, а также информации, не подлежащей разглашению. К разряду конфиденциальных данных могут быть отнесены, например, результаты мониторинга состояния окружающей среды, отчеты о транспортировке и складировании опасных грузов, номера партий, грузовые терминалы, время приготовления, номера счетов, а также национальные идентификационные номера граждан определенных зарубежных стран. Некоторые узкоспециализированные технологические компании могут включать в эту категорию данные о продуктах или ингредиентах строго ограниченного использования.

1.3.13 Риски, связанные с безопасностью систем

Первым шагом по выявлению рисков является точное определение мест хранения чувствительных данных и необходимых мер по их защите. Также необходимо выявить риски, обусловленные спецификой используемых информационных систем. Риски, обусловленные недостаточной защищенностью системы, чреваты компрометацией сети или базы данных. Наличие уязвимостей, во-первых, позволяет сотрудникам, имеющим законный доступ к информации, злонамеренно или по оплошности использовать ее не по назначению, а во-вторых, способствует успеху хакерских атак.

1.3.13.1 ЗЛОУПОТРЕБЛЕНИЕ ИЗБЫТОЧНЫМИ ПРИВИЛЕГИЯМИ

Открывая доступ к данным, процессам или программам, следует неукоснительно придерживаться принципа предоставления пользователям минимума привилегий, необходимого для использования информации по прямому и законному назначению. Всегда есть риск использования сотрудником избыточных прав и полномочий, предоставленных ему сверх требуемого для исполнения прямых должностных обязанностей минимума, во вред организации, причем произойти

это может как по злому умыслу, так и по чистой случайности. Избыточный набор прав доступа (избыточные привилегии — excessive privilege) часто предоставляется пользователям всего лишь потому, что это проще сделать технически, чем детально разбираться с тем, какие права реально нужны пользователю для работы, а какие нет, в каждом конкретном случае. У администраторов БД может просто не хватать времени или метаданных для определения и обновления механизмов детализированного контроля доступа для каждого пользователя. В результате множеству пользователей открывается доступ к данным с набором прав по умолчанию, наделяющий их функциональными возможностями, значительно превосходящими их реальные потребности. Именно из-за рисков, связанных с избыточными привилегиями, во многих регламентирующих документах в области данных отдельно определяются требования по безопасности.

Хорошим решением проблемы избыточных привилегий служит контроль доступа на уровне запросов (query-level access control), поскольку этот механизм ограничивает права доступа к базе данных до минимума, необходимого для обработки SQL-запроса и выдачи только тех данных, которые запрошены. Уровень детализации контроля доступа должен не ограничиваться таблицами, а прорабатываться на уровне отдельных строк и столбцов. Контроль доступа на уровне запросов полезен еще и тем, что позволяет выявлять случаи злоупотребления избыточными привилегиями со стороны недобросовестных или преследующих преступные цели сотрудников.

Большинство программных реализаций баз данных включают те или иные возможности контроля доступа на уровне запросов (триггеры, защиту на уровне строк и таблиц, пользовательские представления), но необходимость настраивать все эти встроенные средства защиты данных вручную ограничивает возможности их практического применения. Сама по себе процедура определения правил доступа на уровне запросов для всех пользователей по всем строкам, столбцам и операциям крайне трудоемка и затратна по времени. Что хуже, роли пользователей имеют свойство изменяться и перераспределяться, что вынуждает регулярно переопределять правила обработки запросов. Большинству администраторов БД бывает затруднительно выкроить достаточно времени на настройку доступа на уровне запросов даже однократно и для небольшого круга пользователей, так что о регулярной поддержке ограничений для сотен пользователей не может быть и речи. В результате во многих организациях для реализации по-настоящему эффективного контроля доступа к данным на уровне запросов требуется применение автоматизированных средств настройки.

1.3.13.2 ЗЛОУПОТРЕБЛЕНИЕ ОБОСНОВАННЫМИ ПРИВИЛЕГИЯМИ

Обоснованные привилегии доступа к базе данных могут быть использованы для совершения несанкционированных действий. Представьте себе, к примеру, склонного к преступной деятельности медработника, обладающего привилегиями на просмотр электронных историй болезни пациентов с помощью специального веб-приложения.

Структура корпоративных веб-приложений обычно не допускает просмотра пользователем больше одной истории болезни за одно обращение и не предусматривает возможности копирования данных. Однако замысливший хищение данных медик вполне может обойти эти ограничения,

подключившись к базе данных с помощью альтернативного клиентского приложения, хотя бы даже просто MS Excel. С помощью MS Excel и своих легальных реквизитов входа в систему такой медработник вполне может выгрузить и сохранить у себя все электронные истории болезни пациентов, имеющиеся в БД.

Следует различать два типа рисков, связанных с несанкционированным использованием привилегий, а именно риски преднамеренного и непреднамеренного злоупотребления. В первом случае работник имеет злой умысел использовать имеющиеся у организации данные не по назначению: например, продать базу данных пациентов или шантажировать пациентов, вымогая у них деньги под угрозой огласки сведений об их заболеваниях. На практике, однако, гораздо чаще приходится сталкиваться с риском непреднамеренной утечки данных вследствие нарушений правил обращения с ними. К примеру, добропорядочный медработник делает ровно то же самое, что и злоумышленник, то есть скачивает множество историй болезни на свой рабочий компьютер, чтобы выполнить какую-то вполне законную работу. Сохраненные на внешнем компьютере данные тут же попадают в зону повышенного риска, поскольку такой компьютер не застрахован от похищения злоумышленниками.

Частичным решением проблемы злоупотребления обоснованным привилегированным доступом к базам данных может служить дополнение контроля на уровне запросов определением политик предоставления удаленного доступа с пользовательских компьютеров строго в рабочее время, мониторингом местонахождения подключенных устройств и объемов загружаемой на них информации и, конечно же, ограничением возможностей по получению любым отдельно взятым пользователем доступа ко всем записям, содержащим чувствительную информацию, если им этого не требуется для выполнения конкретного рабочего задания. Предоставлять такой доступ следует только с разрешения вышестоящего руководства. Например, представителям организации на местах требуется доступ к персональным данным своих клиентов, но всякую возможность скачать к себе на ноутбук всю клиентскую базу данных ради «экономии времени» им следует закрыть.

1.3.13.3 НЕСАНКЦИОНИРОВАННОЕ ПОВЫШЕНИЕ ПРИВИЛЕГИЙ

Злоумышленники могут воспользоваться уязвимостями СУБД для превращения учетной записи рядового пользователя в учетную запись привилегированного пользователя или даже администратора. Уязвимости подобного рода могут иметь место, например, в хранимых процедурах, встроенных функциях, реализациях протоколов и даже в SQL-выражениях. Допустим, разработчик ПО в коммерческом банке обнаруживает уязвимость некой функции и с ее помощью присваивает себе административные привилегии, а затем отключает механизмы аудита и создает фиктивные счета, осуществляет денежные переводы или закрывает существующие счета.

Защиту от эксплуатации уязвимостей с целью несанкционированного повышения привилегий можно реализовать путем сочетания традиционных систем предотвращения вторжений (Intrusion Prevention System, IPS) с системами предотвращения вторжений, применяющими контроль доступа на уровне запросов. Эти системы проверяют трафик базы данных на предмет

выявления структур данных, характерных для известных уязвимостей. Например, если некая функция уязвима для атаки определенного вида, IPS заблокирует доступ либо к ней, либо к процедурам, позволяющим осуществить атаку.

Сочетание IPS с альтернативными средствами мониторинга атак, в частности с контролем доступа на уровне запросов, позволяет повысить точность при выявлении атакующих воздействий. IPS выявляет запросы к БД, использующие уязвимую функцию, а система контроля доступа на уровне запросов проверяет, соответствует ли запрос нормальному пользовательскому поведению. Если при выполнении какого-то запроса одновременно фиксируются обращение к уязвимой функции и необычное поведение, это с очень высокой вероятностью говорит о проведении атаки.

1.3.13.4 ЗЛОУПОТРЕБЛЕНИЕ СЕРВИСНЫМИ И ОБЩИМИ УЧЕТНЫМИ ЗАПИСЯМИ

Использование сервисных учетных записей (пакетных ID — batch IDs) и общих учетных записей (generic IDs) резко повышает риск нарушений ИБ и осложняет возможность определения источника нарушения. В некоторых организациях этот риск еще больше усугубляют, настраивая систему мониторинга таким образом, чтобы она игнорировала уведомления средств защиты, касающихся этих записей. Ответственным за ИБ следует самым тщательным образом рассмотреть и в полной мере задействовать инструменты для безопасного управления сервисными и общими учетными записями.

1.3.13.4.1 Сервисные учетные записи

Сервисные учетные записи очень удобны, поскольку они могут обеспечить повышенный уровень доступа для процессов, которые их используют. Однако в случае их применения не по назначению отследить пользователей или администраторов, которые ими воспользовались, будет невозможно. Пока у сервисных записей нет доступа к ключам дешифрования, они не представляют угрозы для зашифрованных данных. Это особенно важно учитывать при организации работы с серверами, на которых хранятся юридические документы, медицинская информация, информация, относящаяся к коммерческой тайне, а также конфиденциальные планы руководства.

Следует ограничивать права сервисных учетных записей только доступом к определенным задачам или командам в конкретных системах. Кроме того, все действия по выдаче разрешений требуют документирования и подтверждения их правомерности. Целесообразно рассмотреть возможность назначения нового пароля всякий раз при выдаче разрешений, используя для этого ту же процедуру, что и для переназначения паролей учетных записей суперпользователей.

1.3.13.4.2 Общие учетные записи

Общие учетные записи создаются в тех случаях, когда приложение не поддерживает управление достаточным числом индивидуальных учетных записей пользователей или когда добавление пользователей слишком трудоемко либо требует покупки дорогостоящих дополнительных лицензий. Пароли у общих учетных записей зачастую подолгу или вовсе не изменяются, поскольку

уведомить всех фактических пользователей, входящих в систему с использованием таких записей, проблематично. Ввиду того что доступ к общим записям носит, по сути, неуправляемый характер, следует тщательно взвешивать и анализировать все возможные последствия их применения. Во всяком случае, разрешать их использование по умолчанию категорически недопустимо.

1.3.13.5 СЕТЕВЫЕ ВТОРЖЕНИЯ И АТАКИ НА ПЛАТФОРМУ

Обеспечение защиты от вторжений применительно к информационным активам требует сочетания регулярных обновлений ПО и внедрения специальной системы предотвращения вторжений (Intrusion Prevention System, IPS). Часто (но не всегда) IPS дополняется системой обнаружения вторжений (Intrusion Detection System, IDS). Цель создания такого защитного комплекса — предотвратить подавляющее большинство попыток сетевого вторжения и обеспечить немедленное реагирование на всё-таки удавшиеся атаки. Наиболее простая форма защиты от вторжения — межсетевой экран, но в условиях наличия мобильных пользователей, веб-доступа, а также широкого применения во многих корпоративных средах мобильных вычислительных устройств использование обычного межсетевого экрана становится хотя и необходимым, но недостаточным.

Регулярные обновления ПО, предоставляемые поставщиками, уменьшают количество уязвимостей, обнаруживаемых со временем в платформах баз данных. К сожалению, обновления часто устанавливаются в организациях в соответствии с периодическими циклами обслуживания систем, а не сразу же после появления «заплат» (patches) безопасности. В промежутках между циклами обновлений базы данных остаются незащищенными. Кроме того, из-за возникающих после обновлений проблем с совместимостью от них иногда приходится вовсе отказываться. Поэтому для решения проблем ИБ, обусловленных всеми вышеперечисленными факторами, следует реализовывать IPS.

1.3.13.6 SQL-ИНЪЕКЦИИ

SQL-инъекцией (SQL injection) называется внедрение злоумышленником неавторизованного кода в SQL-программы в уязвимых местах приложений баз данных, таких как хранимые процедуры или поля ввода веб-приложений. Внедренные SQL-выражения воспринимаются СУБД как легитимные команды и выполняются. Используя такой прием, хакеры могут получить неограниченный доступ ко всей БД.

SQL-инъекции также применяются для атаки на СУБД посредством передачи SQL-команд как параметра функции или хранимой процедуры. Например, компонент, обеспечивающий резервное копирование, обычно имеет высокий уровень привилегий; обращение к уязвимой в отношении SQL-инъекций функции в процессе выполнения этого компонента может позволить рядовому пользователю повысить свои привилегии до уровня администратора и получить контроль над базой данных.

Для минимизации этого риска весь поток входящих данных должен проходить санитизацию (sanitizing) до передачи на сервер БД.

1.3.13.7 ПАРОЛИ ПО УМОЛЧАНИЮ

Давно сложилась практика поставки программных продуктов в таком виде, что при их установке создаются учетные записи пользователей по умолчанию. В одних случаях они используются в процессе установки ПО, в других — в процессе его тестирования перед началом эксплуатации.

Пароли по умолчанию часто входят в состав демоверсий ПО. Последующая установка еще каких-то приложений других разработчиков повлечет появление новых учетных записей и паролей по умолчанию. К примеру, система управления отношениями с клиентами (CRM) может создать при установке несколько учетных записей в базе данных — инсталляционную и отладочную, а также учетную запись администратора и учетную запись обычного пользователя. Система управления ресурсами предприятия (например, SAP) тоже создает при ее установке ряд учетных записей по умолчанию. Среди разработчиков коммерческих СУБД такой подход также практикуется.

Злоумышленники же постоянно изыскивают простые пути доступа к чувствительным данным. Следует минимизировать эту угрозу с помощью определения уникального имени пользователя с надежным паролем и проверки учетных записей СУБД на отсутствие паролей по умолчанию. Очистка системы от паролей по умолчанию — важный элемент ИБ, и проводиться она должна после любой установки.

1.3.13.8 ЗЛОУПОТРЕБЛЕНИЕ ДАННЫМИ РЕЗЕРВНЫХ КОПИЙ

Резервные копии БД призваны снизить риск утери данных, но увеличивают риски, связанные с безопасностью. То и дело в новостях рассказывают о пропажах носителей с резервными копиями важных данных. Все резервные копии БД должны храниться в зашифрованном виде. Без ключа дешифрования злоумышленники не получат доступа к вашим данным, даже если им удастся похитить носители с резервной копией или перехватить ее при передаче по электронным каналам связи. Ключи следует хранить отдельно в надежно защищенном удаленном месте. Важно, чтобы они были доступны в случае, если потребуется аварийное восстановление БД из резервной копии.

1.3.14 Хакеры и хакинг

Термин *хакинг* (*hacking*) появился еще в те времена, когда найти удачное нетривиальное решение для эффективного выполнения какой-либо вычислительной задачи считалось достижением. Иными словами, «хакер» — отнюдь не синоним «взломщика», а просто изобретательный программист, умеющий изыскивать и использовать мало кому известные операции и подходы к построению хитроумных схем разрешения проблемных ситуаций в сложных компьютерных системах. Другое дело, что хакеры бывают как хорошими, так и плохими.

Этичный, или «белый», хакер (а также «белая шляпа») работает над совершенствованием системы (термин «белая шляпа» отсылает нас к американским вестернам, в которых герой всегда носит такой головной убор). Без этичных хакеров исправимые уязвимости вскрывались бы исключительно по случайности, а чаще в результате их использования злоумышленниками. Систематические обновления систем безопасности компьютеров — результат работы по выявлению уязвимостей, ведущихся в рамках «белого» хакинга.

На другом полюсе спектра существуют злонамеренные, или «черные», хакеры. Это взломщики компьютерных систем, цель которых — хищение конфиденциальных данных или причинение ущерба. «Черные» хакеры обычно охотятся за финансовой информацией, персональными данными или сведениями личного характера с целью незаконного обогащения за чужой счет. Именно они взламывают пароли (чаще всего методом перебора простых вариантов) и отыскивают недокументированные слабые места и тайные входы в имеющихся системах. Иногда их называют «черными шляпами» (в тех же американских вестернах такие шляпы носят злодеи).

1.3.15 Социальные угрозы безопасности / Фишинг

Социальные угрозы (social threats) безопасности выражаются в использовании общения с обладателями доступа к защищенной информации (при личных встречах, по телефону или через интернет) для выведывания у них конфиденциальных данных или способа доступа к ним с целью последующего использования этих данных в неблаговидных целях.

Под *социальной инженерией* (*social engineering*) понимается совокупность приемов и хитростей, используемых хакерами-злоумышленниками для выманивания у людей информации или доступа к ней. Выведав некую частность у одного из сотрудников, хакер может умело использовать ее для убеждения других сотрудников в наличии у него правовых оснований на получение ответов на интересующие его вопросы. Иногда мошенники поочередно общаются со многими сотрудниками низового звена, пока не соберут по крупицам всю информацию, необходимую им для того, чтобы втереться в доверие к вышестоящему сотруднику, после чего переходят к его или ее обработке.

Под *фишингом* (от *англ.* fishing — рыбная ловля, выуживание) понимается деятельность по совершению телефонных звонков, а также отправке мгновенных сообщений или писем по электронной почте с целью «выуживания» у получателей ценной или частной информации, причем такими способами, что получатели даже не сознают, что они ее разглашают. Часто такие звонки или сообщения маскируются под исходящие из вполне легитимного источника. Например, поступает заманчивое предложение купить что-либо по низкой цене или оформить кредит с выгодной процентной ставкой. Если «рыба» клюет на наживку, ей предлагается заполнить анкету или ответить на вопросы, в ходе чего и выведываются персональные и конфиденциальные данные — например, полные имена, адреса и пароли, номера документов или кредитных карт и т. п. Для сведе́ния к минимуму подозрений вопросы часто формулируются как просьба «обновить» или «подтвердить» якобы и без того имеющуюся информацию. Фишинговые мгновенные сообщения или электронные письма также могут использоваться для перенаправления пользователей на мошеннические веб-сайты, где у них обманом выманиваются персональные данные. Особую опасность, однако, представляет фабрикация штучных фишинговых электронных писем в адрес высокопоставленных руководителей с уважительным обращением по имени. На хакерском жаргоне эта разновидность ловли крупной рыбы называется «гарпунной охотой на китов» (spear-phishing for whales). Помимо фишинговых обзвонов и рассылок известны и случаи физического проникновения хакеров на сайты-мишени под видом, например, партнеров или

поставщиков и вступления там в прямые переговоры с сотрудниками с целью получения доступа к чувствительной информации[1].

1.3.16 Вредоносные программы

Понятие *вредоносные программы* (*malware*) используется для собирательного обозначения любых программ, создаваемых с целью повреждения, изменения или хищения данных. К вредоносным программам относятся компьютерные вирусы, «черви», программы-шпионы, включая перехватчики ввода с клавиатуры, рекламные программы и т. п. По большому счету, любую программу, устанавливаемую без ведома и согласия владельца компьютера или системы, можно считать вредоносной хотя бы уже по той причине, что она занимает место на жестком диске и, скорее всего, отбирает часть системных ресурсов. В зависимости от целей (репликация, уничтожение, разведывание или хищение данных, мониторинг процессов или поведения и т. д.) вредоносные программы могут принимать самые разнообразные формы.

1.3.16.1 РЕКЛАМНЫЕ ПРОГРАММЫ

Рекламные программы (adware) — разновидность шпионских программ, которые обычно устанавливаются на компьютер в результате загрузки каких-либо материалов из интернета и отслеживают определенные действия пользователя (например, совершение поисковых запросов и посещение веб-сайтов). Также рекламная программа может встраивать в браузер пользователя какие-либо дополнения или объекты — к примеру, панели инструментов. Сами по себе такие программы противозаконными не являются, но могут использоваться для накопления данных о предпочтениях пользователей, которые могут затем продаваться, допустим, коммерческим маркетинговым фирмам. Самая же главная угроза, исходящая от рекламных программ, заключается в том, что злоумышленники могут использовать их функциональность для внедрения по-настоящему зловредных кодов и кражи идентификационных данных.

1.3.16.2 ШПИОНСКИЕ ПРОГРАММЫ

К шпионским (spyware) относятся любые программы, проникающие на компьютер без ведома владельца и отслеживающие все его действия в режиме онлайн. Программы-шпионы имеют опасное свойство цепляться к любым вполне безобидным программам. Загрузив с веб-сайта пакет установки бесплатного приложения, пользователь рискует получить вместе с ним и шпионскую программу, которая «тихо» установится и будет работать в фоновом режиме без ведома пользователя. В зависимости от типа шпионские программы могут отслеживать различные действия. Какие-то ведут мониторинг посещаемых веб-страниц, какие-то протоколируют и передают злоумышленникам всю последовательность нажатий клавиш с целью хищения номеров кредитных карт, логинов/паролей входа в системы и прочих конфиденциальных данных.

[1] См., например, отчет ФБР о хакерских атаках из-за рубежа на серверы одной из партий в период президентских выборов 2016 г. в США с описанием использования подобных приемов в этом конкретном случае (http://bit.ly/2iKStXO).

Многие открыто функционирующие и формально вполне законопослушные веб-сайты, а также поисковые системы тем не менее не брезгуют установкой на компьютеры пользователей «смягченных» версий шпионских программ (можно отнести их к рекламным).

1.3.16.3 ТРОЯНСКИЕ ПРОГРАММЫ

Название угрозы уходит корнями в античную историю. В «Илиаде» описано, как греки захватили Трою, преподнеся жителям тщетно осаждаемого ранее города-крепости «в дар» выполненное из дерева огромное изваяние коня, а сами удалились от стен города якобы в знак смирения перед его неприступностью. Оборонявшиеся наивно затащили коня в пределы города, а внутри, к их изумлению, оказалось боевое подразделение греков, оперативно овладевших городом.

В информационной безопасности «троянским конем» или просто трояном называется любая вредоносная программа, проникающая в систему под видом или в составе внешне законных программных средств. После установки «троянский конь» начнет удалять файлы, похищать данные, устанавливать и запускать вредоносные программы, изменять конфигурационные настройки компьютера, устанавливать программы-шпионы или даже превратит компьютер в «бота» (сокращение от «робот») или «зомби», то есть орудие и даже оружие в руках злоумышленников, используемое для дальнейшего распространения угроз по сети.

1.3.16.4 ВИРУСЫ

Вирус — встраивающаяся в выполняемый файл или уязвимое приложение программа, делающая использование инфицированного программного обеспечения чреватым негативными последствиями различной степени тяжести — от досадных сбоев до полного выхода из строя информационной системы. Вирус начинает действовать после открытия файла. То есть для выполнения ему требуется какая-то другая сопутствующая программа. Открытие загруженного из интернета файла с инфицированной программой связано с опасностью запуска вируса.

1.3.16.5 «ЧЕРВИ»

Компьютерными «*червями*» (*worms*) называют зловредные программы, имеющие свойство распространения по сети посредством самовоспроизведения и тиражирования. Инфицированный «червем» компьютер начинает безостановочно рассылать по сети потоки зараженных сообщений. Основное назначение «червей» — нанесение вреда сети путем расходования большого количества сетевых ресурсов (вплоть до полного отключения из-за перегрузки), однако они могут использоваться и для выполнения других опасных с точки зрения защиты данных действий.

1.3.16.6 ИСТОЧНИКИ ВРЕДОНОСНЫХ ПРОГРАММ

1.3.16.6.1 Системы мгновенного обмена сообщения (IM)

Системы мгновенного обмена сообщения (Instant Messaging, IM) позволяют пользователям передавать друг другу сообщения в режиме реального времени. При этом IM являют собой и новую угрозу

сетевой безопасности. По причине того, что многие IM-системы сильно запаздывают с добавлением возможностей по обеспечению безопасности, хакеры-злоумышленники приспособились распространять через них шпионские программы, фишинговые сообщения и всяческих «червей». Основную угрозу несут инфицированные вложения и активные ссылки в тексте сообщения.

1.3.16.6.2 Социальные сети

Сайты социальных сетей, таких как Facebook, Twitter, Vimeo, Google+, LinkedIn, Xanga, Instagram, Pinterest и MySpace, представляют собой онлайн-платформы, на которых пользователи создают собственные профили, рассказывают о себе, делятся информацией, мнениями и фотографиями, ведут блоги и тем самым становятся мишенью для всевозможных спамеров и похитителей персональных данных.

Помимо угрозы для держателей аккаунтов со стороны злоумышленников, все эти сайты несут угрозу для работодателей со стороны работников, поскольку в их сообщениях может содержаться немало полезной для конкурентов «инсайдерской» информации. Бывали случаи, когда утечки чувствительных корпоративных данных через соцсети приводили к обвалу биржевых котировок. Обязательно информируйте пользователей об этих опасностях, а также о том, что на самом деле всё, что они выкладывают в соцсети, остается в интернете навсегда. Даже если они спохватятся и удалят свой пост с чувствительными данными, утечка будет необратимой, поскольку множество людей успеют к моменту удаления эти данные скопировать. В некоторых компаниях предпочитают просто блокировать доступ сотрудников к соцсетям в настройках межсетевого экрана, но от риска несанкционированного разглашения данных посредством персональных устройств это не избавляет.

1.3.16.6.3 Спам

Спам, то есть массовые рекламные рассылки сообщений по электронной почте (как правило, на десятки миллионов адресов), в последние годы получил повсеместное распространение. В обозримом будущем прекращения спам-рассылок не предвидится, поскольку затраты окупаются сторицей: для сбыта товаров или услуг на миллионы долларов достаточно и 1% откликнувшихся на рекламу получателей. Большинство корпоративных почтовых серверов сегодня настроены на достаточно успешную фильтрацию потока входящих сообщений от спама с целью снижения непродуктивного трафика во внутренней сети. Основные фильтры исключают доставку электронных писем, которые:

- отправлены с доменных адресов, ранее замеченных в передаче спама;
- содержат подозрительно много получателей в полях «копия» или «скрытая копия»;
- содержат лишь картинку с гиперссылкой в теле письма;
- включают текстовые строки или слова, определенные в настройках спам-фильтра.

Ответив на спам или откликнувшись на содержащуюся в нем рекламу, пользователь тем самым подтверждает, что спам-рассылки по его адресу доходят до адресата и действенны, в результате

чего в будущем потоки спама в почтовом ящике такого пользователя могут возрасти многократно, особенно в свете того, что спамеры активно продают друг другу базы данных действующих адресов.

Спам-сообщения могут к тому же содержать мошенническую рекламу и вредоносные коды, например во вложениях, даже если имена и расширения прикрепленных файлов, равно как и текст сообщения, и картинки создают иллюзию рядового и вполне легитимного почтового отправления. Один из способов определить письмо как вредоносный спам — поочередно навести курсор на содержащиеся в нем гиперссылки и проверить в строке состояния или всплывающей подсказке, куда в действительности ведет ссылка: если не на сайт компании, значащейся в тексте, значит, это точно вредоносный спам. Другой признак: отсутствие функции (ссылки) «отказаться от подписки» (unsubscribe). В США в рекламных объявлениях, распространяемых по электронной почте, наличие такой ссылки для прекращения дальнейшего получения писем от отправителя обязательно.

2. ПРОВОДИМЫЕ РАБОТЫ

Не существует единого рецепта реализации функции обеспечения безопасности данных, который гарантировал бы соблюдение всех необходимых требований по защите информации о частной жизни и конфиденциальных данных. В целом все регламентирующие нормативно-правовые документы детально перечисляют данные, подлежащие защите, но не содержат никаких указаний, за счет чего и какими средствами эта защита должна реализовываться. Организациям надлежит самостоятельно разрабатывать и внедрять средства контроля безопасности данных, демонстрировать их соответствие требованиям законов или регламентов, вести документацию, подтверждающую реализацию всех требуемых мер по обеспечению безопасности, а также осуществлять мониторинг показателей их эффективности, включая динамику изменения показателей со временем. Как и в других областях знаний по управлению данными, направления работ по обеспечению безопасности данных включают выявление требований, оценку текущего положения дел на предмет наличия пробелов и рисков, внедрение инструментов и процессов, необходимых для защиты данных, и регулярное проведение аудита безопасности с целью оценки эффективности принимаемых мер.

2.1 Выявление требований по безопасности данных

Важно проводить различие между бизнес-требованиями, внешними ограничениями нормативно-правового характера и правилами, которые привносятся используемыми прикладными программными продуктами. С одной стороны, прикладные системы служат средствами обеспечения соблюдения бизнес-правил и процедур, а с другой стороны, далеко не редки случаи, когда в тех или иных системах заложены столь строгие механизмы защиты данных, что они значительно превосходят бизнес-требования, а иногда даже служат помехой для реализации бизнес-процессов. Соблюдение строгих требований по обеспечению защиты данных в наши дни становится,

по сути, нормой для готовых и поставляемых «под ключ» комплексных систем ведущих разработчиков. Однако и в этом случае необходимо убедиться, поддерживают ли они также принятые в организации стандарты ИБ.

2.1.1 Бизнес-требования

Планирование комплекса практических мер по защите данных организации начинается с тщательного изучения бизнес-требований. Потребности, миссия, стратегия и масштабы бизнеса, отраслевая принадлежность и размеры организации, — всё это играет немаловажную роль в определении степени жесткости требований по защите данных. Например, на финансовых и фондовых рынках США любая деятельность регулируется настолько жестко, что работающим на них организациям приходится придерживаться строжайших стандартов защиты данных. Противоположный пример: от предприятий мелкой розничной торговли, по сути, ничего не требуется в плане защиты данных; в то же время крупным торговым сетям приходится этим заниматься вплотную, хотя, казалось бы, формально и те и другие относятся к одной и той же отрасли.

Всесторонний анализ бизнес-правил и процессов поможет выявить основные чувствительные моменты в сфере защиты данных и соответствующим образом спланировать направления работы. На любом этапе потока работ могут возникать специфические требования по защите данных. Полезными инструментами могут стать матрица соотнесения данных с процессами (data-to-process) и матрица соотнесения данных с ролями (data-to-role). Они позволяют привязывать потребности в защите данных к процессам и ролям, а затем определять ролевые группы, параметры доступа и разрешения. Сочетайте краткосрочное оперативное планирование с долгосрочным стратегическим целеполаганием для достижения сбалансированности и эффективности функции обеспечения безопасности.

2.1.2 Нормативно-правовые требования

В современной быстро меняющейся в глобальных масштабах нормативно-правовой среде от организаций требуется соблюдение всё большего числа всевозможных законов и регламентов. Этические и юридические проблемы, с которыми сталкиваются организации в информационную эпоху, побуждают правительства принимать всё новые законы и стандарты, и чем дальше, тем больше внимания в них уделяется требованиям строгого контроля безопасности при управлении информацией (см. главу 2).

Создайте и ведите централизованный перечень или реестр всех применимых к вашей организации регламентирующих нормативно-правовых документов и предметных областей данных, на которые распространяются требования каждого из этих документов. По мере утверждения политик безопасности, принятых во исполнение этих требований, и механизмов контроля, реализованных в рамках этих политик, добавляйте в перечень ссылки на них (табл. 13). Требования надзорных органов и отраслевые регламенты, данные, на которые они распространяются, политики ИБ и меры по их осуществлению со временем неизбежно меняются, поэтому лучше вести такой перечень в легко редактируемом формате.

Таблица 13. Пример таблицы учета нормативно-правовых документов

Документ	Предметная область	Ссылки на политики безопасности	Механизмы контроля

Ниже приведены примеры законов, регламентирующих различные аспекты безопасности данных.

- США:
  - Закон Сарбейнса — Оксли (Sarbanes — Oxley Act of 2002);
  - Закон о применении медицинских информационных технологий в экономической деятельности и клинической практике, являющийся частью Закона о восстановлении американской экономики (Health Information Technology for Economic and Clinical Health (HITECH) Act, enacted as part of the American Recovery and Reinvestment Act of 2009);
  - Закон о преемственности и учете данных в медицинском страховании (Health Insurance Portability and Accountability Act of 1996, HIPAA);
  - Закон Грэмма — Лича — Блайли (Gramm — Leach — Bliley I and II);
  - Законы, принятые по инициативе Комиссии по ценным бумагам и биржам (Securities and Exchange Commission, SEC), и Закон о подотчетности в сфере корпоративной ИБ (Corporate Information Security Accountability Act);
  - «Патриотический акт» (Homeland Security Act and USA Patriot Act);
  - Федеральный закон об управлении информационной безопасностью (Federal Information Security Management Act, FISMA);
  - Закон штата Калифорния об обязательном информировании о случаях нарушения защиты данных (California: SB 1386, California Security Breach Information Act).
- ЕС:
  - Директива о защите данных (EU DPD 95/46/): раздел AB 1901, хищение электронных файлов или баз данных (Data Protection Directive (EU DPD 95/46/) AB 1901, Theft of electronic files or databases).
- Канада:
  - Билль 198[1].
- Австралия:
  - Закон о программе экономической реформы корпоративного права (CLERP Act).

Отраслевые нормативные документы, затрагивающие различные аспекты безопасности данных:

[1] Билль 198 (*англ.* Bill 198) — аналог Закона Сарбейнса — Оксли, принятый в 2003 г. в провинции Онтарио, но вступивший в силу лишь с 2006/07 финансового года. — *Примеч. пер.*

- Стандарт безопасности данных индустрии платежных карт (Payment Card Industry Data Security Standard, PCI DSS). Предусматривает подписание соглашения по соблюдению стандарта всеми организациями, имеющими отношение к работе с платежными картами.
- Европейский союз. Соглашение о достаточности капитала (Базель II) устанавливает правила информационного контроля для всех кредитно-финансовых учреждений стран-участниц.
- США. Стандарты защиты данных о потребителях Федеральной торговой комиссии (US Federal Trade Commission, FTC).

Нередко для обеспечения соблюдения корпоративных политик и/или нормативно-правовых требований и ограничений приходится вносить коррективы в бизнес-процессы. Пример: необходимость предоставления доступа к информации о состоянии здоровья граждан (регламентируемые данные) множественным уникальным группам пользователей в соответствии с Законом о преемственности и подотчетности медицинского страхования (HIPAA).

2.2 Определение политики безопасности данных

Организациям следует разработать политики безопасности данных, основанные на требованиях бизнеса и действующих нормативно-правовых документов. Политика — это заявление о выбранном курсе действий и высокоуровневое описание поведения, требуемого для скорейшего достижения поставленных целей. Политики безопасности данных описывают правила поведения, которые определены исходя из интересов организации, желающей защитить свои данные. Для того чтобы оценивать результаты реализации политик, они должны время от времени проверяться.

Корпоративные политики нередко имеют и юридический смысл. Например, суд может принять во внимание, что в организации было сделано всё возможное для приведения внутренних правил в соответствие с требованиями закона. Соответственно, несоблюдение корпоративной политики безопасности данных будет грозить потенциальному нарушителю серьезными правовыми последствиями.

Разработка политики безопасности требует совместного участия администраторов и архитекторов по безопасности ИТ, комитетов (советов) по руководству данными, распорядителей данных, команд по проведению внутреннего и внешнего аудита, представителей юридической службы. Распорядители данных вместе с сотрудниками, отвечающими за соблюдение требований регламентирующих документов (супервизоры, предусматриваемые законом Сарбейнса — Оксли, ответственные, предусматриваемые законом HIPAA, и т. п.), и бизнес-менеджерами, имеющими практический опыт работы с данными, должны вести подготовку метаданных для классификационных категорий информации, соответствующих различным видам нормативно-правовых требований, и контролировать правильность классификации. Все работы по обеспечению соблюдения внешних требований по защите данных должны быть согласованными. Это способствует снижению издержек, а также исключает возникновение сбоев из-за путаницы в инструкциях и тем более из-за бессмысленного дележа полномочий.

2.2.1 Содержание политики безопасности

Чтобы управлять поведением сотрудников с целью обеспечения корпоративной безопасности, в организации должны быть предусмотрены различные уровни политики. Например:

- **Корпоративная политика безопасности.** Включает общие правила доступа сотрудников к объектам инфраструктуры и иным ресурсам организации, а также стандарты и политики в отношении переписки по электронной почте, уровни доступа к защищенным данным в зависимости от должности или звания, политики подачи и рассмотрения жалоб на нарушения режима безопасности.
- **Политика безопасности ИТ.** Стандарты структуры каталогов, политики в отношении паролей, рамочная структура управления идентификацией пользователей и т. п.
- **Политика безопасности данных.** Категории данных, доступных каждому из используемых приложений, роли администраторов и операторов БД, группы пользователей и распределение информации по категориям чувствительности.

Часто политика безопасности ИТ и политика безопасности данных являются составными частями общей политики безопасности организации, однако такое решение нельзя назвать оптимальным. Предпочтительным представляется вариант четкого разделения на три взаимодополняющие политики. Дело в том, что политика безопасности данных по самой своей природе должна содержать значительно более детализированный и специфичный по содержанию свод правил обращения с данными в зависимости от их структуры и содержания, а также предусматривать конкретные механизмы и процедуры защиты каждой их категории. Политика безопасности данных рассматривается и утверждается Советом по руководству данными, а обеспечение ее выполнения и соблюдения поручается руководителю верхнего уровня, отвечающему за управление данными.

Сотрудники должны понимать политики безопасности и строго им следовать. Поэтому разрабатывайте каждую политику таким образом, чтобы все обязательные процедуры и причины их введения были четко определены, а цели их выполнения — достижимы. В целом политики должны быть такими, чтобы следовать им было проще, чем нарушать их. Кроме того, обеспечивая надежную безопасность, политики не должны сильно стеснять доступа к данным, требующимся сотрудникам для выполнения своей работы.

Политики безопасности данных должны быть оформлены в формате, доступном для чтения поставщиками, потребителями и иными заинтересованными лицами. Актуальные версии политик обязательно должны быть выложены в корпоративной интрасети или на портале совместного доступа.

Политики безопасности данных, а также процедуры и состав проводимых работ должны регулярно пересматриваться и обновляться с целью нахождения оптимального баланса требований по ИБ всех заинтересованных сторон.

2.3 Определение стандартов в области безопасности данных

Политики определяют общие правила поведения. Правда, не всякую нештатную ситуацию можно предусмотреть на уровне политик. Поэтому политики дополняются стандартами, детализирующими порядок действий в различных случаях таким образом, чтобы итоговый результат предписываемых действий максимально соответствовал намерениям, заявленным в политиках. Например, политика может предписывать использование надежных паролей; тогда определение понятия «надежный пароль» и требования к нему будут содержаться в стандартах, а соблюдение политики и стандартов будет поддерживаться технологическими средствами, обеспечивающими проверку соответствия создаваемого пароля стандартам.

2.3.1 Определение уровней конфиденциальности данных

Классификация данных по уровням конфиденциальности — важный аспект применения метаданных. Она лежит в основе предоставления пользователям привилегий доступа к данным. Каждая организация создает или заимствует схему классификации, отвечающую ее бизнес-требованиям. Любой метод классификации должен быть ясным и простым в применении. Должно быть предусмотрено достаточное количество уровней конфиденциальности информации, начиная с низкого и до самого высокого. Например, начиная с категории «Для всеобщего пользования» и заканчивая категорией «Строго конфиденциальная, под подписку о неразглашении» (см. раздел 1.3.12.1).

2.3.2 Определение категорий регламентирующих норм и правил

Рост числа резонансных случаев утечки весьма чувствительных конфиденциальных данных, в том числе и личного характера, привел к принятию множества законов, регламентирующих обращение с данными различных специфических категорий. Кроме того, участившиеся инциденты с хищением финансовых данных повлекли принятие по всему миру законов и нормативно-правовых актов, жестко регламентирующих защиту данных в банковских и коммерческих структурах.

В результате образовался целый новый класс данных, который можно назвать *регламентируемой информацией* (*regulated information*). Внешние нормативно-правовые требования можно считать расширением внутренних требований по информационной безопасности организаций. Они требуют принятия дополнительных к уже имеющимся мер по защите данных, необходимых для эффективного соблюдения требований регулирующих органов. При определении конкретных действий, которые следует предпринять организации в соответствии с теми или иными нормами и правилами, бывают полезны консультации с корпоративным юристом. Часто регламентирующие документы формулируют только цели или конечные результаты, а средства, используемые для их достижения, оставляются на усмотрение корпоративного руководства. В связи с этим рекомендуется предпринимать такие меры по защите данных, которые можно предъявить внешним проверяющим в качестве имеющего юридическую силу доказательства соблюдения нормативно-правового соответствия.

Полезный подход к организации работы с действующими нормами и правилами обращения с различными специфическими данными заключается в анализе всех нормативно-правовых документов и группировке схожих норм и правил по категориям, так же как это было проделано при классификации рисков, связанных с безопасностью данных.

Имея по всему миру сотню с лишним отличающихся друг от друга нормативно-правовых актов, регламентирующих порядок обращения с данными в одной и той же узкой области, бессмысленно создавать по отдельной категории данных для каждого из них. То же самое касается и большинства регламентов, за соблюдением которых следят всевозможные контролирующие и надзорные органы. По большому счету, многие регламенты преследуют одну и ту же конечную цель и могут быть отнесены к одной категории. Например, контрактные обязательства по защите конфиденциальных данных клиентов мало чем отличаются от требований американских, японских или канадских законов о защите персональных данных, равно как и от требований неприкосновенности информации частного характера, установленных в ЕС. Подобные параллели легко выявляются, если составить и сравнить перечни проверяемых мер по обеспечению соблюдения требований различных регламентирующих документов. Следовательно, наиболее эффективным подходом к управлению применением указанных мер будет такой, при котором они рассматриваются в рамках единой категории.

Ключевой принцип классификации как по уровням конфиденциальности, так и по категориям норм и правил заключается в том, что бóльшая часть информации может быть объединена в группы в зависимости от уровня чувствительности. Разработчикам необходимо знать, как такое объединение отражается на общих классификациях по уровням конфиденциальности и категориям регламентирующих документов. Например, разрабатывая информационную панель, отчет или представление базы данных, всегда следует обращать внимание, не присутствуют ли в составе требуемой для отображения информации какие-либо чувствительные данные, относящиеся к персональным, инсайдерским или обеспечивающим преимущество над конкурентами. При наличии подобных данных систему следует проектировать таким образом, чтобы права на доступ к ним были исключены из набора пользовательских прав или (если такой доступ всё-таки нужен) чтобы во время авторизации пользователя обеспечивался строгий учет всех требований по безопасности и требований регламентирующих норм и правил.

Результатом работы по классификации должны стать формально утвержденные перечни классов конфиденциальности и категорий регламентирующих норм и правил. Их должен дополнять порядок передачи этих метаданных в центральный репозиторий, для того чтобы сотрудники как бизнес-подразделений, так и технических служб знали, к какому уровню чувствительности относится информация, которую они обрабатывают, передают и к которой авторизуют доступ.

2.3.3 Определение ролей безопасности

В зависимости от потребностей контроль доступа к данным можно организовать на индивидуальном или групповом уровне. К слову, управление правами доступа и привилегиями на уровне индивидуальных учетных записей пользователей сопряжено с массой избыточной работы.

В небольших организациях с малым числом сотрудников это, возможно, и не препятствие, однако в крупных организациях более чем целесообразно применять подход к контролю доступа на основе ролей (role-based access control), предоставляя разрешения ролевым группам и, таким образом, каждому входящему в группу сотруднику.

Ролевые группы позволяют администраторам служб ИБ соотносить привилегии с ролями, а затем наделять этими привилегиями индивидуальных пользователей посредством включения их учетных записей в подходящую им по статусу и/или должности ролевую группу. Технически возможно включение одного и того же сотрудника в различные ролевые группы, но на практике этого следует избегать, чтобы не запутаться, какими привилегиями и полномочиями наделен тот или иной пользователь. По мере возможности старайтесь относить каждого пользователя только к одной ролевой группе. При этом может потребоваться создание различных пользовательских представлений каких-либо данных для разных ролевых групп, с тем чтобы пользователи каждой группы видели только те данные, которые им доступны согласно уровню привилегий и нормативно-правовым требованиям.

Обеспечение согласованности и непротиворечивости данных при управлении пользователями и ролями — задача не из простых. Данные о пользователе (ФИО, должность, ID сотрудника и т. п.) во многих случаях избыточно хранятся в нескольких местах. Как следствие, эти «острова данных» (islands of data) часто содержат противоречивую информацию об одном и том же физическом лице, представляя различные версии «правды» (versions of the «truth»). Во избежание проблемных ситуаций с целостностью данных обеспечьте централизованное управление идентификационными данными пользователей и их распределением по ролевым группам. Это обязательное требование обеспечения качества данных, без соблюдения которого эффективный контроль доступа невозможен. Администраторы безопасности отвечают за создание, изменение и удаление учетных записей пользователей и ролевых групп. Изменения классификации и привилегий групп, а также перевод пользователей из одной группы в другую требуют соответствующего согласования. Все изменения должны отслеживаться с помощью системы управления изменениями.

Непоследовательность или неадекватность мер по защите данных, применяемых в организации, могут повлечь недовольство сотрудников и чреваты значительным риском. Эффективность обеспечения безопасности на основе ролей зависит от того, насколько четко роли определены и насколько последовательно осуществляется их распределение.

Можно выделить два основных альтернативных подхода к определению и организации ролей — на основе решетки (начиная с данных) или на основе иерархии (начиная с пользователей).

2.3.3.1 РЕШЕТКА НАЗНАЧЕНИЯ РОЛЕЙ

Решетку (grid) полезно использовать для сопоставления различных ролей с данными, осуществляемого на основе уровней конфиденциальности, нормативно-правовых требований и функций пользователей. Роль «Пользователь с общим доступом» (Public User) позволяет работать со всеми нерегламентированными данными, предназначенными для всеобщего пользования (general audiences). Пользователям с ролью «Маркетинг» может быть открыт доступ к части персональной

идентификационной информации (PII), необходимой для планирования рекламных кампаний, но для них закрыты конфиденциальные данные о клиентах и информация ограниченного доступа. Таблица 14 содержит очень упрощенный пример такой решетки.

Таблица 14. Пример решетки назначения ролей

	Уровни конфиденциальности данных		
	Для всеобщего пользования	**Конфиденциальные данные о клиентах**	**Конфиденциальные данные ограниченного доступа**
Нерегламентированные	Роль «Пользователь с общим доступом»	Роль «Менеджер по работе с клиентами»	Роль «Ограниченный доступ»
Персональная идентификационная информация (PII)	Роль «Маркетинг»	Роль «Клиентский маркетинг»	Роль «Персонал» (HR)
Данные о держателях платежных карт (PCI)	Роль «Финансы»	Роль «Финансы клиента»	Роль «Финансы» (ограниченный доступ)

2.3.3.2 ИЕРАРХИЯ НАЗНАЧЕНИЯ РОЛЕЙ

Альтернативный вариант определения ролей — определение на уровне рабочих групп или бизнес-единиц с организацией ролей в виде иерархии таким образом, чтобы по мере снижения иерархического уровня объем привилегий дочерних ролей уменьшался за счет удаления части привилегий родительской роли. Постоянное поддержание подобных структур в актуальном состоянии — задача сложная, трудоемкая и требующая наличия систем отчетности, позволяющих отслеживать распределение привилегий доступа по всей иерархии, вплоть до наборов прав отдельных пользователей. Ниже представлен простейший пример иерархической модели назначения ролей (см. рис. 65).

2.3.4 Оценка текущих рисков безопасности данных

Риски безопасности включают элементы, которые могут скомпрометировать сеть и/или базу данных. Первый шаг по выявлению риска всегда заключается в определении мест хранения чувствительных данных и необходимых мер по их защите. Прежде всего оцените каждую систему по следующим параметрам:

- чувствительность хранящихся или передаваемых данных;
- требования по обеспечению защиты этих данных;
- имеющиеся средства защиты этих данных.

Задокументируйте результаты, поскольку они образуют базовый уровень (baseline) для последующих оценок. К тому же ведение подобной документации может быть и обязательным

требованием обеспечения нормативно-правового соответствия, как это, например, предусмотрено директивами ЕС. Выявленные пробелы (gaps) в обеспечении необходимого уровня безопасности подлежат устранению посредством совершенствования процессов, поддерживаемых технологиями. Результативность усовершенствований должна подтверждаться улучшением объективных показателей мониторинга защищенности данных от известных рисков.

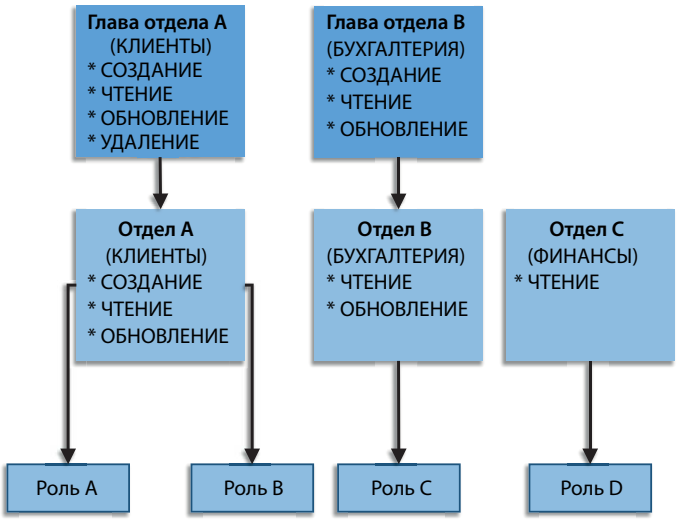


Рисунок 65. Пример иерархии назначения ролей

В крупных организациях имеет смысл периодически прибегать к услугам «белых» хакеров, способных всесторонне проверить системы на уязвимость. Между прочим, подтверждение устойчивости системы ИБ организации к проникновениям, полученное от специалистов, — это еще и отменная реклама, а также средство поднятия репутации на рынке.

2.3.5 Внедрение механизмов контроля и процедур

За внедрение политики безопасности и управление выполнением ее требований прежде всего отвечает администратор безопасности, работающий в координации с распорядителями данных и командой технической поддержки. Например, безопасность баз данных часто входит в сферу ответственности администраторов БД.

Для обеспечения выполнения требований политики безопасности организации необходимо внедрить надлежащие механизмы (элементы) контроля (controls)[1]. Механизмы (элементы) контроля и процедуры должны (как минимум) охватывать:

[1] В ГОСТ Р ИСО/МЭК 27002-2012 «Информационная технология (ИТ). Методы и средства обеспечения безопасности. Свод норм и правил менеджмента информационной безопасности» термин «control» применительно к тематике стандарта переведен как «мера и средство контроля и управления». В данном издании применительно к этой же тематике он для краткости переводится как «механизм контроля» или «элемент контроля» (в зависимости от контекста). — *Примеч. науч. ред.*

- порядок получения и лишения разрешений на доступ пользователей к системам и/или приложениям;
- порядок назначения и снятия пользовательских ролей;
- порядок мониторинга уровней привилегий;
- порядок обработки запросов на изменение прав доступа;
- порядок классификации данных по уровням конфиденциальности и категориям регламентирующих норм и правил;
- порядок выявления нарушений безопасности и порядок действий в случае их выявления.

Документируйте требования, соблюдение которых обязательно для предоставления пользователю прав доступа, с тем чтобы своевременно отменить авторизацию, как только пользователь утратит формальные основания для доступа.

Например, политика «Осуществлять надлежащее ведение привилегий пользователей» может иметь контрольную цель: «Права и привилегии администратора БД и пользователей подлежат ежемесячному пересмотру и подтверждению». Процедура, отвечающая требованиям этого элемента контроля, должна предусматривать внедрение (и сопровождение) в организации процессов, обеспечивающих:

- подтверждение выданных разрешений посредством их сверки с данными системы управления изменениями;
- реализацию автоматизированного потока работ (workflow) по согласованию или процедуры подписания заявки в бумажной форме с целью документирования каждого запроса на изменение;
- реализацию процедуры отмены авторизации лиц, утративших основания для доступа к данным в силу перевода на другую должность или изменения статуса их должности или отдела.

На каком-либо из уровней управления должны быть реализованы формальные процедуры сбора, рассмотрения и утверждения заявок на первоначальную авторизацию и последующих заявок на изменение прав и привилегий пользователей и групп пользователей.

2.3.5.1 НАЗНАЧЕНИЕ УРОВНЕЙ КОНФИДЕНЦИАЛЬНОСТИ

Распорядители данных отвечают за определение уровней конфиденциальности данных в соответствии с классификационной схемой, утвержденной в организации.

При назначении уровня конфиденциальности документов и отчетов должен выбираться наиболее высокий уровень из тех, которые назначены отдельным содержащимся в них элементам данных (см. главу 9). В верхнем или нижнем колонтитуле каждой страницы или экрана просмотра должна присутствовать метка уровня конфиденциальности. Информационные продукты с самым низким уровнем (например, «Для всеобщего пользования») в подобной маркировке не нуждаются. По умолчанию считается, что любой информационный продукт, не содержащий метки уровня конфиденциальности, предназначен для всеобщего пользования.

Авторы документов и разработчики информационных продуктов обязаны оценивать и корректно назначать уровень конфиденциальности, маркируя соответствующей меткой каждый документ, а также каждую базу данных, включая таблицы, столбцы и пользовательские представления реляционных баз данных.

В крупных организациях бóльшую часть работы по классификации в части безопасности и по обеспечению защиты документов и информационных продуктов обычно выполняют специальные подразделения, отвечающие за ИБ. Возможно, сотрудники этих подразделений будут рады делегировать работу по классификации распорядителям данных, но ответственность за обеспечение соблюдения правомерности доступа к данным, а также за физическую защиту сетей они обычно берут на себя.

2.3.5.2 НАЗНАЧЕНИЕ КАТЕГОРИЙ РЕГЛАМЕНТИРУЮЩИХ НОРМ И ПРАВИЛ

Для обеспечения соблюдения требований нормативно-правового соответствия, связанных с ИБ, организациям следует разработать подход к классификации регламентируемых данных (или позаимствовать уже кем-то применяемый) (см. раздел 3.3). Схема классификации предоставляет основу для успешного прохождения организацией процедур внутреннего и внешнего аудита. После ее утверждения вся информация должна оцениваться и классифицироваться строго в рамках этой схемы. Сотрудники подразделений ИБ могут слабо разбираться в концепции такой классификации, поскольку они в основном работают не с нормами и правилами, регламентирующими обращение с отдельными классами данных, а с инфраструктурными системами. Поэтому им потребуются документированные требования по защите данных для каждой регламентированной категории, включая определение мер, которые они могут реализовать.

2.3.5.3 УПРАВЛЕНИЕ И ПОДДЕРЖКА БЕЗОПАСНОСТИ ДАННЫХ

После того как будут определены все требования, а также внедрены политики и процедуры, главной задачей становится обеспечение их соблюдения, включая оперативное выявление и устранение нарушений. Непрерывный мониторинг систем и проведение аудита выполнения процедур безопасности — критически важный аспект обеспечения безопасности данных.

2.3.5.3.1 Контроль доступности данных / Датацентричная безопасность

Контроль доступности данных требует управления наборами прав пользователей, а также средствами (например, маскировкой, созданием пользовательских представлений и т. п.), которые технически реализуют предоставление доступа на основе наборов прав. В некоторых СУБД подобные защитные средства и процессы представлены более широко и реализованы более эффективно, чем в других (см. раздел 3.7).

Менеджеры, обеспечивающие выполнение требований по ИБ, могут непосредственно отвечать за разработку профилей наборов прав пользователей, способствующих бесперебойному ведению бизнеса с одновременным соблюдением всех действующих ограничений.

Определение наборов прав и авторизация пользователей требуют детального учета имеющихся данных, тщательного анализа информационных потребностей, а также документирования данных, которые становятся доступными пользователю в результате применения каждого набора прав. Часто информация с высокой степенью чувствительности смешивается с нечувствительной. Во избежание подобных случаев важно иметь корпоративную модель данных, позволяющую идентифицировать и отслеживать местонахождение чувствительных данных (см. раздел 1.1.1).

Маскировка данных позволяет защитить их даже в случае их непреднамеренного раскрытия. Некоторыми регламентирующими правилами предписывается выдача данных пользователям исключительно в зашифрованном виде, что можно рассматривать как предельный случай маскировки по месту хранения. В таком случае пользователям для ознакомления с данными потребуются ключи дешифрования, а получить их без надлежащей авторизации будет невозможно, что является вполне надежной защитой от несанкционированного ознакомления. Пользователи с правом доступа к ключам дешифрования увидят расшифрованные данные, а остальные — бессмысленный набор символов.

В случае реляционных баз данных можно создавать различные представления для выдачи различным категориям пользователей с целью обеспечения необходимого уровня защиты данных для каждой категории. Из представлений могут исключаться, например, строки, содержащие определенные значения, не подлежащие просмотру, или столбцы с конфиденциальными или регламентируемыми данными.

2.3.5.3.2 Мониторинг аутентификации и поведения пользователей

Предоставление отчетов о выполненных входах в систему и запросах данных — базовое условие, без которого невозможен аудит соблюдения установленных требований и правил. Мониторинг аутентификации и поведения пользователей в отношении доступа (access behavior) позволяет отслеживать, кто подключается и получает доступ к информационным активам. Также мониторинг помогает выявлять необычные, непредвиденные или подозрительные транзакции, требующие расследования. Таким образом, он компенсирует пробелы в планировании, проектировании и реализации средств обеспечения безопасности данных.

Решения о том, что именно нуждается в мониторинге, на протяжении какого срока и какие действия требуются в случае наступления тревожного события, следует принимать по результатам тщательного анализа бизнес-требований и нормативно-правовых ограничений. Мониторинг связан с широким спектром разнообразных действий. Он может проводиться специально в отношении конкретных наборов данных, пользователей или ролей. Мониторинг можно использовать в целях проверки целостности данных, конфигураций или ключевых метаданных. Он может быть реализован в рамках одной информационной системы, а также охватывать взаимосвязанные разнородные систем. Наконец, возможен избирательный мониторинг конкретных привилегий, таких как возможность загружать большие массивы данных или получать доступ к системе в нерабочее время.

Мониторинг может вестись в автоматическом, ручном и полуавтоматическом режимах. Автоматизированный мониторинг создает дополнительную нагрузку на отслеживаемые системы и может снижать их производительность. Иногда полезно проводить периодические моментальные снимки состояния обрабатываемых запросов и запущенных процессов: они позволяют выявлять тенденции и сравнивать их со стандартными критериями. Также может потребоваться итерационная настройка конфигурации для обеспечения оптимальных параметров системы с точки зрения потребностей мониторинга.

Автоматическая регистрация транзакций с участием чувствительных данных и необычных запросов к базам данных должна быть реализована при развертывании любой базы данных. Отсутствие автоматизированного мониторинга подобных действий чревато серьезными рисками, включая следующие.

- **Риски, связанные с нарушением нормативно-правовых требований.** Слабость механизмов аудита баз данных всё чаще ставит организации в ситуации, когда они невольно оказываются не в ладах с законом. Закон Сарбейнса — Оксли (SOX), действующий в секторе финансовых услуг, и Закон о преемственности и подотчетности медицинского страхования (HIPAA) — всего лишь два примера американских нормативных правовых актов с четко определенными требованиями к механизмам аудита баз данных.
- **Риски, связанные с недостаточностью мер и средств обнаружения вторжений и восстановления данных.** Механизм аудита данных — последний рубеж их защиты. Если взломщик обойдет все защитные средства, аудит позволит выявить это, пусть и постфактум. Данные аудита также могут использоваться для установления пользователя, имеющего отношение к произведенной атаке, и определения действий, необходимых для восстановления работоспособности системы.
- **Риск неправомерного использования администраторских и аудиторских полномочий.** Злоумышленник, получивший доступ к серверу базы данных с полномочиями администратора, — вне зависимости от того, были они получены законно или путем взлома, — может отключить функции мониторинга/аудита данных с целью сокрытия мошеннических действий. Поэтому в идеале лучше не допускать совмещения функций аудиторов и администраторов БД или специалистов по поддержке серверных платформ.
- **Риск, обусловленный излишним доверием к встроенным средствам аудита.** Программные платформы поддержки баз данных часто включают базовые средства аудита, но, как правило, они имеют много недостатков, препятствующих их развертыванию в реальных эксплуатационных средах. В тех случаях, когда доступ к базам данных осуществляется с помощью веб-приложений (например, SAP, Oracle E-Business Suite или PeopleSoft), встроенные в эти приложения механизмы аудита не имеют возможности идентифицировать отдельных пользователей и связывают все действия с именем учетной записи веб-приложения. Как следствие, когда с помощью внутренних журналов аудита будут обнаружены случаи мошенничества, отследить по их записям реальных виновников не удастся.

Для снижения рисков рекомендуется внедрить какое-либо ориентированное на сетевые возможности средство аудита, которое устраняет недостатки встроенных инструментов. При этом его использование не освобождает от необходимости регулярного проведения проверок подготовленными аудиторами. Применение сетевых комплексов подобного рода дает следующие преимущества.

- **Высокая производительность.** Сетевые реализации средств аудита могут работать со скоростью передачи данных по сети и почти не влияют на производительность базы данных.
- **Разделение обязанностей.** Сетевые средства аудита работают независимо от администраторов баз данных, что позволяет отделять функции администрирования от функций аудита.
- **Детализированное отслеживание транзакций** способствует оперативному выявлению мошеннических действий, проведению их экспертизы и восстановлению данных. Журналы позволяют выяснять такие детали, как имя приложения-источника, полный текст запроса, атрибуты ответа на запрос, ОС источника, время запроса и системное имя его автора.

2.3.5.4 УПРАВЛЕНИЕ СООТВЕТСТВИЕМ ПОЛИТИКЕ БЕЗОПАСНОСТИ

Управление соответствием политике безопасности включает комплекс текущих мероприятий, обеспечивающих гарантии того, что частные политики соблюдаются и механизмы контроля эффективно поддерживаются. Предполагается также выработка рекомендаций по соблюдению новых требований. Во многих случаях распорядители данных должны действовать совместно с подразделениями ИБ и корпоративным юристом в целях обеспечения согласования операционных политик и технических механизмов контроля.

2.3.5.4.1 Управление нормативно-правовым соответствием

Управление нормативно-правовым соответствием включает:

- оценку соответствия стандартам и процедурам авторизации;
- обеспечение измеримости результатов выполнения требований по работе с данными и, как следствие, возможности проверки (то есть требования наподобие «уделять внимание» не годятся ввиду их неизмеримости);
- обеспечение защиты регламентированных данных в местах хранения и в процессе передачи с использованием стандартных инструментов и процессов;
- использование процедур эскалации и механизмов уведомления в случаях выявления проблемных вопросов в отношении нормативно-правового соответствия или его нарушений.

Механизмы контроля соответствия требуют наличия контрольных журналов (audit trails). Например, если политика устанавливает, что пользователи получают доступ к данным определенной категории после проведения обязательного инструктажа, организация должна иметь возможность подтвердить прохождение инструктажа каждым пользователем, работающим с этими

данными. Без контрольного журнала доказать, было ли соблюдено то или иное требование, невозможно. Все механизмы контроля должны разрабатываться таким образом, чтобы они были проверяемыми.

2.3.5.4.2 Аудит безопасности данных и нормативно-правового соответствия

Внутренние аудиты деятельности организации с целью подтверждения соблюдения политик безопасности данных и нормативно-правового соответствия должны проводиться регулярно и последовательно. Механизмы контроля соответствия сами по себе должны всякий раз заново проверяться и пересматриваться не только в случае принятия любого нового нормативно-правового акта в области работы с данными или внесения изменений в действующие, но и просто с установленной периодичностью, чтобы удостовериться в их пригодности. Проверки могут проводить как внутренние, так и внешние аудиторы. В любом случае аудиторы должны быть полностью независимыми от проверяемых данных и/или процессов во избежание конфликта интересов при проведении мероприятий, а также необъективности и неполноты оценок и заключений по результатам проверки.

Аудиторская проверка — не карательная экспедиция с целью выявления нарушений и наказания виновных. Цель аудита — предоставить руководству организации и Совету по распоряжению данными объективные результаты непредвзятой оценки и рационально обоснованные, практически реализуемые рекомендации.

Положения политики безопасности данных, стандарты, документы, руководства по внедрению, запросы на изменения, журналы мониторинга доступа, сформированные отчеты и иные записи (электронные или твердые копии) образуют входные материалы аудита. Помимо обследования вещественных и документальных свидетельств проведения деятельности аудит часто включает проведение оценок и проверок, таких как:

- анализ политики и стандартов на предмет точности и ясности определения механизмов контроля, а также полноты учета нормативно-правовых требований;
- анализ внедренных процедур и практик авторизации доступа пользователей на предмет соответствия целям, политикам, стандартам и требованиям к результатам, предусмотренным регулирующими нормами и правилами;
- оценка стандартов и процедур авторизации с точки зрения достаточности и соответствия технологическим требованиям;
- оценка эффективности выполнения процедур эскалации и механизмов уведомления в случаях выявления проблемных вопросов в отношении нормативно-правового соответствия или его нарушений;
- проверка контрактов, соглашений о совместном использовании данных и обязательств подрядчиков и поставщиков в отношении соблюдения нормативно-правового соответствия на предмет соблюдения организацией и ее бизнес-партнерами своих обязательств по защите регламентированных данных;

- оценка зрелости действующих в организации практик обеспечения безопасности и подготовка для высшего руководства и других заинтересованных сторон отчета «Состояние обеспечения нормативно-правового соответствия» (State of Regulatory Compliance);
- подготовка рекомендаций по внесению изменений в политику нормативно-правового соответствия и по усовершенствованию операционной деятельности по обеспечению соответствия.

Аудит безопасности данных не заменяет управления безопасностью данных. Это вспомогательный процесс, в ходе которого оценивается, соответствует ли деятельность по управлению поставленным целям.

3. ИНСТРУМЕНТЫ

Инструменты, используемые для управления, в значительной мере зависят от размера организации, сетевой архитектуры, а также политик и стандартов, применяемых ее подразделениями безопасности.

3.1 Антивирусное программное обеспечение

Антивирусные программы защищают компьютеры от вирусов, часто встречающихся в интернете. Поскольку новые вирусы и другие вредоносные программы появляются ежедневно, важно следить за регулярным обновлением средств антивирусной защиты.

3.2 Протокол HTTPS

Если URL-адрес начинается с https://, это указывает на защиту обмена данными с веб-сайтом посредством шифрования. Обычно доступ к веб-сайтам защищен паролем или иными средствами аутентификации пользователей. Доступ к онлайновым банковским и платежным системам, а также к конфиденциальной информации осуществляется только через соединения, защищенные шифрованием. Поэтому необходимо учить пользователей обязательно проверять наличие префикса https:// перед URL-адресом, прежде чем совершать какие-либо чувствительные операции посредством интернета, да и в корпоративной сети тоже. Без шифрования посторонние лица в том же сегменте сети вполне могут прочитать любую передаваемую текстовую информацию.

3.3 Технологии управления идентификацией

Технологии управления идентификацией обеспечивают хранение реквизитов пользователей и их выдачу по запросам систем, например при попытке в них войти. Некоторые приложения ведут собственные репозитории реквизитов пользователей, однако самим пользователям намного удобнее входить в систему единожды, после чего большинство приложений будет определять их права, обращаясь к централизованному репозиторию реквизитов. Имеются специальные

протоколы управления реквизитами пользователей, например LDAP (Lightweight Directory Access Protocol — облегченный протокол доступа к каталогам).

В некоторых компаниях используют принятую на уровне корпорации программу — менеджер паролей Password Safe, которая создает на каждом компьютере зашифрованный файл с паролями. Каждому пользователю нужно помнить лишь одну длинную идентификационную фразу (pass-phrase) для того, чтобы войти в менеджер паролей, и они могут хранить все прочие пароли в зашифрованном файле на своем компьютере. Такую же роль могут выполнять системы идентификации, использующие технологию единого входа (single-sign-on).

3.4 Системы обнаружения и предотвращения вторжений

Инструменты, которые позволяют обнаруживать вторжения и оперативно отказывать в доступе, необходимы в тех случаях, когда злоумышленник всё-таки проник через сетевой экран или другие средства защиты.

Система обнаружения вторжений (IDS) незамедлительно уведомляет ответственных сотрудников о любом выявленном случае несанкционированного доступа. Оптимальным является решение, при котором IDS дополняет система предотвращения вторжений (IPS), автоматически реагирующая как на известные атаки, так и на нелогичные сочетания команд, поступающих от пользователей. В связи с этим полезно анализировать типовые образцы команд и запросов данных в рамках организации. Знание характерных структур обращений к системам позволяет выявлять среди них необычные, то есть потенциально опасные, при обнаружении которых защита сможет отправлять предупреждения ответственным за информационную безопасность.

3.5 Межсетевые экраны

На входе во внутреннюю среду информационных систем предприятия должны устанавливаться сложные и безопасные межсетевые экраны (firewalls), способные осуществлять высокоскоростную передачу данных и одновременно проводить детализированный анализ сетевых пакетов. В случае веб-серверов, доступных в интернете, рекомендуется создавать более сложную структуру защитного межсетевого экрана, поскольку множество злонамеренных хакерских атак маскируются под внешне легитимные обращения к северам баз данных, в то время как в реальности призваны нащупать уязвимости в их защите.

3.6 Отслеживание метаданных

Инструменты отслеживания метаданных позволяют организации осуществлять мониторинг перемещения чувствительных данных. Использование подобных средств, однако, чревато риском выявления внешними шпионскими программами внутренних данных организации по метаданным, связанным с документами. В целом же использование метаданных для идентификации чувствительной информации считается оптимальным способом обеспечения сохранности и надлежащей защиты данных. Подавляющий процент утерь и утечек чувствительной информации связан отнюдь не с наличием соответствующих метаданных, говорящих о степени ее

чувствительности, а, наоборот, с их отсутствием, которое приводит к пренебрежению правилами безопасного хранения и защиты. Поэтому надлежащее ведение и хранение метаданных следует считать абсолютно приоритетным направлением работы, важность которого перекрывает гипотетический риск взлома репозитория метаданных. Этот риск можно считать незначительным, поскольку опытному хакеру на порядок проще отыскать в сети и взломать незащищенное хранилище собственно чувствительной информации, а не разбираться с ее метаданными. К сожалению, чаще всего непонимание необходимости защиты чувствительных данных свойственно именно тем сотрудникам, которые непосредственно с ними работают.

3.7 Маскировка / Шифрование данных

Инструменты маскировки или шифрования полезно использовать для наложения ограничений на представление чувствительных данных при их перемещении (см. раздел 1.3.9).

4. МЕТОДЫ

Методы, применяемые при управлении информационной безопасностью, зависят от размера организации, сетевой архитектуры, типа подлежащих защите данных, политик и стандартов, установленных подразделениями безопасности.

4.1 Использование CRUD-матриц

Создание и использование матриц отношений «данные — процессы» (data-to-process) и «данные — роли» (data-to-role), — так называемых матриц CRUD: от Create (создание), Read (чтение), Update (обновление), Delete (удаление), — помогает отобразить потребности в доступе к данным и определить ролевые группы, параметры и разрешения. Иногда к матрице добавляют пятый параметр — E (Execute — выполнение) — и называют ее CRUDE-матрицей.

4.2 Немедленное развертывание обновлений безопасности

В организации должен быть организован процесс установки обновлений безопасности на всех компьютерах с максимальной оперативностью. Ведь хакеру-злоумышленнику достаточно получить доступ с полномочиями суперпользователя к одному-единственному компьютеру для проведения успешной атаки на сеть. Пользователи не должны иметь возможности задерживать установку обновлений.

4.3 Атрибуты безопасности в метаданных

Репозиторий метаданных — ключевое средство обеспечения целостности и согласованного использования корпоративной модели данных всеми бизнес-процессами организации. Метаданные должны включать классы конфиденциальности данных и категории регламентирующих норм и правил (см. раздел 1.1.3). Включение в состав метаданных атрибутов, относящихся

к безопасности, защищает организацию от риска раскрытия чувствительных данных сотрудниками, которые с ними работают, но могут не знать о том, что эти данные имеют определенную степень чувствительности. Назначая данным уровни конфиденциальности и категории регламентирующих норм и правил, распорядители данных должны следить за тем, чтобы информация о назначении была зарегистрирована в репозитории метаданных. Если позволяет используемая технология, данные должны быть промаркированы с помощью соответствующих меток (см. разделы 3.3.1 и 3.3.2). Сведения об уровнях конфиденциальности и категориях регламентирующих норм и правил можно использовать для определения и управления наборами пользовательских прав, а также для управления авторизацией. Кроме того, эти сведения нужны для информирования команд разработчиков о рисках, связанных с чувствительными данными.

4.4 Метрики

Для подтверждения того, что все процессы обеспечения информационной безопасности функционируют согласно установленным требованиям, нужен набор измеримых параметров (метрик), показывающих их эффективность. Метрики также дают возможность совершенствовать эти процессы. Некоторые метрики отражают прогресс в выполнении каких-либо работ: например, количество проведенных аудиторских проверок, установленных систем безопасности, выявленных инцидентов или объем непроверенных данных. Более сложные метрики сфокусированы на оценке недостатков, выявленных в ходе аудиторских проверок, или уровня зрелости организации в рамках модели зрелости.

В крупных организациях, имеющих службы ИБ, значительное число подобных метрик так или иначе уже введено. Полезно переосмыслить место существующих показателей и рассматривать их в контексте общего комплекса мероприятий по измерению параметров процесса управления угрозами, в частности для того, чтобы избежать дублирования усилий. Определившись с метриками, зафиксируйте для каждой из них исходное значение (базовый уровень — baseline), чтобы затем отслеживать прогресс с течением времени.

Поскольку измерять и отслеживать можно множество условий, требующихся для обеспечения ИБ, фокусируйте внимание на тех, которые имеют практическое значение. Немногочисленные ключевые метрики, сгруппированные по отдельным направлениям, значительно проще поддаются пониманию и управлению, чем не связанные друг с другом показатели. Мероприятия по совершенствованию процессов обеспечения ИБ могут включать тренинги, предполагающие информирование о политиках регулирования в отношении данных и о действиях по соблюдению требований ИБ.

Многие организации сталкиваются со схожими проблемами в области безопасности данных. В выборе подходящих метрик может помочь следующий перечень.

4.4.1 Метрики для оценки эффективности внедрения мер безопасности

Следующие общие метрики безопасности представляют собой показатели, исчисляемые в процентах.

- Доля компьютеров организации, на которых установлены все самые последние обновления безопасности.
- Доля компьютеров, на которых установлены и функционируют самые последние версии антивирусного и другого ПО защиты от вредоносных программ.
- Доля новых сотрудников, успешно прошедших проверку на благонадежность.
- Доля сотрудников, прошедших ежегодное тестирование в области практик ИБ с количеством правильных ответов более 80%.
- Доля бизнес-единиц, где проведена формальная оценка рисков в сфере ИБ.
- Доля бизнес-процессов, успешно протестированных на предмет аварийного восстановления данных в случае пожара, землетрясения, урагана, наводнения, взрыва или иной чрезвычайной ситуации.
- Доля устраненных недостатков, выявленных в ходе аудиторской проверки.

Для выявления и отслеживания тенденций можно использовать статистические показатели или показатели, выражаемые в виде перечней.

- Метрики производительности всех систем ИБ.
- Проведенные проверки анкетных данных сотрудников и их результаты.
- Статусы планов действий в непредвиденных ситуациях и обеспечения непрерывности бизнеса.
- Случаи инцидентов криминального характера и результаты их расследования.
- Проверки обеспечения «должной осмотрительности» (Due Diligence) в отношении соблюдения требований ИБ и количество выявленных нарушений, которые необходимо устранить.
- Проведенные мероприятия по аналитическому исследованию вопросов управления рисками (с указанием количества мероприятий, повлекших действенные изменения).
- Проверки соблюдения политики ИБ и их результаты (например, проверки соблюдения политики «чистого стола» (clean desk) при вечерних обходах помещений сотрудниками службы безопасности).
- Статистика спецопераций, мероприятий по обеспечению физической безопасности и безопасности помещений.
- Количество задокументированных и доступных для ознакомления действующих стандартов ИБ (они же «политики» — policies).
- Мотивированность различных сторон к соблюдению политик безопасности также может быть оценена.
- Анализ сложившейся практики ведения бизнеса в контексте репутационных рисков, включая проведение тренингов.
- Профилактика нечистоплотности в бизнесе и рисков инсайдерских угроз в отношении использования специфических категорий информации (финансовая, медицинская, относящаяся к коммерческой тайне, инсайдерская и т. п.).

- Показатели, отражающие доверие к службе информационной безопасности со стороны менеджеров и сотрудников, а также ее влияние в организации (служат косвенными индикаторами того, как воспринимаются усилия по обеспечению ИБ и политики в этой области).

Выберите из перечисленных метрик разумное количество наиболее действенных, относящихся к различным категориям, и осуществляйте их регулярное отслеживание с целью соблюдения существующих требований, выявления проблем до их перехода в критическую фазу и демонстрации вышестоящему руководству четкого стремления обеспечить защиту ценной корпоративной информации.

4.4.2 Метрики для оценки осведомленности сотрудников в области безопасности

Для выбора наиболее подходящих метрик следует рассмотреть следующие общие категории.

- **Результаты оценки рисков** позволяют получить обоснованные и заслуживающие доверия данные, которые необходимо передать руководству и сотрудникам соответствующих бизнес-единиц, с тем чтобы они были лучше осведомлены о своей ответственности.
- **Рисковые события и профили рисков**: выявляйте неконтролируемые области, подвергаемые рискам и требующие принятия мер по защите. С помощью последующих проверок инициатив в отношении информирования сотрудников определяйте степень (или отсутствие) продвижения в части снижения подверженности рискам или соблюдения политики безопасности. Это позволяет определить, насколько эффективны мероприятия по информированию.
- **Формальные обследования или интервью с целью получения откликов** позволяют определить общий уровень информирования в области безопасности. Кроме того, следует оценивать количество сотрудников из состава целевых групп, которые прошли тренинги, направленные на информирование в области ИБ.
- **Расследование инцидентов, учет извлеченных уроков и интервью со сторонами, понесшими ущерб**, являются хорошим источником информации о пробелах в информировании в области безопасности. Для измерения могут использоваться показатели того, насколько снизилась уязвимость.
- **Аудиторские проверки эффективности обновлений безопасности** проводятся на отдельных компьютерах, имеющих доступ к конфиденциальным и регламентированным данным. (Рекомендуется использовать систему автоматического обновления на всех компьютерах, где это возможно.)

4.4.3 Метрики для оценки защиты данных

Выберите актуальные для вашей организации показатели, исходя из действующих требований.

- **Классификация критичности** отдельных видов данных и информационных систем, выход из строя которых окажет существенное влияния на деятельность организации.

- **Оценочная величина среднегодовых потерь** от компрометации, повреждения или утраты данных вследствие чрезвычайных происшествий, взломов, хищений или аварий.
- **Риск утери специфических данных**, относящихся к определенной категории регламентируемой информации, и классификация приоритетности устранения последствий.
- **Соотнесение рисков с конкретными бизнес-процессами**: например, риск, связанный с хищением данных при помощи платежных терминалов, можно включить в профиль рисков, присущих системе проведения платежных операций.
- **Оценки угроз** проводятся на основе оценки вероятности атак на определенные информационные ресурсы, содержащие ценные данные, а также на средства, по которым эти данные передаются.
- **Оценки уязвимости** отдельных частей бизнес-процессов, в которых чувствительная информация может быть раскрыта случайно или преднамеренно.

Результатом должен стать доступный для аудиторской проверки перечень мест организации, через которые осуществляется распространение чувствительных данных.

4.4.4 Метрики для оценки инцидентов безопасности

- Выявленные случаи обнаружения и предотвращения вторжения.
- Показатель возврата инвестиций (Return on Investment) для затрат на безопасность, определяемый на основе оценки денежных средств, сохраненных в результате предотвращения вторжений.

4.4.5 Предотвращение необоснованного распространения конфиденциальных данных

Количество копий конфиденциальных данных должно отслеживаться во избежание их необоснованного распространения. Чем больше мест, где хранятся одни и те же конфиденциальные данные, тем выше риск их утечки.

4.5 Учет потребностей в безопасности данных в проектных требованиях

Каждый проект, предусматривающий использование данных, должен учитывать требования по информационной безопасности. Детальные требования по безопасности данных и приложений выявляются на стадии анализа. Чем раньше они будут сформулированы, тем более целенаправленно будет осуществляться проектирование, что позволит избежать внесения изменений в процессы обеспечения безопасности. Если команды по реализации с самого начала понимают требования по защите, они могут встроить механизмы обеспечения их соблюдения в базовую архитектуру системы. Эта информация также может быть использована при выборе программного обеспечения сторонних разработчиков.

4.6 Эффективный поиск в массиве зашифрованных данных

Поиск в массиве зашифрованных данных предполагает их дешифрование. Единственный способ, позволяющий сократить количество данных, требующих дешифрования, заключается

в шифровании критерия поиска (например, строки) с использованием того же метода, который использовался для шифрования данных в массиве, и осуществлении поиска совпадений с зашифрованным критерием. В результате потребуется дешифрование лишь совпадающих данных, а это менее затратно (и рискованно). После этого можно обычным текстовым поиском отобрать из дешифрованных результатов точные совпадения.

4.7 Санитизация документов

Санитизация документов (document sanitization) — это процесс удаления из документов метаданных (например, таких, как зафиксированная история изменений) перед предоставлением к ним общего доступа. Санитизация снижает риск случайной утечки конфиденциальной информации, которая может содержаться в записях метаданных. При заключении контрактов доступ контрагентов к подобной информации может негативно сказаться на ходе переговоров.

5. РЕКОМЕНДАЦИИ ПО ВНЕДРЕНИЮ

Внедрение практик обеспечения безопасности данных зависит от корпоративной культуры, природы и характера рисков, степени чувствительности данных, которыми управляет компания, и типов используемых ею систем. При внедрении системных компонентов следует руководствоваться стратегическим планом обеспечения безопасности и учитывать базовую архитектуру.

5.1 Оценка готовности / Оценка рисков

Обеспечение безопасности данных неразрывно связано с корпоративной культурой. Часто организации лишь реагируют на критические события, вместо того чтобы предотвращать их за счет проактивного управления ответственностью и обеспечения возможности проверки. Поскольку достичь идеального уровня безопасности данных почти невозможно, наиболее подходящим способом избежать нарушений является выстраивание практики улучшения информированности и понимания требований, политик и процедур безопасности.

Организации могут повысить степень соответствия требованиям безопасности, прилагая усилия в следующих направлениях.

- **Обучение.** Продвижение стандартов посредством обучения и разъяснения инициатив в области безопасности на всех уровнях организации. Любые учебные программы должны дополняться механизмами оценки, включая, например, онлайновое тестирование, направленное на повышение осведомленности сотрудников. Подобные тренинги и тесты следует сделать обязательными компонентами процесса оценки эффективности сотрудников.
- **Согласованные и последовательные политики.** Определение политик безопасности и политик обеспечения нормативно-правового соответствия для рабочих групп и функциональных подразделений, которые должны дополнять и быть согласованными с корпоративными

политиками. Ориентация сотрудников на образ мышления «действуй локально» (act local)[1] способствует более активному вовлечению в процессы обеспечения безопасности.

- **Измерение выгод от мер по обеспечению безопасности данных.** Выгоды от мер по обеспечению ИБ следует увязывать с корпоративными инициативами. Организации должны включать объективные метрики для оценки деятельности в области безопасности в свои сбалансированные системы показателей (balanced scorecard) и системы оценки проектов.
- **Определение требований по безопасности данных для поставщиков.** Требования по ИБ следует включать в соглашения об уровне обслуживания (Service Level Agreement, SLA) и аутсорсинговые контракты. Требования, включаемые в SLA, должны охватывать всю деятельность по обеспечению безопасности.
- **Формирование «чувства крайней необходимости».** Следует постоянно акцентировать внимание сотрудников на требованиях законодательства и регулирующих органов, а также договорных обязательствах в части безопасности данных. Это способствуют формированию «чувства крайней необходимости» (sense of urgency), а также внутренней рамочной структуры управления безопасностью данных.
- **Непрерывные коммуникации.** Необходима реализация непрерывной программы обучения, информирующей сотрудников о практиках безопасной работы с данными и актуальных угрозах. Постоянно действующая программа способствует выработке у руководителей понимания важности процессов управления безопасностью данных и необходимости их поддержки.

5.2 Организационные и культурные изменения

Организации должны разрабатывать политики в области данных таким образом, чтобы они, с одной стороны, позволяли им успешно достигать поставленных целей, а с другой — обеспечивали надежную защиту чувствительной и регламентированной информации от ненадлежащего использования или несанкционированного доступа. При этом учет интересов всех заинтересованных сторон и минимизация рисков не должны существенно осложнять доступ к данным, необходимым для текущей работы. Часто техническая архитектура должна адаптироваться к архитектуре данных с целью обеспечения баланса всех потребностей и создания эффективной и безопасной электронной среды. В большинстве организаций для эффективной защиты данных требуется изменение поведения как руководителей, так и рядовых сотрудников.

Во многих крупных компаниях служба ИБ уже имеет политики, механизмы, инструменты обеспечения безопасности, системы контроля доступа, а также устройства и системы защиты информации. Важно четкое понимание того, в какой части эти элементы дополняют работу распорядителей данных и администраторов БД. Распорядители данных обычно отвечают за классификацию данных в отношении безопасности. Команды по обеспечению безопасности оказывают помощь в принятии мер по поддержке соответствия требованиям и внедрению операционных

[1] Имеется в виду широко распространенный принцип «Думай глобально, действуй локально» (Think global, act local). — *Примеч. науч. ред.*

процедур, основанных на политиках безопасности данных, а также классах конфиденциальности и категориях регламентирующих норм и правил.

Реализация мер по обеспечению безопасности данных без учета ожиданий клиентов и сотрудников чревата всплеском недовольства среди тех и других, а это серьезный организационный риск. Для того чтобы стимулировать соблюдение требований по безопасности, необходимо при планировании соответствующих мер рассмотреть позиции всех, кто будет непосредственно работать с данными и системами. Хорошо спланированный и всесторонний комплекс технических мер по обеспечению ИБ должен сделать защищенный доступ к данным максимально простым и удобным для всех заинтересованных лиц.

5.3 Доступность информации о наборах прав пользователей

Каждый набор прав (определяющий совокупность элементов данных, которые становятся доступными для пользователя после его авторизации) должен тщательно проверяться на стадии внедрения системы на предмет наличия в нем прав доступа к какой-либо регламентируемой информации. Для того чтобы знать, кто и к каким данным имеет доступ, необходимо управление метаданными, описывающими классы и категории данных в части конфиденциальности и требований регулирующих норм и правил, а также управление правами доступа и авторизациями. Классификация данных по уровням конфиденциальности и категориям нормативно-правовых требований должна быть стандартной частью процесса определения данных.

5.4 Обеспечение безопасности данных в условиях аутсорсинга

На аутсорсинг может быть передано всё, за исключением ответственности за результаты деятельности организации.

Аутсорсинг ИТ-услуг сопряжен с дополнительными проблемами и обязанностями в отношении безопасности данных. Привлечение сторонних организаций увеличивает количество людей, разделяющих ответственность за данные в рамках новых организационных и географических границ. В такой ситуации считавшиеся ранее неформальными роли и обязанности должны быть точно определены в виде контрактных обязательств. Договоры аутсорсинга должны содержать подробные описания обязанностей и ожиданий по каждой роли.

Любая форма аутсорсинга подвергает организацию дополнительному риску, в том числе и вследствие частичной потери контроля над технической средой и людьми, работающими с данными организации. Меры и процессы обеспечения безопасности данных должны рассматривать риск, исходящий от привлеченного поставщика услуг, одновременно и как внешний, и как внутренний.

С повышением уровня зрелости аутсорсинга в области ИТ всё больше организаций пересматривают свое отношение к таким услугам. Выработалось достаточно устойчивое общее мнение относительно того, что вопросы архитектуры и владения ИТ, включая архитектуру безопасности данных, должны оставаться в ведении организации. Иными словами, организация владеет

и управляет корпоративной архитектурой и архитектурой безопасности. Ответственность за реализацию архитектуры может быть возложена на привлеченного партнера.

Передача оперативного управления, но не ответственности, требует более строгого управления рисками и механизмов контроля. К таким механизмам относятся, например:

- соглашения об уровнях обслуживания;
- наличие в договорах аутсорсинга положений об ограничении ответственности;
- наличие в договорах аутсорсинга положений о праве на проведение проверок;
- четко определенные последствия нарушения договорных обязательств;
- регулярные отчеты поставщика услуг о состоянии безопасности данных;
- независимый мониторинг системной активности поставщика услуг;
- регулярные тщательные аудиторские проверки безопасности данных;
- непрерывные коммуникации с поставщиком услуг;
- знание и учет различий в юридических трактовках договоров в разных странах в случае возникновения споров с зарубежным поставщиком.

В среде, находящейся на аутсорсинге, критически важное значение обретает отслеживание происхождения (lineage) данных (или потока данных), перемещающихся между системами и отдельными людьми, в целях поддержки «цепочки поставок» (chain of custody), документирования цепочки операций хранения, контроля, передачи, анализа и использования данных. Аутсорсинговым организациям особенно полезно разрабатывать матрицы CRUD (операций создания, чтения, обновления и удаления), которые отражают распределение ответственности за данные в рамках бизнес-процессов, приложений, ролей и организаций, отслеживая преобразования, происхождение и «цепочку поставок» данных. Кроме того, матрица должна описывать распределение прав по принятию бизнес-решений или использованию функциональности приложений — например, права подтверждения счетов или заказов.

Еще одним полезным средством выработки более четкого понимания ролей, а также их сфер ответственности и распределения обязанностей, включая обязанности в области обеспечения безопасности, является матрица RACI[1].

Матрицу RACI можно включать, например, в договоры с поставщиками, а также в политики безопасности данных. С помощью разработки матриц, подобных RACI, вносится ясность относительно подотчетности и владения, что позволяет точно определить обязанности сторон договора аутсорсинга и обеспечить согласованное внедрение и поддержку политик безопасности.

Аутсорсинг ИТ-услуг, однако, не снимает с организации конечной ответственности за сопровождение данных. Поэтому критически важно иметь соответствующие механизмы

[1] RACI (*сокр.* от *англ.* Responsible/Accountable/Consulted/Informed) используются для назначения ответственного исполнителя (R), куратора (A), консультантов (C) и информируемых (I) в рамках проекта или направления работы. — *Примеч. пер.*

обеспечения соответствия действующим требованиям по безопасности и оставаться реалистами, отдавая себе отчет в том, что в этом вопросе нельзя всецело полагаться на привлеченных исполнителей.

5.5 Обеспечение безопасности данных в облачных средах

Стремительное развитие веб-приложений, сопровождающееся появлением таких подходов к взаимодействию, как «бизнес для бизнеса» (Business-to-Business, B2B) и «бизнес для клиента» (Business-to-Consumer, B2C), резко расширило границы распространения данных далеко за пределы четырех стен организации. Наблюдающийся в последние годы прогресс облачных вычислений раздвинул эти границы еще шире. Слова «как услуга» (as-a-service), указывающие на модель обслуживания с использованием облачных технологий, сегодня с легкостью добавляются ко многим технологическим и бизнес-понятиям. Термины «данные как услуга», «программное обеспечение как услуга», «платформа как услуга» стали общеупотребимыми. Облачные вычисления, обеспечивающие использование распределенных по сети ресурсов для обработки данных и информации, являются логичным завершением перехода к предоставлению «всего как услуги» (Anything-as-a-Service, XaaS).

Политики защиты данных в такой ситуации должны учитывать распределение данных по различным моделям услуг. Это подразумевает максимально возможное применение внешних стандартов в области безопасности данных.

Вопросы разделения ответственности, определения «цепочки поставок» данных, а также определения прав по владению и распоряжению данными особенно важны в случае облачных вычислений. Инфраструктурные соображения (например, определение стороны, ответственной за межсетевой экран при предоставлении ПО как услуги, или стороны, назначающей права доступа к серверам) оказывают непосредственное влияние на управление безопасностью данных и на политики в области данных.

Тонкая настройка или даже создание новой политики управления безопасностью данных, ориентированной специально на облачные вычисления, сегодня необходимы организациям любых размеров. Даже если сама организация напрямую не задействует облачные ресурсы, их вполне могут использовать ее бизнес-партнеры. А в мире сетевого обмена данными использование бизнес-партнером облачных технологий автоматически означает, что данные организации размещены в облаке. В целом для облачных сред справедливы те же принципы защиты чувствительных/конфиденциальных данных, которые применяются для обычной среды эксплуатации. В частности, они не допускают необоснованного распространения данных, ограничивая количество копий необходимым минимумом.

Архитектура внутреннего центра обработки облачных данных, включая виртуальные машины (потенциально, возможно, даже более защищенная), тем не менее должна соответствовать требованиям той же политики безопасности, что и остальные компоненты ИТ-инфраструктуры организации.

6. РУКОВОДСТВО БЕЗОПАСНОСТЬЮ ДАННЫХ

Обеспечение безопасности информационных систем предприятия и хранящихся в них данных требует согласованного взаимодействия заинтересованных сторон, представляющих бизнес и ИТ. В основе руководства безопасностью данных должны лежать строгие и четкие политики и процедуры.

6.1 Безопасность данных и корпоративная архитектура

Корпоративная архитектура определяет состав информационных активов и других компонентов организации, их взаимосвязи и бизнес-правила, которые учитывают преобразования данных, а также принципы и руководящие указания. Архитектура безопасности данных — составная часть корпоративной архитектуры, описывающая реализацию мер по обеспечению безопасности данных внутри организации, направленных на выполнение требований бизнес-правил и внешних регулирующих органов. Архитектура безопасности данных влияет на следующие аспекты обеспечения ИБ.

- Инструменты для управления безопасностью данных.
- Стандарты и механизмы шифрования данных.
- Правила доступа к ресурсам внешних поставщиков и других контрагентов.
- Протоколы передачи данных через интернет.
- Требования к документации.
- Стандарты удаленного доступа.
- Процедуры информирования о случаях нарушения ИБ.

Архитектура безопасности данных особенно важна для обеспечения интеграции данных между:

- внутренними системами и бизнес-единицами;
- организацией и ее внешними бизнес-партнерами;
- организацией и регулирующими органами.

Например, архитектурный шаблон сервис-ориентированного интеграционного механизма для внутренних и внешних участников обмена данными потребует иной реализации мер по обеспечению безопасности, нежели традиционная архитектура интеграции на основе электронного обмена данными (Electronic Data Interchange, EDI).

Для крупной организации необходимым условием защиты данных от злоупотреблений, хищений, утечек и потерь является введение формальной функции, обеспечивающей координацию между различными направлениями деятельности по внедрению архитектуры информационной безопасности. Представители каждого направления при этом должны как можно лучше понимать интересы других сторон, что позволит им разговаривать на общем языке и вырабатывать взаимно приемлемые подходы к решению общих задач.

7. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Andress, Jason. *The Basics of Information Security: Understanding the Fundamentals of InfoSec in Theory and Practice*. Syngress, 2011. Print.

Calder, Alan, and Steve Watkins. *IT Governance: An International Guide to Data Security and ISO27001/ISO27002*. 5th ed. Kogan Page, 2012. Print.

Fuster, Gloria González. *The Emergence of Personal Data Protection as a Fundamental Right of the EU*. Springer, 2014. Print. Law, Governance and Technology Series / Issues in Privacy and Data Protection.

Harkins, Malcolm. *Managing Risk and Information Security: Protect to Enable (Expert's Voice in Information Technology)*. Apress, 2012. Kindle.

Hayden, Lance. *IT Security Metrics: A Practical Framework for Measuring Security and Protecting Data*. McGraw-Hill Osborne Media, 2010. Print.

Kark, Khalid. «Building A Business Case for Information Security». *Computer World*. 2009-08-10, http://bit.ly/2rCu7QQ Web.

Kennedy, Gwen, and Leighton Peter Prabhu. *Data Privacy: A Practical Guide*. Interstice Consulting LLP, 2014. Kindle. Amazon Digital Services.

Murdoch, Don GSE. *Blue Team Handbook: Incident Response Edition: A condensed field guide for the Cyber Security Incident Responder*. 2nd ed. CreateSpace Independent Publishing Platform, 2014. Print.

National Institute for Standards and Technology (US Department of Commerce website), http://bit.ly/1eQYolG

Rao, Umesh Hodeghatta and Umesha Nayak. *The InfoSec Handbook: An Introduction to Information Security*. Apress, 2014. Kindle. Amazon Digital Services.

Ray, Dewey E. *The IT professional's merger and acquisition handbook*. Cognitive Diligence, 2012.

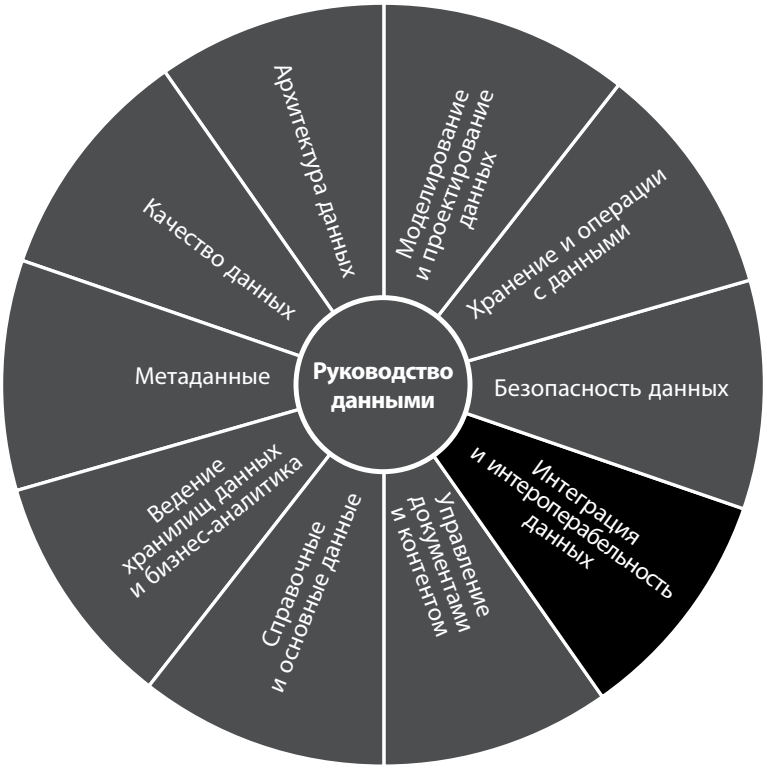
Schlesinger, David. *The Hidden Corporation: A Data Management Security Novel*. Technics Publications, LLC, 2011. Print.

Singer, P. W. and Allan Friedman. *Cybersecurity and Cyberwar: What Everyone Needs to Know®*. Oxford University Press, 2014. Print. What Everyone Needs to Know.

Watts, John. *Certified Information Privacy Professional Study Guide: Pass the IAPP's Certification Foundation Exam with Ease!* CreateSpace Independent Publishing Platform, 2014. Print.

Williams, Branden R., Anton Chuvakin Ph. D. *PCI Compliance: Understand and Implement Effective PCI Data Security Standard Compliance*. 4th ed. Syngress, 2014. Print.

Интеграция и интероперабельность данных



DAMA-DMBOK2 Data Management Framework

1. ВВЕДЕНИЕ

Интеграция и интероперабельность данных (Data Integration and Interoperability, DII) — область знаний по управлению данными, которая описывает процессы, связанные с перемещением и консолидацией данных как внутри хранилищ, приложений и организаций, так и в рамках обеспечения их взаимодействия. Интеграция позволяет объединять данные в согласованные физические или виртуальные формы. Под интероперабельностью данных подразумевается способность двух или более систем к обмену информацией. Решения в области DII необходимы для реализации

базовых функций управления данными, которые используются в большинстве организаций. К ним относятся:

- миграция и конвертация данных;
- консолидация данных в концентраторах (или хабах — hubs) или витринах;
- интеграция программных продуктов сторонних поставщиков в единый комплекс приложений организации;
- совместное использование данных различными приложениями как в рамках одной организации, так и в рамках группы организаций;
- распространение данных по хранилищам и ЦОДам;
- архивирование данных;
- управление интерфейсами обмена данными;
- получение внешних данных и подготовка их к использованию;
- интеграция структурированных и неструктурированных данных;
- предоставление оперативной информации и поддержка управленческих решений.

DII находится в зависимости от других областей управления данными:

- **руководство данными** — в части определения правил преобразования данных и структуры сообщений;
- **архитектура данных** — в части разработки архитектуры DII-решений;
- **безопасность данных** — в части обеспечения соответствия DII-решений требованиям по безопасности данных, как постоянно хранимых (persistent), так и виртуальных (virtual), а также «данных в движении» (in motion), которые перемещаются между приложениями и организациями;
- **метаданные** — в части отслеживания такой информации, как техническое описание данных (постоянно хранимых, виртуальных и передаваемых), описание их значения для бизнеса, описание бизнес-правил преобразования данных, а также история операций и сведения о происхождении (lineage) данных;
- **хранение и операции с данными** — в части физической реализации решений по хранению данных;
- **моделирование и проектирование данных** — в части проектирования структур данных (постоянно хранимых, виртуальных, а также сообщений, которые перемещаются между приложениями и организациями).

Интеграция и интероперабельность данных критически важны для ведения хранилищ данных и бизнес-аналитики, а также для управления справочными и основными данными, поскольку обе эти области управления данными сфокусированы на преобразовании и интеграции данных из систем-источников в консолидационных хабах, с последующей передачей консолидированных данных в целевые системы, которые предоставляют их потребителям (людям и другим системам).

Интеграция и интероперабельность данных занимают центральное место в недавно появившейся области управления большими данными. Эта область подразумевает интеграцию различных видов данных, включая структурированные данные из всевозможных БД, неструктурированные текстовые данные из документов или файлов, а также неструктурированные данные других видов, такие как аудио, видео и потоковые. Большие данные могут быть объектом интеллектуального анализа, использоваться для построения предиктивных моделей и получения оперативной информации.

1.1 Бизнес-драйверы

Потребность в управлении перемещением данных — основной драйвер DII. Поскольку в большинстве организаций имеются сотни, а то и тысячи всевозможных баз и хранилищ данных, управление процессами перемещения данных между местами хранения внутри организации и обмена данными с другими организациями становится одной из главных сфер ответственности любой ИТ-службы. Без надлежащего управления процесс перемещения данных быстро исчерпает все их ресурсы и возможности, лишив при этом необходимой поддержки традиционные приложения и области управления данными.

Повсеместный переход организаций на использование покупного прикладного ПО вместо разработки собственного усилил потребность в обеспечении интеграции и интероперабельности на корпоративном уровне. Каждое коммерческое приложение добавляет собственный набор хранилищ основных данных, транзакционных данных и данных отчетов, и все их приходится интегрировать с другими хранилищами данных, уже имеющимися в организации. Даже системы планирования ресурсов предприятия (ERP), обеспечивающие выполнение общих функций организации, практически никогда не охватывают всех необходимых хранилищ данных. Они также должны интегрировать свои данные с другими данными организации.

Потребность в управлении сложностью и связанные со сложностью затраты требуют корпоративного подхода к построению архитектуры интеграции. Очевидно, что корпоративная архитектура более эффективна и менее затратна, чем распределенные решения или решения «точка-точка». Разработка решений «точка-точка» для связи между приложениями может потребовать тысяч или даже миллионов интерфейсов обмена, что быстро исчерпает возможности даже самой эффективной службы ИТ-поддержки.

Информационные хабы (hubs), такие как хранилища данных (data warehouses) и решения по управлению основными данными, значительно облегчают преодоление этой проблемы, обеспечивая консолидацию данных, которые требуются многим приложениям, а также унификацию представления этих данных. Аналогичным образом и задачи управления операционными и транзакционными данными, требующими общего доступа в рамках всей организации, значительно упрощаются, если использовать методы интеграции данных корпоративного уровня: например, по схеме звезды с интеграцией в центре (hub-and-spoke, дословно — «ступица и спица») в сочетании с каноническими моделями обмена сообщениями.

Рисунок 66.
Контекстная диаграмма: интеграция и интероперабельность данных

Еще один важный бизнес-драйвер интеграции — управление затратами на поддержку. Перемещение данных с использованием множества технологий, каждая из которых требует специфических навыков разработки и обслуживания, способно привести к непомерному росту стоимости поддержки. Внедрение стандартных инструментов позволяет значительно сократить потребности в обслуживании и персонале, а также повысить эффективность поиска и устранения неполадок. Снижение сложности управления интерфейсами обмена данными может сократить затраты на их обслуживание и дать возможность перераспределить ресурсы сопровождения на решение других приоритетных задач организации.

Проведение работ в области DII также помогает организации соблюдать действующие стандарты и регламенты обработки данных. DII-системы корпоративного уровня позволяют повторно использовать коды, обеспечивающие соответствие требованиям нормативных документов, и упрощают проверку их соблюдения.

1.2 Цели и принципы

Внедрение практик и решений в области интеграции и интероперабельности данных преследует следующие цели:

- своевременное предоставление требуемых данных потребителям (как пользователям, так и приложениям) в нужном им формате;
- физическая или виртуальная консолидация данных в хабах;
- снижение стоимости и сложности решений по управлению данными за счет разработки общих моделей и интерфейсов;
- выявление значимых событий (возможностей и угроз) и автоматический запуск процедур выдачи уведомлений и принятия мер;
- поддержка функций BI, аналитики, управления основными данными и обеспечение операционной эффективности.

Внедряя решения в области DII, организация должна руководствоваться следующими принципами.

- При проектировании надлежит придерживаться корпоративного подхода, обеспечивающего возможность последующего расширения и масштабирования, но реализацию проводить итерационно, методом пошагового ввода новых решений в эксплуатацию.
- Должны сбалансированно учитываться локальные и корпоративные потребности в данных, а также в поддержке и сопровождении.
- Следует обеспечивать ответственность бизнеса за проектирование решений и проведение других работ в области DII. Эксперты со стороны бизнеса должны привлекаться к разработке и модификации правил преобразования данных — как постоянно хранимых, так и виртуальных.

1.3 Основные понятия и концепции

1.3.1 Извлечение, преобразование и загрузка

В основе любых решений в области интеграции и интероперабельности данных лежит процесс извлечения, преобразования и загрузки (Extract, Transform, and Load, ETL). Вне зависимости от того, выполняются они физически или виртуально, в пакетном режиме или режиме реального времени, эти шаги непременно присутствуют при перемещении данных внутри или между приложениями и организациями.

В зависимости от требований по интеграции данных процедуры ETL могут выполняться в режиме периодической (пакетной) обработки или обработки по мере доступности новых или обновленных данных (в режиме реального времени или управляемой на основе событий — event driven). Обработка данных о текущих операциях обычно проводится в режиме реального времени или в режиме, близком к реальному времени (near real-time), а данных, требуемых для анализа и отчетности, — по графику, в пакетном режиме.

Требованиями по интеграции также определяется, сохраняются или нет извлеченные и преобразованные данные физически в промежуточных структурах временного хранения (staging structures). Физическое хранение позволяет вести контрольный журнал шагов преобразований и в случае сбоя перезапускать процесс с промежуточной точки. Однако структуры временного хранения требуют дисковой памяти и времени на запись и чтение. Поэтому там, где требуется интеграция данных с минимальным запаздыванием, физического сохранения промежуточных результатов преобразования данных обычно не предусматривается.

1.3.1.1 ИЗВЛЕЧЕНИЕ

Процесс извлечения включает выбор требуемых данных и выгрузку их из источника. Извлеченные данные сохраняются на дисковый накопитель или в оперативную память. В первом случае временное хранилище может находиться как вместе с хранилищем-источником, так и вместе с целевым хранилищем, или и с тем и с другим.

Если этот процесс выполняется в среде системы, обеспечивающей операционную деятельность, потребность в ее ресурсах стараются минимизировать, чтобы не обделять ими текущие рабочие процессы. Часто практикуется вариант пакетной обработки сложных запросов на выбор и извлечение обновленных данных из исходного хранилища в период минимальной рабочей нагрузки на систему.

1.3.1.2 ПРЕОБРАЗОВАНИЕ

Процесс преобразования переводит выбранные данные в структуру, совместимую с целевым хранилищем. Следует различать случаи преобразования данных при перемещении из исходного хранилища в целевое, копировании из исходного хранилища в несколько целевых, а также использования преобразованных данных для запуска каких-либо процессов без последующего их сохранения.

Примеры преобразований могут включать следующее.

- **Изменения формата.** Перекодировка данных из одного технического формата в другой — например, из EBCDIC в ASCII.
- **Изменения структуры.** Внесение изменений в структуру данных — например, из ненормализованных в нормализованные записи.
- **Семантическая конверсия.** Преобразование данных для поддержки соответствующего семантического представления. Пример: в исходном наборе данных допустимые значения атрибута Пол — 0, 1, 2 или 3; в целевом наборе им соответствуют текстовые значения: НЕ ИЗВЕСТЕН, МУЖСКОЙ, ЖЕНСКИЙ или ДАННЫЕ НЕ ПРЕДОСТАВЛЕНЫ.
- **Дедупликация.** Если правила требуют, чтобы ключевые значения или записи были уникальными, осуществляется сканирование целевого набора данных на предмет дублирующихся строк — и они удаляются.
- **Переупорядочивание.** Изменение порядка элементов данных или записей в соответствии с определенным шаблоном.

Преобразование может проводиться в режимах реального времени или пакетной обработки, результаты сохраняются либо физически в области временного хранения, либо виртуально в памяти до момента готовности данных к стадии загрузки. Данные, полученные в результате стадии преобразования, должны быть готовы к интеграции с данными в целевой структуре.

Рисунок 67. Поток операций процесса ETL

1.3.1.3 ЗАГРУЗКА

Завершающий этап процесса ETL — физическое сохранение или представление преобразованных данных в целевой системе. В зависимости от характера произведенных преобразований, назначения целевой системы и предполагаемого применения данных они могут или требовать дополнительной обработки для интеграции с другими данными, или же быть готовыми к использованию потребителями.

1.3.1.4 ИЗВЛЕЧЕНИЕ, ЗАГРУЗКА, ПРЕОБРАЗОВАНИЕ (ELT)

Если целевая система располагает значительно большими возможностями трансформации, нежели исходная или промежуточная прикладные системы, порядок процессов может быть изменен на ELT (Extract, Load, and Transform — извлечение, загрузка и преобразование). ELT позволяет выполнять преобразование данных после загрузки в целевое хранилище или нередко как часть процесса загрузки. ELT также позволяет сохранять в целевой системе копию сырых (raw) данных, если они могут пригодиться в каких-то иных процессах. Такой подход широко распространен в средах обработки больших данных, где ELT используется для загрузки озера данных (data lake) (см. главу 14).

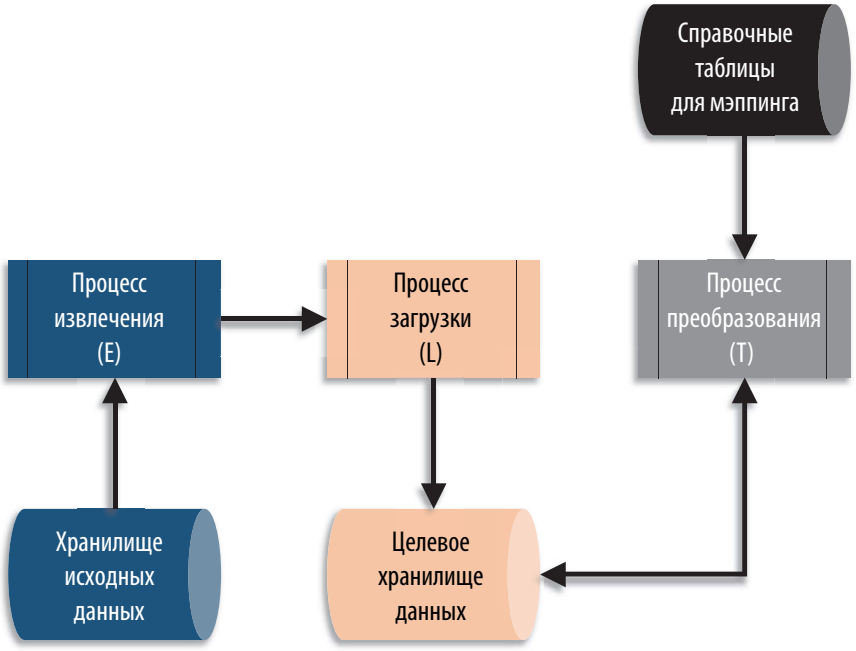


Рисунок 68. Поток операций процесса ELT

1.3.1.5 МЭППИНГ

Мэппинг (*mapping*), являясь синонимом преобразования, означает как процесс преобразования (lookup matrix) матрицы из исходной структуры в целевую, так и результат этого процесса. Мэппинг определяет источники, откуда извлекаются данные, правила идентификации данных,

подлежащих извлечению, целевые хранилища для загрузки и правила идентификации целевых строк, подлежащих замене (если это предполагается), а также правила преобразования или формулы вычислений, которые должны быть применены. Многие программные средства интеграции предлагают функции визуализации мэппинга, позволяющие разработчикам использовать графический интерфейс для создания кода, реализующего преобразование.

1.3.2 Задержка

Задержка (latency) — это разница во времени между моментом, когда данные были сгенерированы в системе-источнике, и моментом, когда они стали доступны в целевой системе. Различные подходы к обработке данных определяют различную степень задержки. Задержка может быть высокой (при пакетной обработке), низкой (при запуске процедур переноса на основе событий) или очень низкой (при использовании синхронизации в режиме реального времени).

1.3.2.1 ПАКЕТНАЯ ОБРАБОТКА

Основной объем данных перемещается между приложениями или организациями в массивах или файлах, отправка которых выполняется по запросу пользователей либо в автоматическом режиме по запрограммированному графику. Такой вид обмена данными называется *пакетным* (*batch*) или *ETL*.

Пакет может включать либо полный набор данных по состоянию на момент перед отправкой — например, балансовые ведомости, подготовленные по завершении отчетного периода, — либо выборку, которая отражает только те изменения, которые произошли за период с момента отправки предыдущего пакета, — например, изменения адресов, зарегистрированные за прошедшие сутки. В первом случае передаваемый набор данных называют «*снимком*» (или *снапшотом — snapshot*), во втором — «*дельтой*» (*delta*).

В решениях, использующих пакетную интеграцию данных, обычно предполагается весьма значительная задержка обновления данных в целевой системе. Пакетную обработку полезно использовать для выполнения операций с большими объемами данных за короткий выделенный отрезок времени. В больших хранилищах данных обычно используется именно этот метод даже в тех случаях, когда технически поддерживаются и альтернативные решения с меньшим временем запаздывания.

Для ускорения обработки и сокращения задержки обновления некоторые интеграционные решения предусматривают обработку микропакетов (micro-batch) с большей частотой: например, каждые пять минут, а не раз в сутки, как в большинстве стандартных решений.

Пакетная интеграция данных широко используется в целях конвертирования, миграции и архивирования, а также для обмена информационными массивами с хранилищами и витринами данных. Чем длительнее задержка обновления данных, тем выше риск некорректной работы использующих эти данные приложений. Для минимизации подобных эффектов составляйте расписание обмена обновлениями между приложениями с периодичностью не реже чем ежесуточно: например, по завершении рабочего дня или в ночное время следом за проведением какой-либо

специальной обработки данных. Во избежание передачи в корпоративные хранилища данных неполных наборов данных плановые выполнения заданий по их копированию должны быть привязаны к завершению цикла ежедневной, еженедельной или ежемесячной отчетности.

1.3.2.2 СБОР ДАННЫХ ИЗМЕНЕНИЙ

Сбор данных изменений (Change Data Capture) — метод, позволяющий уменьшить загрузку каналов связи за счет передачи лишь тех данных, которые были изменены за определенный период. С помощью этого метода осуществляется мониторинг информационных массивов на предмет изменений (добавления, корректировки, удаления) и передача изменений («дельт») в другие массивы, приложения и организации — потребители данных. Данные также могут быть помечены в рамках процесса специальными идентификаторами, например флажками или временны́ми метками. Сбор данных изменений может осуществляться на основе данных (data-based) или на основе журнала (log-based) (см. главу 6).

Известны три метода сбора данных изменений на основе данных.

- Система-источник расставляет специальные элементы данных, например временны́е метки в рамках временного ряда, коды или флажки, которые и служат индикаторами изменений. Процесс извлечения использует правила, основанные на этих элементах, для определения строк, которые следует извлечь.
- Процессы системы-источника при внесении изменений в данные добавляют соответствующие объекты или идентификаторы в специальный список, который потом используется для определения данных, подлежащих извлечению.
- Процессы системы-источника после внесения изменений копируют измененные данные в отдельный объект (в рамках выполняемой транзакции), который потом используется в ходе процесса извлечения. Этот объект не обязательно должен создаваться в среде СУБД.

Все три метода используют встроенные возможности приложения-источника, которые могут оказаться слишком затратными в отношении ресурсов и предполагают модификацию приложения.

При сборе данных изменений на основе журнала СУБД создает журнал операций с базой данных, который копируется и обрабатывается с целью поиска изменений, переносимых в целевую базу данных. Сложные изменения бывают трудны для переноса. В таких случаях — с целью их временного хранения и обработки — могут использоваться промежуточные структуры, схожие с исходными объектами.

1.3.2.3 РЕЖИМ, БЛИЗКИЙ К РЕАЛЬНОМУ ВРЕМЕНИ, И УПРАВЛЕНИЕ НА ОСНОВЕ СОБЫТИЙ

В большинстве решений по интеграции данных, не использующих пакетную обработку, реализована обработка в режиме, близком к реальному времени, или управляемая на основе событий. Данные обрабатываются небольшими порциями в течение суток — либо по заданному расписанию, либо после того, как происходит событие, приводящее к изменению данных, например

обновление. Обработка в режиме, близком к реальному времени, осуществляется с меньшей задержкой по сравнению с пакетным режимом и часто снижает рабочую нагрузку на систему (поскольку работа распределена во времени), однако уступает по скорости решениям с синхронизованной интеграцией данных. Решения, работающие в режиме, близком к реальному времени, обычно реализуются на базе корпоративной сервисной шины.

Информация о состоянии и зависимость между процессами должны отслеживаться процессом загрузки приложения — получателя данных. В частности, для получения корректного результата может потребоваться загрузка данных в строго определенном порядке. Например, сначала загружаются основные данные или данные измерений, и лишь после этого можно переходить к обработке транзакционных данных, которые их используют.

1.3.2.4 АСИНХРОННАЯ ОБРАБОТКА

В случае асинхронных потоков данных система-источник продолжает обработку данных, не дожидаясь подтверждения получения обновления от принимающей системы. Рассинхронизация подразумевает, что одна из двух систем (передающая или принимающая данные) может работать в автономном режиме, в то время как другая будет подключена к сети.

Асинхронная интеграция данных не препятствует продолжению выполнения процессов приложения-источника, а также не делает его недоступным, если недоступны какие-либо из приложений-получателей. Поскольку при асинхронной конфигурации обновления данных в целевых приложениях производятся с запаздыванием, такой тип интеграции называется *интеграцией в режиме, близком к реальному времени*. Задержка между обновлением данных в источнике и отражением изменений в целевых наборах данных для такого режима обычно составляет от нескольких секунд до нескольких минут.

1.3.2.5 ОБРАБОТКА В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ, СИНХРОННАЯ

В некоторых ситуациях недопустима даже малейшая рассогласованность между исходным и целевым наборами данных. В таких случаях обязательно использование решений, обеспечивающих синхронизацию данных в режиме реального времени.

В синхронных интеграционных решениях выполняющийся процесс дожидается подтверждения обновления от всех приложений или процессов-адресатов, прежде чем перейти к осуществлению следующей операции или к обработке очередной транзакции. Это означает, что решение может обработать меньше транзакций из-за простоев, вызванных ожиданием получения всех требуемых подтверждений синхронизации. Если хотя бы одно из приложений, требующих обновления данных, недоступно, приложение-источник не может завершить транзакцию. В такой ситуации данные остаются синхронизованными, но существует опасность зависимости стратегически важных приложений от второстепенных приложений с низкой степенью критичности.

Решения, использующие подобную архитектуру, требуют оценки того, насколько экономически оправданным является использование дорогостоящих технических решений,

обеспечивающих минимальную разницу между синхронизируемыми в реальном времени наборами данных. Синхронизацию вполне можно обеспечить средствами СУБД — например, с помощью механизма двухфазного подтверждения транзакции, гарантирующего: либо все изменения данных в рамках бизнес-транзакции проведены успешно, либо не проведено ни одно из них. Например, в финансовых учреждениях двухфазное подтверждение транзакций используется для того, чтобы обеспечить полную синхронизацию таблиц с данными транзакций и таблиц с балансовыми данными. При этом в большинстве прикладных программ двухфазное подтверждение не применяется. Как следствие, сохраняется небольшой риск того, что в результате неожиданного прерывания приложения один набор данных обновится, а другие — нет.

Будучи затратными по ресурсам, решения, обеспечивающие синхронизацию в режиме реального времени, требуют меньше внимания к управлению состоянием баз данных, чем асинхронные решения, поскольку порядок обработки транзакций четко контролируется приложениями, которые обновляют данные. Однако работа в таком режиме может привести к блокировке и задержке других транзакций.

1.3.2.6 ОБРАБОТКА С НИЗКОЙ ЗАДЕРЖКОЙ ИЛИ ПОТОКОВАЯ ОБРАБОТКА

Сегодня достигнуты огромные успехи в сфере разработки решений, которые обеспечивают сверхбыструю интеграцию данных. Однако их реализация сопряжена с серьезными затратами на аппаратное и программное обеспечение. Использование столь дорогостоящих решений бывает оправданным в тех случаях, когда организации требуется сверхскоростная передача данных на большие расстояния. Потоковые данные (streaming data) «вытекают» из компьютерных систем в непрерывном режиме по ходу событий. Потоки данных фиксируют такую информацию о событиях, как сведения о покупках товаров или ценных бумаг, комментарии в соцсетях или показания датчиков, отслеживающих координаты местоположения, а также значения температуры, нагрузки и прочих характеристик.

Решения, обеспечивающие интеграцию данных с низкой задержкой, разрабатываются для поддержания максимально быстрого отклика на события. В них часто задействуются аппаратные (например, твердотельные диски) или программные (например, БД в оперативной памяти) средства хранения данных, избавляющие от потерь времени на запись или считывание с традиционных жестких дисков. Обработка данных в оперативной памяти или на SSD-накопителе требует в тысячи раз меньше времени, чем операции обращения к жесткому диску.

Решения с низкой задержкой обычно используют асинхронный обмен данными, при котором транзакциям не требуется ждать подтверждения от последующих процессов, перед тем как начать обрабатывать следующую порцию данных.

Обычно в решениях с низкой задержкой используются мощные мультипроцессорные системы или конфигурации с параллельной обработкой данных, обеспечивающие распределение операций по обработке входных данных по множеству процессоров, поскольку производительности, достигаемой на малом количестве процессоров, для решения практических задач оказывается недостаточно.

1.3.3 Репликация

Некоторые приложения обеспечивают малое время отклика на обращения пользователей со всех уголков мира за счет ведения идентичных копий наборов данных во многих географически разнесенных местах. Решения с использованием репликации позволяют минимизировать влияние операций по анализу и сложных запросов на производительность среды, обеспечивающей в первую очередь обработку транзакций.

Такие решения предполагают синхронизацию физически распределенных копий наборов данных. Большинство СУБД имеют для этого специальные утилиты, обеспечивающие репликацию. Для эффективной работы таких утилит желательно, чтобы все базы данных находились под управлением СУБД, реализованных на одной технологии. Репликационные решения обычно осуществляют мониторинг журнала изменений набора данных, а не сам набор данных. Они оказывают минимальное влияние на работу операционных приложений, поскольку не конкурируют с ними за доступ к набору данных. Обмен данными между копиями сводится лишь к передаче данных из журнала изменений. Стандартные репликационные решения обеспечивают обмен обновлениями в режиме, близком к реальному времени: задержка между проведением изменений в одном экземпляре набора данных и проведением этих же изменений в других экземплярах минимальна.

Преимущества репликации — минимум влияния на исходный набор данных и минимум передаваемой информации — делают такие решения крайне привлекательными, и поэтому они часто используются для интеграции даже таких наборов данных, которые не удалены друг от друга на большие расстояния. Утилиты СУБД не требуют большого объема программирования, поэтому количество программных ошибок обычно невелико.

Репликационные утилиты работают оптимально, когда исходные и целевые наборы данных являются точными копиями друг друга. Если же набор-источник и целевой набор различаются, то синхронизация подвержена риску. Если исходный и целевой наборы данных не являются точными копиями, необходимо организовать промежуточную область для размещения и ведения точной копии набора — источника данных. Это требует дополнительного дискового пространства и, возможно, применения дополнительных технологий баз данных.

Решения, основанные на репликации, не являются оптимальными в ситуациях, когда требуется синхронизировать многочисленные наборы данных, в каждый из которых могут вноситься изменения. Если есть вероятность, что в разные экземпляры набора данных могут быть одновременно внесены изменения в одной и той же его части, то существует риск рассинхронизации данных. В измененную часть одного из экземпляров могут быть без предупреждения записаны изменения, внесенные в другой экземпляр. Изменения, внесенные в первый экземпляр, при этом пропадут (см. главу 6).

1.3.4 Архивирование

Данные, используемые нечасто или малоактивно, можно перенести в альтернативную структуру или на устройство хранения, которое требует от организации меньших затрат. Для переноса данных в архив (возможно, с преобразованием) обычно используются функции ETL. Переносу

в архив подлежат данные приложений, выводимых из эксплуатации, а также данные действующих систем среды эксплуатации, не использовавшиеся на протяжении длительного времени, поскольку это будет способствовать повышению операционной эффективности.

Критически важно следить за работоспособностью технологий архивирования и, в частности, контролировать совместимость новых технологий со старыми архивами, чтобы данные в них оставались доступными и после внедрения технологических изменений. Если новое приложение для работы с архивами не поддерживает старые форматы, это может быть сопряжено с рисками, особенно если не истек установленный законом обязательный срок хранения архивных данных (см. главу 9).

1.3.5 Корпоративный формат сообщений / Каноническая модель

Каноническая модель данных — общая модель (используемая организацией или группой, отвечающей за обмен данными), стандартизирующая формат, в котором осуществляется распространение данных. Она нужна, в частности, при использовании звездообразной схемы обмена данными с интеграцией в центре (hub-and-spoke), когда системы общаются друг с другом исключительно через центральный информационный хаб. Все передаваемые данные преобразуются в общий формат сообщений, принятый в организации (каноническую модель) (см. главу 5). Использование канонической модели ограничивает количество преобразований данных при обмене между системами или организациями. Каждой системе достаточно реализовать преобразование данных только в каноническую модель (при передаче) или из нее (при приеме), вместо того чтобы разрабатывать отдельные средства преобразования для множества систем, с которыми осуществляется обмен.

Хотя разработка и согласование единого унифицированного формата сообщений — задача ответственная и трудоемкая, внедрение канонической модели значительно уменьшает сложность решений по обеспечению интероперабельности данных организации и тем самым существенно сокращает затраты на их поддержку. При этом следует учитывать, что создание и сопровождение канонической модели, распространяющейся на все случаи обмена данными между системами организации, предполагает проведение комплекса затратных мероприятий по реализации и поддержке информационного взаимодействия по звездообразной схеме. Эти усилия можно считать оправданными лишь при необходимости обеспечения взаимодействия более трех разнородных систем. В средах, где обмениваются данными более ста прикладных систем, интеграционное решение на основе канонической модели является единственно возможным и незаменимым.

1.3.6 Модели взаимодействия

Модели взаимодействия описывают способы обеспечения связи между системами с целью обмена данными.

1.3.6.1 МОДЕЛЬ «ТОЧКА-ТОЧКА»

В большинстве случаев взаимодействие между системами происходит в режиме прямого обмена данными по схеме «точка-точка» (point-to-point). При небольшом количестве систем это вполне

рациональный подход. Однако по мере роста числа систем, обращающихся к одним и тем же источникам данных, он становится неэффективным и рискованным для организации.

- **Влияние на обработку данных при выполнении операций.** Если системы-источники обслуживают операционную деятельность, то выполнение внешних запросов отнимает ресурсы и, в целом, замедляет обработку данных при выполнении операций.
- **Управление интерфейсами.** Количество интерфейсов обмена данными, необходимых для реализации модели «точка-точка», приблизительно равно квадрату числа систем, участвующих в обмене (s^2, где s — число систем). И все эти интерфейсы нужно сначала создать, а затем обслуживать и поддерживать. С ростом количества систем трудозатраты на эту деятельность быстро превышают трудозатраты на обслуживание и поддержку самих систем.
- **Возможное рассогласование данных.** Проблемы такого рода возникают, когда различные системы используют одни и те же данные в различных вариантах представления или форматах. Использование множества интерфейсов приводит к возникновению противоречий в данных, которые тиражируются во все системы, участвующие в их обработке.

1.3.6.2 ЗВЕЗДООБРАЗНАЯ МОДЕЛЬ С ИНТЕГРАЦИЕЙ В ЦЕНТРЕ

Звездообразная модель с интеграцией в центре (hub-and-spoke) является альтернативой модели «точка-точка». Она позволяет консолидировать (физически или виртуально) данные совместного использования в центральном хабе данных, к которому обращаются все приложения. То есть их взаимодействие друг с другом по протоколам «точка-точка» заменяется обращением к общей центральной системе, контролирующей данные. Хорошо известными примерами хабов данных являются хранилища данных (Data Warehouses), витрины данных (Data Marts), хранилища операционных данных (Operational Data Stores) и хабы для управления основными данными.

Центральные хабы обеспечивают единообразное и согласованное представление данных при ограниченном влиянии на производительность поставляющих их систем. Они минимизируют количество систем и операций извлечения, требующих доступа к исходным данным, минимизируя таким образом влияние на ресурсы систем-источников. Добавление в состав информационных систем предприятия нового приложения требует построения всего лишь одного интерфейса — обеспечивающего взаимодействие с хабом данных. Звездообразная модель становится эффективнее и дешевле модели «точка-точка» уже при относительно небольшом числе систем, а когда их счет идет на сотни и тысячи, централизация становится неизбежной.

Корпоративные сервисные шины (Enterprise Service Buses, ESB) — интеграционные решения, которые обеспечивают синхронизацию данных в режиме, близком к реальному времени, между многими системами. Такие решения используют понятие хаба данных, предоставляющего стандартный формат или каноническую модель для совместного использования данных организацией.

Звездообразные модели не всегда оптимальны. Иногда они неприемлемы из-за слишком длительной задержки, иногда — из-за недостаточной производительности. Использование

в звездообразной архитектуре центрального хаба влечет слишком высокие накладные расходы, и при небольшом количестве систем предпочтительнее обойтись взаимодействием «точка-точка». Однако преимущества хаба чаще всего перевешивают недостатки, когда требуется оперативный обмен данными между тремя и более системами. Кроме того, применение при реализации обмена данными шаблона проектирования, соответствующего звездообразной модели, уменьшает распространение и снижает количество отдельных специфических трансформаций и интеграционных решений. Организационная поддержка деятельности по интеграции данных при этом сильно упрощается.

1.3.6.3 МОДЕЛЬ «ПУБЛИКАЦИЯ-ПОДПИСКА»

Модель публикации и подписки (publish and subscribe) предусматривает наличие систем, поставляющих данные («издателей»), и систем, получающих эти данные («подписчиков»). Системы, поставляющие данные, вносятся в каталог сервисов данных, а системы, которым эти данные требуются, должны подписываться на услуги провайдера. После публикации данные автоматически рассылаются подписчикам.

При наличии множества потребителей одних и тех же наборов данных или данных в одном и том же формате подготовка этих данных (с последующим открытием доступа к ним) в централизованном порядке позволяет обеспечивать использование потребителями согласованных наборов данных и их регулярное своевременное обновление.

1.3.7 Понятия и концепции, используемые в архитектуре DII

1.3.7.1 СВЯЗЫВАНИЕ ПРИЛОЖЕНИЙ

Связывание (coupling) описывает, насколько тесно две системы «переплетены» между собой. Сильно связанные (tightly coupled) системы обычно имеют синхронизованный интерфейс: то есть, отправив запрос другой системе, первая приостанавливает обработку данных до получения требуемого ответа. Сильная увязка рискованна: если одна система оказывается недоступной, то фактически недоступны будут обе системы, — и план обеспечения непрерывности бизнеса (business continuity plan) в отношении этих двух систем должен быть общим (см. главу 6).

Слабое связывание (loose coupling) — более предпочтительный подход к проектированию интерфейса (там, где это возможно). При таком подходе получение ответов на запросы, обращенные к другой системе, не является обязательным условием продолжения работы первой системы, то есть доступность каждой из слабо связанных систем не зависит от доступности другой системы. Слабое связывание может быть реализовано с использованием различных средств: например, посредством сервисов, интерфейсов прикладного программирования (API) или очередей сообщений (см. рис. 69).

Примером шаблона проектирования для слабо связанного информационного взаимодействия служит сервис-ориентированная архитектура (Service-Oriented Architecture, SOA), использующая корпоративную сервисную шину.

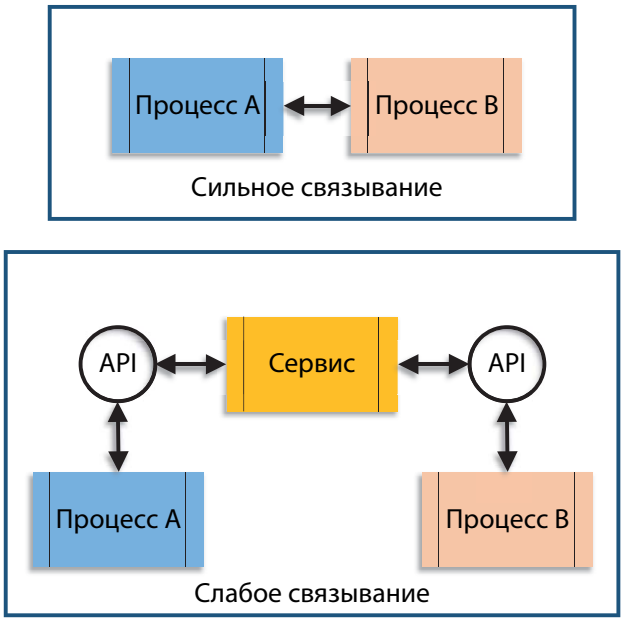


Рисунок 69. Варианты связывания приложений

Когда системы слабо связаны, замена любой из них на альтернативную теоретически даже не потребует переделки остальных систем, с которыми она взаимодействует, поскольку все точки взаимодействия четко определены.

1.3.7.3 ИНТЕГРАЦИЯ КОРПОРАТИВНЫХ ПРИЛОЖЕНИЙ (EAI)

В модели интеграции корпоративных приложений (Enterprise Application Integration, EAI) программные модули взаимодействуют друг с другом только посредством точно определенных вызовов функций интерфейса прикладного программирования (API). Хранилища данных используют для проведения обновлений только свои собственные программные модули, а другие программные средства могут получить доступ к данным хранилищ, только используя их API.

Модель EAI построена на основе объектно-ориентированного подхода, который акцентирует внимание на повторном использовании и возможности замены любого модуля без оказания какого-либо влияния на работу других модулей.

1.3.7.4 КОРПОРАТИВНАЯ СЕРВИСНАЯ ШИНА (ESB)

Корпоративная сервисная шина (ESB) представляет собой специальную систему, которая выполняет функции посредника между прикладными системами, передавая сообщения от одной системы к другой. Приложения могут отправлять и получать сообщения или файлы, используя ESB; при этом они изолированы от других ее процессов. ESB является примером реализации подхода к построению интеграционных решений, основанного на слабом связывании. Она действует как сервис обмена данными между приложениями (см. рис. 70).

Рисунок 70. Корпоративная сервисная шина (ESB)

1.3.7.5 СЕРВИС-ОРИЕНТИРОВАННАЯ АРХИТЕКТУРА (SOA)

Наиболее зрелые корпоративные стратегии интеграции приложений используют концепцию сервис-ориентированной архитектуры (SOA), в которой функциональность по предоставлению или обновлению данных (или другие сервисы данных) может быть представлена в виде точно определенных вызовов сервисов, используемых приложениями в процессе их взаимодействия. При таком подходе приложениям не нужно ни взаимодействовать друг с другом напрямую, ни знать что-либо о внутренней структуре и работе других приложений. SOA обеспечивает независимость приложений и возможность замены той или иной системы в организации без необходимости внесения существенных изменений в системы, которые с ней взаимодействуют.

Цель сервис-ориентированной архитектуры — организация строго определенного взаимодействия между отдельными независимыми программными модулями. Каждый модуль выполняет функции (часто говорят «предоставляет сервисы») в интересах других программных модулей или людей. Ключевым концептуальным моментом SOA является то, что предоставляемые сервисы независимы: сервис не знает ровным счетом ничего об обращающемся к нему приложении, равно как для приложения реализация сервиса является «черным ящиком». Сервис-ориентированная архитектура может быть реализована с помощью различных технологий, включая веб-сервисы, обмен сообщениями, RESTful API[1], и т. п. Сами сервисы обычно реализуются как интерфейсы прикладного программирования (API), доступные для

[1] REST (*сокр.* от *англ.* Representational State Transfer — передача состояния представления) — набор архитектурных принципов построения сервис-ориентированных приложений. RESTful — прилагательное, употребляющееся по отношению к сервисам, которые соответствуют принципам REST. — *Примеч. науч. ред*

вызова прикладным системам или пользователям (потребителям). Регистрационная запись точно определенного API описывает доступные опции, необходимые параметры запроса и выдаваемую в ответ на обращение информацию.

Сервисы данных, обеспечивающие обработку запросов на добавление, удаление, обновление или выборку данных, регистрируются в каталоге доступных сервисов. Для достижения корпоративных целей в части обеспечения масштабируемости информационных систем (поддержка интеграции всех приложений организации без необоснованно высоких затрат ресурсов) и повторного использования (использования одних и тех же сервисов для работы с конкретным видом данных всеми потребителями данных этого вида) требуется внедрение строгой и четкой модели руководства, охватывающей разработку и регистрацию сервисов и API. Прежде чем разрабатывать новый сервис данных, необходимо убедиться, что для обработки предполагаемых запросов нельзя использовать какой-то из уже имеющихся сервисов. Кроме того, всякий новый сервис должен разрабатываться с учетом как можно более широкого круга требований, чтобы помимо решения текущей задачи он мог впоследствии многократно использоваться и для решения других похожих задач.

1.3.7.6 ОБРАБОТКА СЛОЖНЫХ СОБЫТИЙ (CEP)

Обработкой событий (event processing) называют метод отслеживания и анализа (обработки) потоков информации (данных) о происходящем (событиях) с целью формирования заключений. Обработка сложных событий (Complex Event Processing, CEP) предполагает объединение данных из множества источников для выявления значимых событий (например, угроз или возможностей) с целью прогнозирования поведения наблюдаемых объектов и автоматического реагирования в режиме реального времени (примером такой реакции является предложение потенциальным покупателям подходящих товаров). Для управления обработкой событий и определения последовательности действий задаются правила.

Организации могут использовать довольно сложные модели и алгоритмы обработки событий с целью прогнозирования поведения и/или развития событий и автоматического запуска упреждающих, встречных или ответных действий в режиме реального времени. События (например, выявление потенциальных клиентов, переходы по ссылкам, заказы, обращения в службу клиентской поддержки и т. д.) могут происходить в рамках организации на различных ее уровнях. Кроме того, потоки событий могут включать новостные сообщения, текстовые уведомления, сообщения в соцсетях, данные с лент биржевых котировок, информацию о пробках на дорогах, прогнозы погоды и т. д. Событие может быть определено и как изменение состояния объекта наблюдения: например, по достижении заданного порогового значения времени, температуры или какой-либо иной измеримой характеристики.

Реализация технологий CEP сопряжена с некоторыми трудностями в части сбора данных. Во многих случаях частота событий такова, что нереалистично рассчитывать на возможность получения дополнительных данных, требующихся для их оперативной интерпретации. Следовательно, для эффективной динамической обработки потоков входных данных обязательно

наличие в памяти приложения, реализующего CEP, шаблонных заготовок блоков данных, соответствующих пусковым событиям различных типов.

Поддержка CEP предполагает наличие среды, которая позволяла бы интегрировать массу разнородных данных. Из-за колоссальных объемов и структурного разнообразия данных, обычно используемых для прогнозирования, обработка сложных событий часто бывает тесно связана с областью больших данных. Во многих случаях она требует применения технологий, обеспечивающих сверхмалую задержку, — например, обработки входящих потоков данных в режиме реального времени или баз данных в оперативной памяти (см. главу 14).

1.3.7.7 ФЕДЕРАЛИЗАЦИЯ И ВИРТУАЛИЗАЦИЯ ДАННЫХ

Данные из разрозненных хранилищ могут сводиться воедино и обобщаться и без физической интеграции. Федерализация заключается в обеспечении доступа к объединению отдельных хранилищ данных, которые могут иметь различную структуру. Виртуализация обеспечивает поддержку распределенных баз данных, а также возможность доступа к множеству хранилищ разнородных данных как к единой базе данных (см. главу 6).

1.3.7.8 ДАННЫЕ КАК УСЛУГА (DAAS)

Программное обеспечение как услуга (Software-as-a-Service, SaaS) — модель поставки и лицензирования ПО, в рамках которой приложения предоставляют сервисы, но сами они и данные физически находятся в ЦОДах поставщика ПО, а не приобретающей лицензию организации. Аналогичные концепции теперь используются при предоставлении доступа к различным уровням информационно-вычислительной инфраструктуры в порядке абонентского обслуживания (ИТ как услуга, платформы как услуга, базы данных как услуга и т. п.).

В частности, понятие «данные как услуга» (Data-as-a-Service, DaaS) используется для описания ситуации, когда поставщик предоставляет организации, купившей лицензию, данные по запросам, что избавляет последнюю от необходимости хранить и вести эти данные в собственном ЦОДе. Распространенный пример — информация о ценных бумагах и их котировках (текущих и исторических), которая предлагается к продаже фондовыми биржами.

Хотя в прямом смысле данные как услуга подразумевают лишь предоставление платного доступа к данным различным клиентам в соответствующей отрасли, в последнее время понятие «услуга» стало всё чаще использоваться внутри организаций применительно к предоставлению доступа к корпоративным данным или сервисам данных различным функциям и системам, поддерживающим операционную деятельность. Сервисные организации публикуют каталоги предлагаемых услуг, информацию об уровнях обслуживания, расценки и т. п.

1.3.7.9 ИНТЕГРАЦИЯ НА ОСНОВЕ ОБЛАКА

Интеграция на основе облака (cloud-based integration), также известная под названием «интеграционная платформа как услуга» (Integration Platform-as-a-Service, IPaaS), — это форма интеграции систем, предоставляемая как облачный сервис, реализующий различные варианты

использования (use cases) данных, процессов, сервис-ориентированной архитектуры (SOA) и интеграции приложений.

До появления облачных вычислений интеграцию можно было четко разделить на внутреннюю (internal) и интеграцию «бизнес для бизнеса» (Business-to-Business, B2B). Внутренняя интеграция реализуется посредством развернутой на предприятии программной платформы промежуточного слоя (middleware platform), — как правило, с использованием корпоративной сервисной шины (ESB), обеспечивающей управление обменом данными между системами. Интеграция B2B осуществляется с помощью шлюзов EDI (Electronic Data Interchange — электронный обмен данными), а также сетей с дополнительными услугами (сетей с «добавленной стоимостью» — Value-Added Networks, VAN) или электронных торговых площадок.

С приходом SaaS-приложений сформировался спрос на интеграцию данных, которые физически находятся за пределами ЦОДа организации. Их объединение обеспечивается с помощью интеграции на основе облака. После появления решений такого типа многие из них были доработаны с целью использования для интеграции локальных приложений, функционирующих в организациях, а также в качестве шлюзов EDI.

Интеграционные решения на основе облака обычно реализуются как SaaS-приложения и развертываются в ЦОДах поставщиков услуг, а не организаций — владельцев данных, которые подлежат интеграции. Интеграция на основе облака включает взаимодействие с SaaS-приложениями с целью объединения данных с использованием SOA-сервисов (см. главу 6).

1.3.8 Стандарты обмена данными

Стандарты обмена данными задают формальные правила для определения структуры элементов данных. Стандарты такого рода разрабатываются как Международной организацией по стандартизации (ISO), так и во многих отраслях. Спецификация обмена данными — это общая модель (используемая организацией или группой, отвечающей за обмен данными), стандартизирующая формат, в котором осуществляется распространение данных. Шаблон, применяемый для обмена, определяет структуру, в которую должна преобразовать данные каждая из взаимодействующих организаций. Необходимо осуществить мэппинг данных в соответствии со спецификацией обмена.

Хотя разработка и согласование единого для всех участников информационного взаимодействия формата сообщений — задача трудоемкая, но, обеспечив единообразие структуры обмена, мы значительно упрощаем обеспечение интероперабельности данных в организации, снижаем затраты на ее поддержку и предоставляем возможности для лучшего понимания данных.

В США для обмена документами и транзакционными данными между правительственными учреждениями разработана Национальная модель обмена информацией (National Information Exchange Model, NIEM). Цель ее создания — обеспечение общего однозначного понимания смысла данных отправителем и получателем. Соблюдение требований NIEM гарантирует четкую и не допускающую вольной трактовки интерпретацию базового набора сведений, что обеспечивает единообразие восприятия данных самыми различными сообществами и, следовательно, интероперабельность.

Модель NIEM для определения схем и представления элементов использует язык XML, что максимально упрощает структуру данных и их понимание за счет применения простых, но тщательно определенных правил синтаксиса.

2. ПРОВОДИМЫЕ РАБОТЫ

Обеспечение интеграции и интероперабельности данных (DII) подразумевает, что данные всегда должны оказываться в нужное время в нужном месте и в требуемой форме. Работы по интеграции данных (разработке интеграционного решения) соответствуют фазам жизненного цикла разработки систем. Они начинаются с планирования и далее проходят фазы проектирования, разработки, тестирования и внедрения. По завершении внедрения интегрированные системы требуют надлежащего управления, мониторинга и совершенствования.

2.1 Планирование и анализ

2.1.1 Определение требований к интеграции и жизненному циклу данных

Определение требований к интеграции данных начинается с осмысления бизнес-задач организации, а также потребностей в данных и информационных технологиях, за счет которых эти задачи могут быть выполнены. Кроме того, необходимо тщательно собирать и учитывать требования всех нормативно-правовых актов и отраслевых регламентов, распространяющихся на данные, которые планируется использовать. Какие-то работы, возможно, потребуется вести в режиме строгой конфиденциальности из-за характера данных, и лучше все подобные требования знать заранее во избежание проблем в будущем. Требования также могут учитывать политику организации в отношении сроков сохранения данных (data retention) и других аспектов их жизненного цикла. Требования в отношении сроков сохранения часто сильно зависят от предметной области и вида данных.

Обычно требования к интеграции и жизненному циклу данных определяются бизнес-аналитиками, распорядителями данных и архитекторами, работающими в различных областях деятельности организации, включая ИТ, которые заинтересованы в том, чтобы данные оказались в нужных местах, в требуемых форматах и были интегрированы с другими данными. Требования определяют вид модели взаимодействия при обеспечении DII (DII interaction model), которая, в свою очередь, обусловливает выбор технологий и сервисов, необходимых для ее реализации с целью выполнения требований.

В процессе определения требований создаются и выявляются полезные метаданные. Они подлежат управлению на протяжении всего жизненного цикла данных, начиная с обнаружения источников и заканчивая выполнением операций. Чем полнее и точнее метаданные организации, тем больше она способна управлять рисками и затратами, связанными с интеграцией данных.

2.1.2 Исследование данных

Исследование данных (data discovery) необходимо проводить перед проектированием. Цель исследования — определение потенциальных источников данных, которые могут быть использованы при выполнении работ по интеграции. Оно должно выявить, где данные могут быть получены и где они должны интегрироваться. Процесс исследования объединяет технический поиск, использующий инструменты, которые сканируют метаданные и/или реальное содержимое наборов данных организации, с экспертизой данных по той или иной предметной области (то есть интервьюированием профильных специалистов, работающих с данными).

Исследование также включает высокоуровневую оценку качества данных с целью определения их пригодности к использованию в рамках реализуемой интеграционной инициативы. Такая оценка требует не только детального анализа имеющейся документации и интервьюирования специалистов в предметных областях, но и проверки собранной информации на предмет ее соответствия реальным данным путем профилирования или других видов анализа (см. раздел 2.1.4). Почти во всех случаях выявляются расхождения между предполагаемым состоянием набора данных и результатами проверки.

В процессе исследования создается или дополняется реестр данных организации. Этот реестр должен вестись в репозитории метаданных. Следует рассматривать его ведение как стандартную часть работ по интеграции и обеспечивать своевременное обновление его содержимого в ходе проводимой деятельности, например при добавлении или удалении хранилищ данных или при изменении структуры документов.

Большинству организаций требуется интеграция данных из их внутренних систем. Однако некоторые интеграционные решения предусматривают получение данных от сторонних поставщиков. Объемы полезной информации — как бесплатной, так и предоставляемой поставщиками на платной основе — лавинообразно нарастают. Данные из внешних источников могут оказаться крайне полезными при их объединении с данными организации, — но лишь при условии тщательного планирования их приобретения и интеграции.

2.1.3 Документирование происхождения данных

В процессе исследования данных выявляется также информация о том, как данные перемещаются в организации. Эта информация может быть использована для документирования происхождения данных (имеется в виду только высокоуровневое описание): из какого источника они берутся или как генерируются; где внутри организации они перемещаются и изменяются; как и для чего используются (аналитика, принятие решений, запуск процессов). Детализированное описание происхождения данных может включать правила, в соответствии с которыми они изменяются, и сведения о частоте изменений.

Анализ происхождения данных иногда позволяет выявлять необходимость внесения изменений в документацию на эксплуатирующиеся системы. ETL-решения собственной разработки и другие унаследованные средства манипулирования данными должны быть документированы, чтобы организация имела возможность анализировать влияние любых изменений в потоках данных.

Анализ также позволяет выявить потенциальные возможности для оптимизации потоков данных. Например, может выясниться, что какой-нибудь ресурсоемкий программный модуль обработки данных можно заменить вызовом стандартной функции, а то и вовсе исключить его из процесса обработки за ненадобностью. Иногда при использовании старых инструментальных средств возникают ситуации, когда на последующих этапах обработки потока данных производится обратное преобразование, сводящееся к отмене изменений, внесенных ранее. Выявление и удаление подобных непроизводительных элементов может оказать существенную помощь в успешной реализации проекта и повышает способность организации эффективно использовать свои данные.

2.1.4 Профилирование данных

Успешная интеграция данных невозможна без понимания их содержания и структуры, для чего полезно использовать такой аналитический прием, как профилирование данных (data profiling). Реальные данные по структуре и содержанию всегда отличаются от нашего представления о них. Хорошо, если эти отличия несущественны; однако нередко они достаточно велики, чтобы свести к нулю все усилия по интеграции данных. Профилирование помогает командам по интеграции выявить такие расхождения и использовать полученные знания для принятия решений относительно оптимизации источников и подходов к проектированию. Если опустить этап профилирования данных, информация, которую следовало бы изначально учитывать при проектировании, не выявится, пока не начнется тестирование или эксплуатация.

Базовое профилирование данных включает анализ следующих аспектов:

- формат данных, определенный в описании структур данных и выявленный на основе реальных данных;
- заполнение полей данных, включая уровни наличия неопределенных и пустых значений, а также значений по умолчанию;
- фактические значения данных и степень их соответствия определенным наборам допустимых значений;
- паттерны и связи в наборе данных, такие как связанные поля и правила мощности связей;
- связи с другими наборами данных.

Однако, чтобы понять, насколько данные соответствуют требованиям конкретной интеграционной инициативы, требуется более обширное и глубокое профилирование потенциальных наборов — источников данных и целевых наборов. Профилирование и исходных и целевых наборов позволяет составить представление о том, каким образом следует преобразовать данные, чтобы обеспечить выполнение требований.

Одной из целей профилирования является оценка качества данных. В частности, для проведения оценки пригодности данных к использованию по конкретному целевому назначению требуется наличие четко сформулированных бизнес-правил и измеримых показателей

соответствия данных этим правилам. Для оценки точности данных нужно иметь эталонный набор для сравнения, данные в котором признаны точными. Такие наборы имеются далеко не всегда, поэтому оценка точности может быть невозможна, особенно в рамках усилий по профилированию.

Как и в случае высокоуровневого исследования данных, профилирование данных включает проверку соответствия предположений по поводу данных реальным данным. Результаты профилирования данных следует зафиксировать в репозитории метаданных, чтобы использовать их в будущих проектах. Следует также использовать информацию, полученную в процессе профилирования, для повышения точности метаданных (Olson, 2003; см. также главу 13).

Требование профилирования данных не должно вступать в противоречие с установленными в организации правилами информационной безопасности и внешними требованиями по защите конфиденциальных данных (см. главу 7).

2.1.5 Сбор и систематизация бизнес-правил

Бизнес-правила — критически важное подмножество требований. Бизнес-правило — это утверждение, которое определяет или ограничивает тот или иной аспект бизнес-процесса. Бизнес-правила устанавливаются для утверждения определенной деловой структуры и контроля/влияния на нее. Можно выделить четыре основные категории бизнес-правил:

- определения бизнес-терминов;
- взаимосвязи между различными бизнес-терминами;
- ограничения или предписываемые действия;
- производные правила.

Используйте бизнес-правила для поддержки различных функциональных элементов обеспечения интеграции и интероперабельности данных, в частности с целью:

- определения порядка доступа к данным в исходном и целевом наборах;
- маршрутизации потоков данных в организации;
- мониторинга операционных данных организации;
- определения пороговых значений и сигналов для автоматического запуска событий и/или выдачи предупреждений.

В области управления основными данными бизнес-правила включают правила соответствия, слияния, наследования и утверждения. Для архивирования и других процессов, связанных с использованием различного рода хранилищ данных, бизнес-правила также включают правила сохранения данных.

Процесс сбора бизнес-правил иногда называют также «пожинанием» (harvesting) или «добычей» (mining) бизнес-правил. Действительно, бизнес-аналитикам или распорядителям данных

приходится извлекать правила из имеющейся документации (сценариев использования, спецификаций, системного кода и т. п.) и/или проводить множество рабочих встреч и интервью с экспертами в предметных областях.

2.2 Проектирование решений по интеграции данных

2.2.1 Разработка архитектуры интеграционного решения

Решения по интеграции данных должны определяться как на корпоративном уровне, так и на уровне отдельных приложений (см. главу 4). Устанавливая общие корпоративные стандарты, организация тем самым экономит время на реализацию будущих отдельных решений, поскольку все необходимые экспертные оценки и согласования проводятся единожды и заранее. Подход к проектированию архитектуры в масштабах организации дает еще и денежную экономию при покупке лицензий на программное обеспечение за счет групповых скидок, а также снижение эксплуатационных расходов благодаря упрощению и согласованности решений. Наконец, совместимость операционных сред позволяет обеспечивать взаимную подстраховку, при необходимости перераспределять ресурсы и даже объединять их в совместно используемый пул.

При разработке интеграционного решения следует руководствоваться зафиксированными требованиями, но при этом стараться максимально использовать уже имеющиеся компоненты DII. Архитектура решения определяет методы и технологии, которые будут использоваться. Она должна включать: реестр задействованных структур данных (как существующих, так и вновь создаваемых; как для обеспечения постоянного хранения, так и промежуточных); описание оркестровки и интенсивности потоков данных; описание мер по обеспечению нормативно-правового соответствия и безопасности данных, а также средств решения возможных проблем; описание подходов к выполнению операций в части: резервного копирования и восстановления, обеспечения доступности, архивирования и контроля соблюдения сроков хранения данных.

2.2.1.1 ВЫБОР МОДЕЛИ ВЗАИМОДЕЙСТВИЯ

Определите, какая модель взаимодействия лучше всего соответствует требованиям — звездообразная с интеграцией в центре, «точка-точка», «публикация-подписка». Если требованиям удовлетворяет какое-либо уже внедренное решение, постарайтесь в максимальной степени обеспечить его повторное использование, насколько это возможно, — чтобы минимизировать затраты времени и усилий на новую разработку.

2.2.1.2 ПРОЕКТИРОВАНИЕ СЕРВИСОВ ДАННЫХ ИЛИ ШАБЛОНОВ ОБМЕНА

Спроектируйте новые или используйте существующие интеграционные потоки. Новые сервисы данных должны быть максимально совместимыми с похожими сервисами, которые уже имеются. При этом старайтесь всячески избегать создания множества почти идентичных сервисов и максимально используйте уже действующие, поскольку избыток сервисов усложняет поиск проблем и поддержку. Если существующий поток данных можно модифицировать таким образом, чтобы

он дополнительно поддерживал и новые потребности, то может оказаться целесообразным использовать эту возможность вместо того, чтобы создавать новый сервис.

Любое проектирование обмена данными должно изначально вестись с учетом отраслевых стандартов или иных уже существующих шаблонов обмена. При внесении изменений в действующие шаблоны старайтесь, насколько это возможно, ориентироваться на общее применение, чтобы впоследствии их можно было использовать и в других системах. Архитектура, при которой с каждым процессом обмена сопоставлен специфический шаблон, приводит к воспроизведению ровно тех же проблем, которые присущи соединению «точка-точка».

2.2.2 Моделирование хабов данных, интерфейсов, сообщений и сервисов данных

Структуры данных, необходимые для обеспечения интеграции и интероперабельности, включают компоненты интеграционных решений, в которых данные хранятся постоянно (например, хабы для управления основными данными, хранилища и витрины данных, хранилища операционных данных), а также структуры временного хранения, используемые для перемещения данных или их преобразования (например, интерфейсы, структуры для вывода сообщений и канонические модели). И те и другие должны быть смоделированы (см. главу 5).

2.2.3 Мэппинг исходных структур данных на целевые

Практически любое интеграционное решение включает преобразование данных из исходной структуры (источник) в целевую (получатель). Мэппинг определяет правила такого преобразования.

Для каждого преобразуемого атрибута данных спецификация мэппинга определяет:

- технические форматы источника и получателя;
- требуемые операции по преобразованию в каждой промежуточной (а также в целевой) точке временного хранения на пути от источника к получателю;
- правила заполнения полей атрибута в каждой промежуточной (а также в целевой) структуре;
- операции по преобразованию значений атрибута (если такое преобразование требуется) — например, с помощью таблицы сопоставления исходных значений и значений в целевой структуре;
- требуемые вычисления (если они нужны).

Преобразование может производиться по расписанию в пакетном режиме или запускаться автоматически по мере регистрации определенных событий в режиме реального времени. Результаты преобразований могут либо физически сохраняться в целевом формате, либо отображаться в этом формате в виде виртуального представления данных.

2.2.4 Проектирование оркестровки данных

Потоки данных в интеграционном решении должны быть спроектированы и документально оформлены. Оркестровка данных как раз и представляет собой описание (шаблон — pattern)

потоков данных от «старта» до «финиша», включая промежуточные шаги, требуемые для выполнения преобразования и/или транзакции.

Оркестровка пакетной интеграции данных должна также предоставлять сведения о частоте перемещения и преобразования данных. Отдельные задачи, с помощью которых реализуется пакетная интеграция, обычно описываются в планировщике (scheduler), который и запускает их в указанное время, с указанной периодичностью или по наступлении заданного события. Расписание задач может включать множество взаимозависимых шагов.

Оркестровка интеграции данных в режиме реального времени, как правило, предусматривает запуск задач по событию — например, добавлению или обновлению данных. Такая оркестровка обычно сложнее, чем в пакетном режиме, и реализуется посредством применения многих инструментов. Она по своей природе не может быть линейной.

2.3 Разработка решений по интеграции данных

2.3.1 Разработка сервисов данных

Сервисы доступа, преобразования и выдачи данных разрабатываются в соответствии со спецификациями и выбранной моделью взаимодействия. Чаще всего для реализации интеграционных решений, таких как преобразование данных, управление основными данными, ведение хранилищ данных и т. п., используются специальные инструменты или поставляемые крупными поставщиками программные пакеты. Применение во всех указанных решениях единых в рамках всей организации взаимосогласующихся инструментов и стандартных пакетов значительно упрощает техническую поддержку и снижает операционные расходы.

2.3.2 Разработка потоков данных

Интеграционные или ETL-потоки данных обычно разрабатываются с помощью специализированного лицензионного программного обеспечения. Состав же последних варьируется в зависимости от разработчика. Пакетные потоки данных разрабатываются с помощью планировщика (как правило, это стандартный корпоративный планировщик), который управляет очередностью и частотой выполнения отдельных функциональных элементов интеграционного решения, а также отслеживает их зависимости.

Требования по интероперабельности могут предусматривать разработку мэппинга или точек координации (coordination points) между хранилищами данных. Некоторые организации используют корпоративную сервисную шину (ESB) для подписки пользователей на созданные или измененные данные, а другие приложения — для публикации изменений. ESB постоянно опрашивает приложения на предмет наличия у них данных для публикации и доставляет им свежие данные, на которые они подписаны.

В случае разработки потоков данных для интеграции в режиме реального времени необходимо также предусмотреть применение средств мониторинга событий, запускающих на выполнение сервисы получения, преобразования или публикации данных. Подобные схемы обычно

реализуются на базе одной или нескольких лицензионных технологий, поэтому лучше всего внедрять решение, способное осуществлять сквозное управление операциями в каждой из применяемых технологических сред.

2.3.3 Выработка подхода к миграции данных

При внедрении новых приложений, а также при выводе из эксплуатации или объединении существующих возникает необходимость в переносе данных. Этот процесс включает их преобразование в формат приложения-получателя. В том или ином объеме миграция присутствует в любых проектах по разработке программного обеспечения, пусть даже речь идет всего лишь о заполнении таблиц справочными данными, требующимися новому приложению. Миграция не является одноразовым процессом, поскольку обычно ее нужно провести во время фазы тестирования, а затем — при окончательном внедрении.

Проекты миграции данных часто недооценивают и, как следствие, «недопроектируют» из-за того, что программистов просят просто перенести данные в новую среду; они не привлекаются к анализу и проектированию операций, необходимых для интеграции данных. Данные, перенесенные без предварительного анализа, часто по некоторым аспектам отличаются от данных, поступивших в рамках стандартных процессов обработки. Иногда такие данные оказываются просто непригодными для использования приложением. Профилирование данных ключевых рабочих приложений обычно позволяет выявить массивы информации, которые были перенесены из среды старых программ, функционирующих под управлением операционных систем предыдущих версий. Такие массивы обычно не соответствуют современным стандартам представления данных в отличие от массивов, созданных новыми приложениями (см. главу 6).

2.3.4 Выработка подхода к публикации

Системы, в которых создаются или ведутся критически важные данные, должны сделать их доступными другим системам организации. Публикация новых или изменившихся данных, которые требуются другим системам (прежде всего хабам данных или корпоративным шинам данных), должна осуществляться либо сразу же после изменения данных (управление на основе событий), либо с установленной периодичностью.

Оптимальной считается практика определения общих для всех систем форматов сообщений (каноническая модель) по каждому виду данных, используемому в организации, и уведомление потребителей (приложений и/или пользователей), подписавшихся на интересующие их данные, о внесенных изменениях.

2.3.5 Разработка потоков обработки сложных событий

При разработке решений по обработке сложных событий необходимо:

- подготовить исторические данные о физических лицах, организациях, продуктах, или рынках и произвести предварительное информационное наполнение предиктивной модели;

- обеспечить обработку потока данных в режиме реального времени, для того чтобы произвести информационное наполнение предиктивной модели и выявить значимые события (возможности или угрозы);
- обеспечить автоматический запуск выполнения действий в ответ на значимые события.

Подготовка и предварительная обработка исторических данных, используемых в предиктивной модели, могут производиться либо в ночное время в пакетном режиме, либо в режиме, близком к реальному времени. Обычно некоторое информационное наполнение предиктивной модели может быть произведено до начала обработки событий (например, на основании имеющихся сведений о том, какие продукты или услуги практически всегда заказываются покупателями в комплекте с основным предметом покупки).

Некоторые потоки обработки в режиме реального времени запускают выполнение соответствующих действий в ответ на каждое поступающее событие (например, добавление товара в корзину); другие ориентированы на выявление особо значимых событий, требующих вмешательства (например, проведение операций, подозрительно напоминающих попытку списания средств с кредитной карты мошенническим путем).

Действия в ответ на выявленные значимые события могут варьироваться от простейших (отправка сообщения с просьбой ввести код подтверждения операции) до крайне сложных (автоматическая выдача серии приказов о развертывании вооруженных сил).

2.3.6 Ведение метаданных, необходимых для обеспечения DII

Как отмечалось ранее (см. раздел 2.1), в процессе разработки решений по обеспечению интеграции и интероперабельности данных организации создают и выявляют полезные метаданные. Метаданные требуют тщательного учета и управления с целью обеспечения правильного понимания данных в информационных системах, а также для того, чтобы избежать необходимости их повторного выявления при последующих разработках. Надежные метаданные повышают способность организации управлять рисками, минимизировать издержки и получать отдачу от имеющихся данных.

Обязательно документируйте структуры данных всех систем, участвующих в процессах интеграции в качестве источников, получателей или пунктов их промежуточного хранения и обработки. Включайте в документацию как бизнес-, так и технические определения (структура, формат, размер), а также описания всех преобразований данных на пути из одних мест хранения в другие. Вне зависимости от того, где хранятся метаданные (в документах или в репозитории метаданных), они не могут быть изменены без проведения процедуры согласования со всеми заинтересованными сторонами, представляющими как бизнес-подразделения, так и технические службы.

Большинство поставщиков ETL-инструментов дополняют свои репозитории метаданных специальной функциональностью по поддержке процессов руководства и распоряжения метаданными. Если же репозиторий метаданных применяется в качестве одного из инструментов

обеспечения операционной деятельности, тогда он может регистрировать даже операционные метаданные о том, откуда данные были скопированы и где они были преобразованы в процессе их перемещения между системами.

Особую важность для DII-решений имеет реестр SOA (SOA registry), который предоставляет контролируемый доступ к постоянно обновляющемуся каталогу имеющихся сервисов для работы с данными и приложениями.

2.4 Внедрение и мониторинг

Активация сервисов данных осуществляется после полного завершения работ по их разработке и тестированию. Обработка данных в режиме реального времени требует мониторинга проблемных ситуаций (проводимого также в режиме реального времени). Следует определить параметры, указывающие на появление проблем в обработке, и реализовать непосредственную автоматическую передачу сообщений о проблемных ситуациях всем заинтересованным сторонам. Должен быть организован как автоматический, так и автоматизированный (с участием специалиста-оператора) мониторинг, особенно в сложных и рискованных случаях, когда автоматическое инициирование выполнения действий в ответ на событие чревато непоправимыми последствиями. Например, известны прецеденты, когда сбои при автоматическом исполнении алгоритмов обеспечения безопасности финансовых сделок приводили к обрушению целых рынков или банкротству компаний.

Информационное взаимодействие должно поддерживаться (и контролироваться с помощью средств мониторинга) на том же уровне обслуживания, который предусмотрен для наиболее требовательных целевых приложений или потребителей данных.

3. ИНСТРУМЕНТЫ

3.1 Программный комплекс для преобразования данных / ETL-инструмент

Программный комплекс («движок» — engine[1]) для преобразования данных (или ETL-инструмент) — основной инструмент в наборе интеграционных программных средств, являющийся главным подспорьем при выполнении корпоративной программы интеграции данных. Такие комплексы обычно поддерживают как непосредственное выполнение операций по преобразованию данных, так и их проектирование.

Существуют чрезвычайно сложные инструменты для разработки и выполнения ETL-операций (физически или виртуально) как в пакетном режиме, так и в режиме реального времени.

[1] Под термином «engine» (мотор, двигатель) в сфере ИТ обычно понимается выделенная часть программного кода (программа / часть программы / комплекс программ / библиотека), которая реализует набор базовых функций, необходимых для решения конкретной прикладной задачи (среди специалистов, как правило, употребляется жаргонизм «движок»). В данном издании этот термин чаще всего переводится как «программный комплекс». — *Примеч. науч. ред.*

Интеграционные решения для разового преобразования данных, перемещаемых из одного места в другое по схеме «точка-точка», часто реализуются с помощью индивидуального программирования (custom coding). Решения корпоративного уровня обычно требуют применения инструментов, обеспечивающих преобразование и перемещение данных стандартными способами, используемыми в рамках всей организации.

Основные соображения при выборе программного комплекса для преобразования данных включают рассмотрение вопроса о необходимости поддержки обработки как в пакетном режиме, так и в режиме реального времени, а также вопроса о необходимости охвата процессами интеграции неструктурированных данных помимо структурированных, поскольку самые зрелые из существующих инструментов предназначены только для пакетной обработки структурированных данных.

3.2 Сервер виртуализации данных

Программные комплексы для преобразования данных, как правило, извлекают, преобразуют и загружают данные на физическом уровне, в то время как серверы виртуализации данных позволяют выполнять эти операции виртуально и при этом объединять структурированные данные с неструктурированными. Входом для сервера виртуализации часто служит хранилище данных (data warehouse), но полностью заменить хранилище данных в корпоративной информационной архитектуре такой сервер не может.

3.3 Корпоративная шина данных (ESB)

Корпоративной шиной данных (ESB) называют как модель программной архитектуры, так и разновидность ориентированного на передачу сообщений промежуточного программного обеспечения. ESB предназначена для реализации обмена сообщениями в режиме, близком к реальному времени, между гетерогенными хранилищами данных, приложениями и серверами, функционирующими в одной и той же организации. Большинство внутренних интеграционных решений, требующих согласования данных чаще, чем раз в сутки, строятся на основе именно этой архитектуры и технологии. Обычно ESB используется в асинхронном режиме, обеспечивающем свободный поток данных, однако в некоторых ситуациях могут требоваться и решения со строгой синхронизацией.

Корпоративная шина данных реализует очереди входящих и исходящих сообщений для каждой из систем, участвующих в информационном обмене и подключенных к шине с помощью адаптера или агента. Центральный процессор, выполняющий операции ESB, обычно функционирует на сервере, который отделен от остальных обменивающихся данными систем. Процессор отслеживает, какие системы подписаны на те или иные виды сообщений. Он непрерывно опрашивает каждую систему — участника обмена на предмет наличия исходящих сообщений, а также помещает входящие сообщения в очередь для сообщений, на которые система подписана, и сообщений, которые адресованы непосредственно ей.

Такая модель обработки данных называется обработкой в режиме, «близком к режиму реального времени» (near real-time), поскольку задержка доставки сообщения от системы-отправителя

системе-получателю может составлять не более нескольких минут. Эта модель является слабо связанной, то есть в ней система, отправившая данные, продолжает работу, не дожидаясь подтверждения получения и обновления данных от системы-получателя.

3.4 Программный комплекс для управления бизнес-правилами

Многие интеграционные решения зависят от бизнес-правил. Будучи важной категорией метаданных, бизнес-правила могут использоваться как при осуществлении самой простой интеграции, так и для реализации решений, включающих обработку сложных событий, которая позволяет организации реагировать на значимые события в режиме, близком к реальному времени. С помощью программного комплекса для управления бизнес-правилами пользователи, не являющиеся специалистами в сфере ИТ, могут самостоятельно управлять правилами, реализованными с помощью ПО. Это крайне полезный инструмент, предоставляющий возможность модификации созданного решения и его дальнейшего развития ценой минимальных затрат. Разработанные предиктивные модели могут быть скорректированы без внесения изменений в программный код. Например, модели для предварительного отбора товаров, которые должны заинтересовать покупателя, достаточно легко определяются с помощью описания бизнес-правил, не требуя программирования.

3.5 Инструменты моделирования данных и процессов

Инструменты моделирования данных следует использовать для проектирования не только целевых, но и промежуточных структур данных, необходимых для реализации решений по интеграции. Структура сообщений, которые перемещаются между системами или организациями и, как правило, не сохраняются, тем не менее подлежит моделированию. Потоки данных и потоки обработки сложных событий также должны быть спроектированы.

3.6 Инструменты профилирования данных

Профилирование данных предполагает статистический анализ содержимого информационных массивов с целью получения более полного представления о формате, полноте, согласованности, достоверности, актуальности и структуре данных. Все проекты по созданию решений в области обеспечения интеграции и интероперабельности должны предусматривать оценку потенциальных источников и получателей с целью определения пригодности существующих данных для использования в рамках проектируемого решения. Поскольку большинство проектов интеграции подразумевают работу с информационными массивами значительных объемов, наиболее эффективным подходом к проведению такого анализа является применение инструментов профилирования данных (см. раздел 2.1.4 и главу 13).

3.7 Репозиторий метаданных

Репозиторий метаданных содержит информацию о данных организации с описанием их структуры, содержимого и бизнес-правил управления данными. В ходе интеграционного проекта (с

целью документирования сведений о технической структуре и применении в бизнесе исходных, преобразуемых и конечных данных) может быть использован один или несколько репозиториев.

Правила в отношении процессов преобразования, отслеживания происхождения и обработки данных, используемые инструментами интеграции, обычно также хранятся в репозитории метаданных, равно как и инструкции по обеспечению выполнения процессов согласно расписанию (условия для запуска, периодичность и т. п.).

Каждый инструмент обычно имеет собственный репозиторий метаданных. Пакеты инструментальных средств, поставляемых одним и тем же разработчиком программного обеспечения, могут использовать для хранения метаданных общий репозиторий. Для консолидации всех метаданных, создаваемых и поддерживаемых в различных операционных инструментах, какой-либо из имеющихся репозиториев метаданных может быть выделен в качестве центрального (см. главу 12).

4. МЕТОДЫ

Отдельные наиболее важные методы проектирования решений в области интеграции данных описаны в разделе «Основные понятия и концепции» в начале этой главы. Ключевыми целями являются обеспечение слабой связанности приложений, ограничение количества разрабатываемых интерфейсов и сложности управления за счет применения модели взаимодействия по схеме звезды с интеграцией в центре, а также создание стандартных (или канонических) интерфейсов.

5. РЕКОМЕНДАЦИИ ПО ВНЕДРЕНИЮ

5.1 Оценка готовности / Оценка рисков

В любой организации всегда в том или ином виде присутствуют решения в области DII. Поэтому необходимо проводить оценку готовности/рисков в отношении внедрения корпоративных интеграционных инструментов, а также расширения возможностей по обеспечению интероперабельности.

Внедрение *корпоративных* решений в области интеграции данных обычно экономически оправдано, поскольку позволяет обеспечить совместимость одновременно для целого ряда систем. Следует осуществлять интеграционные проекты по поддержке обмена данными именно между многими системами и организациями, а не просто внедрять интеграционное решение на первом попавшемся участке.

Многие организации тратят время на переделку и доработку имеющихся решений, вместо того чтобы проводить работы по интеграции, которые могут принести дополнительные выгоды. Основное внимание следует уделять проектам, которые должны обеспечить интеграцию данных там, где она до сих пор отсутствует (или реализована в недостаточной степени), вместо того чтобы заменять удовлетворительно работающие средства единым корпоративным решением в масштабах организации.

В рамках определенных проектов бывает целесообразным построение интеграционного решения, обеспечивающего потребности конкретного приложения, например хранилища данных или хаба для управления основными данными. В таких случаях любое дополнительное направление применения интеграционного решения повышает отдачу от сделанных в него инвестиций, потому что его использование по прямому назначению так или иначе уже оправданно.

Команды по поддержке приложений часто предпочитают управлять интеграционными решениями на локальном уровне. Им кажется, это будет проще и дешевле, чем развертывать корпоративное решение. Поставщики программного обеспечения, которое используется такими командами, также крайне заинтересованы в применении средств интеграции, которые они продают. Следовательно, необходимо, чтобы внедрение корпоративной программы интеграции данных было поддержано руководителями, которые обладают достаточным влиянием на процессы проектирования решений и закупки технологий, — например, корпоративным ИТ-архитектором. Кроме того, могут потребоваться дополнительные меры по поощрению включения прикладных систем в корпоративную программу интеграции. Сюда можно отнести позитивное стимулирование (например, централизованное финансирование проектов по интеграции) и негативное стимулирование (например, запрет на согласование применения новых альтернативных технологий интеграции данных).

В проектах разработки, предусматривающих внедрение новых технологий интеграции данных, зачастую всё внимание фокусируется на самих технологиях и упускаются из виду бизнес-цели. Поэтому необходимо принимать меры по обеспечению соответствия разрабатываемых проектных решений бизнес-целям и требованиям, в том числе и путем включения в проектные команды по всем проектам участников, ориентированных на интересы бизнеса и использование приложений, а не одних лишь экспертов в области интеграции данных.

5.2 Организационные и культурные изменения

Организации должны прежде всего определить, будет ли управление внедрением решений в области интеграции данных осуществляться централизованно, или же ответственность за управление будет распределена между командами по внедрению отдельных приложений. Локальным командам более понятны те данные, с которыми работают их приложения. Централизованные команды могут обладать более глубокими знаниями инструментов и технологий. Многие организации формируют у себя Центр компетенции (Center of Excellence), специализирующийся на проектировании и развертывании корпоративных решений по интеграции данных предприятия. Центр компетенции и локальные команды взаимодействуют с целью разработки оптимальных

вариантов подключения приложений к корпоративному решению по интеграции данных. Локальная команда несет основную ответственность за управление решением и устранение проблем, эскалируя их центру компетенции, если это необходимо.

Решения по интеграции данных часто воспринимаются как чисто технические, однако для успешного получения от них ощутимой отдачи они должны разрабатываться на основе глубокого знания специфики бизнеса. Деятельность по анализу и моделированию данных должна осуществляться лицами, ориентированными прежде всего на интересы бизнеса. Разработка канонической модели сообщений (или согласованного стандарта по обеспечению совместного доступа к данным в организации) требует привлечения значительных ресурсов, включая как специалистов, необходимых для моделирования бизнеса, так и специалистов в области ИТ. Все проектные решения (и их изменения) в части мэппинга данных должны проверяться экспертами в предметных областях, представляющих каждую из участвующих в обмене систем.

6. РУКОВОДСТВО DII

Проектные решения по структуре сообщений, моделям данных и правилам их преобразования оказывают прямое влияние на способность организации использовать свои информационные массивы. Эти решения должны быть «управляемыми бизнесом» (business-driven). Несмотря на наличие множества технических соображений, которые следует учитывать при внедрении бизнес-правил, чисто технический подход к обеспечению DII чреват ошибками в мэппинге и преобразованиях данных. Такие ошибки могут возникнуть на различных этапах перемещения данных между внутренними и внешними (по отношению к организации) участниками информационного взаимодействия.

Определение правил, на основании которых осуществляются моделирование и преобразование данных, входит в сферу ответственности заинтересованных сторон, представляющих бизнес-подразделения (business stakeholders). Они же должны утверждать все вносимые в правила изменения. Правила должны фиксироваться в качестве метаданных и консолидироваться с целью проведения сквозного анализа в рамках организации. Создание и проверка достоверности предиктивных моделей (вместе с определением действий, выполнение которых должно быть инициировано в ответ на те или иные события) также относятся к бизнес-функциям.

Без уверенности в том, что проектные решения в области интеграции или интероперабельности данных будут соответствовать заданным требованиям и обеспечат надежное и безопасное выполнение реализованных функций, не может быть и речи о получении выгоды для бизнеса. Поэтому структура элементов системы руководства областью DII, обеспечивающая необходимый уровень доверия к внедряемым решениям, может быть довольно сложной и детализированной. Один из возможных подходов заключается в четком определении событий (исключений или критических инцидентов), требующих проведения проверок органами руководства. Каждое

такое событие должно быть привязано к проверке, которую назначает тот или иной орган руководства. Контроль событий может быть включен в жизненный цикл разработки систем (SDLC) как один из элементов принятия решения о переходе проекта на следующую стадию, в соответствии с моделью управления проектом «Стадия — Переход» (Stage-Gate), или учитываться в пользовательских историях (User Stories). Например, чек-лист соответствия архитектуры решения предъявляемым требованиям может включать вопросы следующего рода: «Используется ли ESB или другие специальные инструменты (там, где это возможно)?», «Был ли проведен анализ возможностей повторного использования имеющихся сервисов?» и т. п.

Могут быть использованы механизмы контроля, уже применяемые в рамках текущей практики руководства, такие как обязательное рецензирование моделей, аудит метаданных, контроль выходных результатов и обязательное согласование изменений правил преобразования.

Решения по интеграции операционных данных в режиме реального времени необходимо включить в соглашение об уровне обслуживания и планы обеспечения непрерывности бизнеса / аварийного восстановления. Для них должен быть определен такой же уровень обслуживания в части резервного копирования и восстановления, какой предусмотрен для наиболее критичных из систем, которые участвуют в обмене данными с помощью этих решений.

Следует ввести комплекс политик, направленных на обеспечение получения выгоды от применения корпоративного подхода к DII. Например, могут быть введены политики, требующие, чтобы обязательно соблюдались принципы SOA, чтобы проектирование новых сервисов инициировалось только после анализа имеющихся на предмет повторного использования и чтобы обмен данными между системами осуществлялся через шину служб предприятия.

6.1 Соглашения о совместном доступе к данным

Прежде чем создавать интерфейсы или реализовывать предоставление данных в электронном виде, следует разработать соглашение о совместном использовании данных (data sharing agreement) или меморандум о взаимопонимании (Memorandum Of Understanding, MOU). Документ должен включать основные требования в отношении ответственности и условий использования данных, которыми предполагается обмениваться, согласованные с соответствующими распорядителями бизнес-данных.

Соглашения о совместном использовании данных должны описывать возможные варианты использования данных и доступа к ним, ограничения на использование, а также ожидаемые уровни обслуживания, включая требования к времени непрерывной работы и времени отклика систем. Подобные соглашения особенно важны в строго регламентированных отраслях и в случае использования персональных или других защищенных данных.

6.2 DII и происхождение данных

При разработке DLL-решений полезно учитывать сведения о происхождении данных. Эти сведения также часто требуются потребителям при использовании данных, но еще большую важность

они приобретают, когда интегрируются данные нескольких организаций. Для того чтобы документирование информации о первоисточниках и перемещении данных гарантированно осуществлялось, требуется контроль со стороны функции руководства данными.

Соглашения о совместном использовании данных могут накладывать ограничения на их использование, поэтому для соблюдения этих соглашений необходимо знать, как данные перемещаются и где они хранятся. В последнее время появляются нормативно-правовые документы, регулирующие подобного рода вопросы (например, директива EC Solvency II) и требующие от организаций готовности отчитываться перед надзорными органами о том, откуда появились данные и как они изменялись при перемещении между различными системами.

Кроме того, информация о происхождении данных требуется при внесении изменений в информационные потоки. Она является критически важной частью метаданных интеграционного решения. Наличие возможности проследить происхождение данных (где данные используются и откуда они поступили) является необходимым условием проведения полноценного анализа влияния планируемых изменений структур, потоков или процедур обработки данных.

6.3 Метрики для оценки эффективности интеграции данных

Для оценки полноты внедрения решения по интеграции данных и отдачи от него используются показатели доступности, объемов, скорости, затрат и применимости.

- Доступность данных:
  - доступность запрошенных данных.
- Объемы и скорости обработки данных:
  - объемы переданных и преобразованных данных;
  - объемы проанализированных данных;
  - скорость передачи данных;
  - задержка между временем завершения обновления данных и временем выдачи обновленных данных;
  - задержка между временем начала события и временем начала выполнения ответных действий;
  - время, требуемое для предоставления доступа к новым источникам данных.
- Стоимость и сложность решения:
  - стоимость разработки и эксплуатации решений;
  - простота получения новых данных;
  - сложность решений и операций;
  - количество систем, использующих решение по интеграции данных.

7. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Aiken, P. and Allen, D. M. *XML in Data Management*. Morgan Kaufmann, 2004. Print.

Bahga, Arshdeep, and Vijay Madisetti. *Cloud Computing: A Hands-On Approach*. CreateSpace Independent Publishing Platform, 2013. Print.

Bobak, Angelo R. *Connecting the Data: Data Integration Techniques for Building an Operational Data Store (ODS)*. Technics Publications, LLC, 2012. Print.

Brackett, Michael. *Data Resource Integration: Understanding and Resolving a Disparate Data Resource*. Technics Publications, LLC, 2012. Print.

Carstensen, Jared, Bernard Golden, and JP Morgenthal. *Cloud Computing — Assessing the Risks*. IT Governance Publishing, 2012. Print.

Di Martino, Beniamino, Giuseppina Cretella, and Antonio Esposito. *Cloud Portability and Interoperability: Issues and Current Trend*. Springer, 2015. Print. Springer Briefs in Computer Science.

Doan, AnHai, Alon Halevy, and Zachary Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012. Print.

Erl, Thomas, Ricardo Puttini, and Zaigham Mahmood. *Cloud Computing: Concepts, Technology and Architecture*. Prentice Hall, 2013. Print. The Prentice Hall Service Technology Ser. from Thomas Erl.

Ferguson, M. *Maximizing the Business Value of Data Virtualization*. Enterprise Data World. Atlanta, GA: Dataversity, 2012. Print.

Giordano, Anthony David. *Data Integration Blueprint and Modeling: Techniques for a Scalable and Sustainable Architecture*. IBM Press, 2011. Print.

Haley, Beard. *Cloud Computing Best Practices for Managing and Measuring Processes for On-demand Computing*, Applications and Data Centers in the Cloud with SLAs. Emereo Publishing, 2008. Print.

Hohpe, Gregor and Bobby Woolf. *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*. Addison-Wesley Professional, 2003. Print.

Inmon, W. *Building the Data Warehouse*. 4th ed. Wiley, 2005. Print.

Inmon, W., Claudia Imhoff, and Ryan Sousa. *The Corporate Information Factory*. 2nd ed. Wiley 2001, Print.

Jamsa, Kris. *Cloud Computing: SaaS, PaaS, IaaS, Virtualization, Business Models, Mobile, Security and More*. Jones and Bartlett Learning, 2012. Print.

Kavis, Michael J. *Architecting the Cloud: Design Decisions for Cloud Computing Service Models (SaaS, PaaS, and IaaS)*. Wiley, 2014. Print. Wiley CIO.

Kimball, Ralph and Margy Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. 2nd ed. Wiley, 2002. Print.

Linthicum, David S. *Cloud Computing and SOA Convergence in Your Enterprise: A Step-by-Step Guide*. Addison-Wesley Professional, 2009. Print.

Linthicum, David S. *Enterprise Application Integration*. Addison-Wesley Professional, 1999. Print.

Linthicum, David S. *Next Generation Application Integration: From Simple Information to Web Services*. Addison-Wesley Professional, 2003. Print.

Loshin, David. *Master Data Management*. Morgan Kaufmann, 2009. Print.

Majkic, Zoran. *Big Data Integration Theory: Theory and Methods of Database Mappings, Programming Languages, and Semantics*. Springer, 2014. Print. Texts in Computer Science.

Mather, Tim, Subra Kumaraswamy, and Shahed Latif. *Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance*. O'Reilly Media, 2009. Print. Theory in Practice.

Reese, George. *Cloud Application Architectures: Building Applications and Infrastructure in the Cloud*. O'Reilly Media, 2009. Print. Theory in Practice (O'Reilly).

Reeve, April. *Managing Data in Motion: Data Integration Best Practice Techniques and Technologies*. Morgan Kaufmann, 2013. Print. The Morgan Kaufmann Series on Business Intelligence.

Rhoton, John. *Cloud Computing Explained: Implementation Handbook for Enterprises*. Recursive Press, 2009. Print.

Sarkar, Pushpak. *Data as a Service: A Framework for Providing Reusable Enterprise Data Services*. Wiley-IEEE Computer Society Pr, 2015. Print.

Sears, Jonathan. *Data Integration 200 Success Secrets — 200 Most Asked Questions On Data Integration — What You Need to Know*. Emereo Publishing, 2014. Kindle.

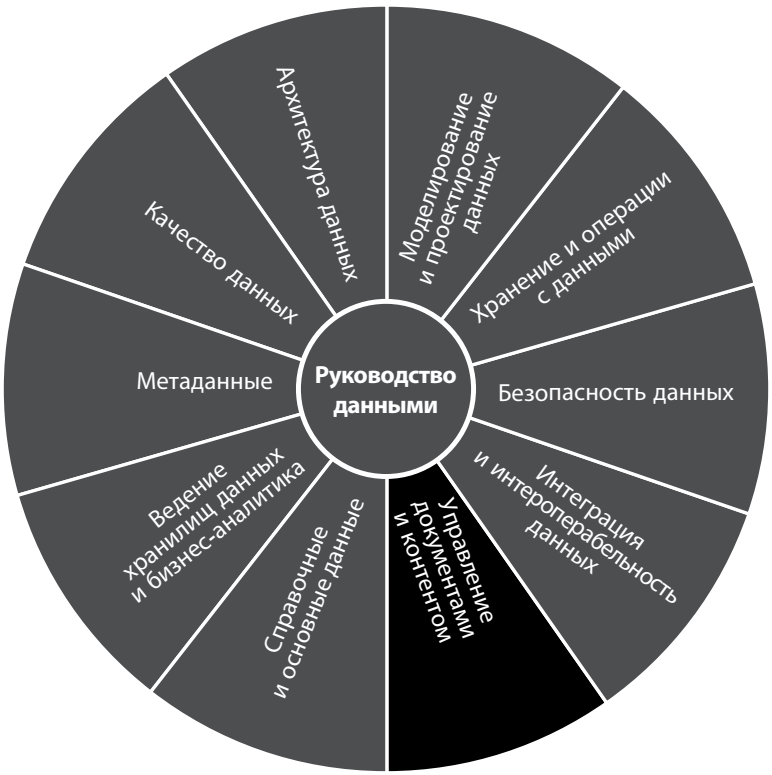
Sherman, Rick. *Business Intelligence Guidebook: From Data Integration to Analytics*. Morgan Kaufmann, 2014. Print.

U. S. Department of Commerce. *Guidelines on Security and Privacy in Public Cloud Computing*. CreateSpace Independent Publishing Platform, 2014. Print.

Van der Lans, Rick. *Data Virtualization for Business Intelligence Systems: Revolutionizing Data Integration for Data Warehouses*. Morgan Kaufmann, 2012. Print. The Morgan Kaufmann Series on Business Intelligence.

Zhao, Liang, Sherif Sakr, Anna Liu, and Athman Bouguettaya. *Cloud Data Management*. Springer, 2014. Print.

Управление документами и контентом



DAMA-DMBOK2 Data Management Framework

1. ВВЕДЕНИЕ

Управление документами и контентом (Document and Content Management) подразумевает наличие средств, позволяющих контролировать создание, регистрацию, хранение, защиту и использование самых разнородных данных и информации, доступ к которым не может быть организован

средствами традиционных реляционных СУБД[1]. Главная задача в этой области управления данными — обеспечение сохранности и целостности документов и других неструктурированных / полуструктурированных (semi-structured) данных, а также регулирование доступа к ним. В этом плане управление документами и контентом в общих чертах мало чем отличается от операционного управления реляционными базами данных. Однако помимо текущих задач в этой области есть и стратегические. Во многих организациях неструктурированные данные имеют прямое отношение к структурированным. Соответственно, и управленческие решения должны приниматься согласованно и применяться последовательно. Кроме того, как и другие типы данных, документы и неструктурированный контент требуют надежной защиты и контроля качества. А обеспечение информационной безопасности и качества таких данных, в свою очередь, невозможно без руководства, надежной архитектуры и налаженного управления метаданными.

1.1 Бизнес-драйверы

Главными бизнес-драйверами управления документами и контентом являются обеспечение соблюдения требований нормативно-правового регулирования, способность адекватно отвечать на запросы судебно-арбитражных и надзорных органов об электронном раскрытии информации, а также требования по обеспечению непрерывности бизнеса. Качественное управление записями (records) оказывает влияние на эффективность работы организаций. Хорошо структурированные веб-сайты с мощными поисковыми возможностями позволяют эффективно управлять онтологиями и другими структурами, максимально упрощающими клиентам и сотрудникам поиск нужного контента, повышая тем самым уровень их удовлетворенности.

Законы и регламенты требуют от организаций учета определенных видов деятельности и хранения соответствующих записей на протяжении установленных сроков. Кроме того, в большинстве организаций действуют внутренние правила, стандарты и методики ведения записей. К записям относятся как бумажные документы, так и информация, сохраняемая в электронном виде (Electronically Stored Information, ESI). Хорошее управление записями является обязательным для обеспечения бесперебойной и устойчивой работы любой организации. Оно же позволяет надежно доказывать свою правоту в случае каких-либо судебных или арбитражных разбирательств.

Процедура электронного раскрытия информации (e-discovery) позволяет выявлять электронные записи, которые могут быть предъявлены в судах различных инстанций и юрисдикций в качестве документальных доказательств. По мере развития технологий создания, хранения и использования данных объемы ESI растут по экспоненте. При этом все их нужно хранить и учитывать, поскольку невозможно заранее определить, что именно из массы накопленных вашей организацией электронных записей со временем пригодится в суде или будет истребовано надзорными органами.

[1] Число типов неструктурированных данных с начала нулевых годов выросло неимоверно, первопричиной чему — колоссальный рост емкости хранилищ и производительности систем регистрации и обработки цифровой информации. Понятие *неструктурированные данные* (*unstructured data*) при этом по-прежнему относится к любым данным, структура которых не описывается предопределенной моделью данных — будь то реляционная или некая иная модель.

Рисунок 71. Контекстная диаграмма: управление документами и контентом

Способность организации адекватно и оперативно откликаться на запросы в отношении e-discovery зависит от того, насколько предусмотрительно и активно велось управление накапливающимися электронными записями, такими как почта, чаты, веб-сайты и электронные документы, а также рабочими данными приложений и метаданными. Повсеместное распространение практики накопления, обработки и анализа гигантских массивов данных (большие данные) также стимулирует организации к более эффективному учету имеющихся в хранилищах массивов электронной информации, соблюдению требований относительно сроков хранения записей и созданию развитой структуры руководства данными.

Потребность в прибавке в эффективности по всем направлениям — лучший бизнес-драйвер совершенствования управления документами. Прогресс в области технологий управления документами позволяет организациям автоматизировать и оптимизировать документооборот, устранять дублирующие друг друга ручные операции, налаживать внутриорганизационное сотрудничество и внешние партнерские связи. Дополнительным преимуществом этих технологий выступает упрощение и ускорение поиска, доступа и публикации нужных документов. Наконец, они же способствуют предотвращению утери важных документов. Всё это крайне важно и для обеспечения электронного раскрытия информации, и для экономии денежных средств за счет высвобождения офисного пространства и снижения затрат на ведение и обработку документации.

1.2 Цели и принципы

Основными целями реализации лучших практик в сфере управления данными и контентом являются:

- обеспечение возможностей для эффективного накопления, получения и использования данных, сохраняемых в неструктурированных форматах;
- интеграция структурированных и неструктурированных данных;
- соблюдение действующего законодательства и соответствие ожиданиям клиентов.

Руководящие принципы управления данными и контентом таковы:

- Все сотрудники так или иначе отвечают за безопасность своей организации во избежание потенциальных неприятностей. Следовательно, каждый должен строго соблюдать правила и процедуры создания, получения, использования и ликвидации записей, имеющихся в распоряжении организации.
- Эксперты по обработке записей и контента должны привлекаться к выработке политики и планированию в качестве полноправных участников. Нормативно-правовое поле и лучшие практики в этой области сильно зависят от отрасли и страны юрисдикции организации.

Даже если в организации нет штатных специалистов по управлению записями, азам грамотного управления в этой области можно обучить всех сотрудников, чтобы они в полной мере понимали

важность проблемы. По завершении тренинга распорядители данных совместно с другими заинтересованными лицами могут выработать общий подход к эффективному управлению электронными документами и записями.

В 2009 году ARMA International, некоммерческая международная ассоциация специалистов по управлению записями и информацией, опубликовала Общепринятые принципы ведения записей (Generally Acceptable Recordkeeping Principles, GARP)[1], описывающие, как именно следует вести деловую документацию. Ассоциация также предлагает тщательно проработанную рамочную структуру руководства информацией и соответствующие метрики. Ниже приведены вводные положения каждого из принципов, а детальные разъяснения каждого из них можно найти на веб-сайте ARMA.

- **Принцип подотчетности.** В организации должен быть выбран глава программы руководства информацией из числа руководителей высшего звена, отвечающий за назначение ответственных, выработку политики, определение правил и процедур, которыми должны руководствоваться штатные сотрудники, и обеспечение контролируемого и проверяемого хода реализации программы.
- **Принцип целостности.** Программа руководства информацией должна быть выстроена таким образом, чтобы записи и информация, генерируемая организацией или имеющаяся у нее, в том числе и через сторонних провайдеров информационных услуг, имела разумные гарантии аутентичности и достоверности.
- **Принцип защиты.** Программа руководства информацией должна быть выстроена таким образом, чтобы обеспечивать надлежащий уровень защиты персональных данных, сведений личного характера и иной информации, не подлежащей разглашению.
- **Принцип соблюдения нормативно-правового соответствия.** Программа руководства информацией должна быть выстроена таким образом, чтобы обеспечивалось соблюдение всех применимых законов и подзаконных актов, а также внутриорганизационной политики и правил.
- **Принцип доступности.** Организация должна обеспечивать максимальное удобство, оперативность и эффективность доступа к нужной информации и точность ее предоставления для всех, кто заинтересован в ее получении и имеет право доступа к ней.
- **Принцип соблюдения сроков хранения.** Организация обязана хранить информацию на протяжении сроков, установленных действующими законами, подзаконными актами и прочими регламентирующими документами и требованиями надзорных органов, не допуская ни преждевременного уничтожения отчетности, ни сверхнормативного удержания чувствительных данных. Конкретные требования зависят от страны.
- **Принцип ответственного распоряжения.** Организация обязана обеспечивать безопасное и надлежащее использование имеющейся в ее распоряжении информации в соответствии с требованиями внутриорганизационной политики и всех применимых внешних законов, подзаконных актов и требований надзорных органов.

[1] ARMA International, ARMA Generally Accepted Recordkeeping Principles®.

- **Принцип прозрачности.** Организация обязана документировать политики, процессы, процедуры и меры по обеспечению надлежащего порядка делопроизводства, включая программу руководства информацией, понятным и доступным для сотрудников и заинтересованных сторон образом.

1.3 Основные понятия и концепции

1.3.1 Контент

Контент (*content*) — это содержимое документа, его информационное наполнение. Документ соотносится с контентом, как стакан соотносится с водой: документ — это контейнер, а контент — то, что в этом контейнере содержится. Под контентом понимают данные и информацию, размещенную внутри файла, документа или на веб-сайте. Контентом часто управляют исходя из степени концептуальной важности документов, в которых он содержится, а также в зависимости от типа или статуса документов. У контента также имеется свой жизненный цикл. В своей завершенной форме часть контента становится содержимым записей организации. Официальные записи требуют особого обращения по сравнению с прочим контентом.

1.3.1.1 УПРАВЛЕНИЕ КОНТЕНТОМ

Управление контентом (*content management*) включает процессы, методы и технологии упорядочения, классификации и структурирования информационных ресурсов с целью обеспечения возможности их хранения, публикации и многократного, многоцелевого использования.

Жизненный цикл контента может быть высокоактивным и предусматривать ежедневные изменения посредством контролируемых процессов создания, добавления или изменения информации. Существует также статичный контент, вовсе не меняющийся или изменяемый крайне редко и в минимальных пределах. В свою очередь, управление контентом может варьироваться от строго формализованного (в соответствии с жесткими правилами хранения, доступа, обращения и аудита, контролем сроков хранения и ликвидации) до полностью неформального добавления и изменения контента пользователями.

Особую важность управление контентом имеет для веб-сайтов и порталов, но принципы индексирования по ключевым словам и таксономической организации широко применяются при использовании самых разнообразных технологических платформ. Если управление контентом ведется в масштабах организации, такой подход называется управлением корпоративным контентом (Enterprise Content Management, ECM).

1.3.1.2 МЕТАДАННЫЕ КОНТЕНТА

Без метаданных управлять неструктурированной информацией невозможно, будь то традиционный контент и документы или то, что мы теперь называем «большими данными». Ни инвентаризировать, ни упорядочить контент без метаданных не получится. Метаданные неструктурированного контента могут отражать следующее.

- **Формат.** Часто формат определяет метод доступа к данным (например, метаданные могут включать электронный индексный указатель содержания массива неструктурированных данных).
- **Возможности поиска.** Существуют ли программные инструменты поиска, разработанные для работы с неструктурированными данными соответствующего вида.
- **Самодокументирование.** Являются ли метаданные самодокументируемыми (как в файловых системах). Если да, то разработка требуется минимальная, поскольку можно легко адаптировать для работы с контентом существующие программные средства.
- **Существующие шаблоны.** Можно ли применить или адаптировать к контенту существующие методы и шаблоны (например, каталоги библиотек).
- **Предметный указатель.** Очерчивает тематику контента, помогая пользователям ориентироваться в неструктурированных данных.
- **Требования.** Насколько осторожно и даже щепетильно следует относиться к открытию доступа к контенту (например, как в фармацевтической промышленности или ядерной энергетике[1]); в подобных случаях детальные метаданные на уровне контента могут оказаться незаменимым подтверждением выполнения всех требований, и, следовательно, может потребоваться специальное программное средство маркировки контента.

В целом ведение метаданных неструктурированной информации, по сути, сводится к управлению схемой перекрестных ссылок между различными локальными структурами хранения контента и официальным набором корпоративных метаданных. И менеджеры записей, и специалисты по ведению метаданных должны в полной мере отдавать себе и друг другу отчет в том, что в любой организации, включая их собственную, всегда существуют сложившиеся в долгосрочной перспективе и неявным образом внедренные методы многолетнего хранения и обращения документов, записей и прочего контента, но как-то реорганизовать и упорядочить эти устоявшиеся методы, если они кажутся неэффективными, зачастую бывает крайне сложно и дорого. В некоторых организациях даже идут по пути создания централизованной команды, занимающейся исключительно определением и сопровождением сложных структур перекрестных ссылок между относящимися к управлению записями индексами, таксономиями и версиями тезаурусов.

1.3.1.3 МОДЕЛИРОВАНИЕ КОНТЕНТА

Моделирование контента — процесс преобразования логических представлений о контенте в четко описанные структуры типов и атрибутов контента и типов данных со связями. Атрибут контента описывает какое-либо его отличительное свойство или характеристику. Тип данных

[1] В этих отраслях предприятия обязаны вести строгий учет оборота определенных веществ и материалов и отчитываться об их использовании и местонахождении. В частности, производители лекарств, например, должны отчитываться о происхождении, лабораторной проверке и соблюдении технологий обработки всех компонентов и синтеза лекарств, чтобы они были сертифицированы для использования людьми в медицинских целях.

ограничивает по типу данные, которые может содержать атрибут, предоставляя возможность для его проверки и автоматической обработки. При моделировании контента используются стандартные приемы управления метаданными и моделирования данных.

Моделирование контента — процесс двухуровневый. На первом уровне (уровне информационного продукта) создается осязаемый результат — например, веб-сайт. На втором уровне (компонентном) осуществляется дальнейшая детализация и прорабатываются составные элементы, образующие модель информационного продукта. Степень детализации (гранулированности) зависит от требований по повторному использованию и структуризации.

Модели контента помогают реализации стратегии управления контентом за счет упорядочения его создания и стимулирования повторного использования. Также модели могут поддерживать создание адаптивного контента, свободного от форматов и независимого от устройств. В конечном счете модели воплощаются в виде спецификаций контента, представленных как описания XML-схем (XSD), формы или таблицы стилей.

1.3.1.4 МЕТОДЫ ДОСТАВКИ КОНТЕНТА

Контент должен быть модульным, структурированным, повторно используемым и независимым от устройств и платформ. Методы доставки (предоставления) контента пользователям включают веб-страницы, печатные материалы, мобильные приложения и файлы всевозможных форматов вплоть до электронных книг с интерактивными видео и аудио. Чем раньше в процессе разработки контент конвертируется в XML, тем проще дается обеспечение его многоцелевого использования в широком спектре каналов мультимедиа.

Системы доставки контента подразделяются на три основных класса, определяемых по роли пользователя.

- **Push-системы (выдача по подписке).** Пользователь выбирает тип контента, который желает получать, и желаемый график или частоту выдачи, после чего система регулярно выдает подписчику контент на клиентское устройство или в пользовательское приложение. При этом возможно как агрегирование контента, так и распространение контента из одного источника по многим каналам. Простейший пример push-системы доставки — настраиваемые RSS-ленты новостей и прочего веб-контента.
- **Pull-системы (выдача по запросу).** Пользователи ищут и выбирают контент, который хотят скачать из интернета. Простейший пример pull-системы — интернет-магазин.
- **Интерактивные системы** выдают контент через всевозможные пользовательские интерфейсы: например, приложения электронных точек продаж (Electronic Point Of Sale, EPOS) сторонних разработчиков или предназначенные для клиентов веб-сайты (например, предлагающие записываться в какие-либо программы в качестве участников, и т. п.). Подобные решения предполагают обмен большими объемами данных в режиме реального времени между корпоративными приложениями. Варианты архитектурных решений, обеспечивающих совместное использование данных приложениями, включают интеграцию корпоративных приложений

предприятия (EAI), регистрацию изменений данных, интеграцию данных и интеграцию корпоративной информации (EII) (см. главу 8).

1.3.2 Контролируемые словари

Контролируемым словарем (*controlled vocabulary*) называют в явном виде определенный перечень слов, которые допустимо использовать в индексах, названиях категорий, документов, файлов и иных объектов, а также тегах метаданных с целью обеспечения возможности поиска, извлечения и просмотра контента. Подобный регламентированный лексикон необходим также и для систематизации документов, записей и контента в каталогах библиотек. Сложность структуры словарей может варьироваться в пределах от простого списка или меню до более сложных кругов синонимов или нормативных словарей, еще более сложных таксономий и вплоть до сложнейших онтологий и тезаурусов. Примером служит так называемое «Дублинское ядро» (Dublin Core), используемое в каталогах публикаций.

Специальные политики организации должны четко определять порядок добавления терминов в словарь и ответственного за его ведение (это может быть, например, штатный специалист по классификации или индексации либо библиотекарь). С профессиональной точки зрения лучше всего обучены проработке контролируемых словарей специалисты по библиотечному делу. Пользователи, имеющие доступ к тому или иному списку допустимых терминов, могут их только использовать в своей предметной области, но никак не дополнять или редактировать (см. главу 10).

В идеале контролируемые словари должны логически и семантически согласовываться с именами и определениями сущностей корпоративной концептуальной модели данных. При этом оптимально подходить к составлению словаря, начиная с лексикографического сбора и анализа данных о фактически используемых на низовом уровне понятиях и терминах, а затем обобщать их в «народную таксономию», чтобы максимально упростить простым пользователям последующую ориентацию в названиях разделов, тегах документов и т. п

Контролируемые словари относятся к категории справочных данных. Соответственно, важно следить за их полнотой и актуальностью. В то же время словари можно считать и разновидностью метаданных, поскольку они помогают разбираться в содержании и назначении других данных. В настоящую главу раздел, посвященный контролируемым словарям, отнесен по той причине, что чаще всего они используются всё-таки применительно к управлению документами и контентом.

1.3.2.1 УПРАВЛЕНИЕ СЛОВАРЯМИ

Поскольку общепринятая терминология со временем меняется, эволюционируют также и словари, и, как следствие, они нуждаются в управлении. В США стандарты управления словарями определены стандартом ANSI/NISO Z39.19-2005 «Руководство по составлению, форматированию и поддержке многоязычных контролируемых словарей» (Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies), где прямо указано, что

управление словарями требуется для «повышения эффективности работы информационно-поисковых систем, систем для веб-навигации и иных сред, запрашивающих желаемый идентифицируемый контент посредством его словесного описания. Главное назначение контроля словаря — обеспечить единообразие и согласованность описания содержимого объектов контента и упростить их нахождение».

Управление словарями включает функции определения, изыскания, импорта и ведения записей (словарных статей) раздельно по каждому словарю. Ключевые вопросы управления словарями касаются их назначения, использования, целевой аудитории, применимых стандартов и порядка ведения:

- Какие информационные концепции будут поддерживаться данным словарем?
- На какую аудиторию он ориентирован? Какие процессы призван обеспечивать? Какие функциональные роли играют потребители?
- Зачем нужен этот словарь? Предназначен ли он для поддержки приложения, управления контентом или аналитики?
- Какой ответственный за принятие решений орган отвечает за выбор предпочтительных терминов из рядов синонимов?
- Какие словари уже существуют и используются различными подгруппами целевой аудитории для классификации информации? Где они находятся? Как создаются и ведутся? Кто отвечает за управление ими и/или выступает в роли экспертов по предметным областям? Не сопряжено ли их использование с проблемами в области информационной безопасности или конфиденциальности?
- Не имеется ли действующего стандарта, который можно применить для решения рассматриваемого вопроса? Стоит ли доверять стандарту стороннего происхождения, или для внутриорганизационного применения лучше создать собственный? Как часто и насколько полно обновляется стандарт? Насколько прост доступ к внешнему стандарту и его импорт? Не слишком ли дорого обойдется обеспечение его совместимости с собственными системами?

Лишь проработав все вышеперечисленные вопросы, вы получите возможность перейти к интеграции данных. Кроме того, их решение способствует определению и совершенствованию внутриорганизационных стандартов, включая соответствующие им словари предпочтительных терминов и функции управления связями и отношениями между терминами.

Если же подобную комплексную экспертизу не проводить, то словари предпочтительной терминологии так или иначе будут у вас в организации «самоопределяться», вот только делаться это будет хаотично и несогласованно, в узкой и ограниченной перспективе разрозненных и разделенных организационно-функциональными перегородками проектов, а результатом станет неизбежное удорожание интеграции систем плюс проблемы с обеспечением качества данных (см. главу 13).

1.3.2.2 ПРЕДСТАВЛЕНИЯ СЛОВАРЯ И МИКРОКОНТРОЛИРУЕМЫЙ СЛОВАРЬ

Представлением словаря (*vocabulary view*) называется подмножество терминов из контролируемого словаря предметной области, ограниченное выбранной тематикой. Представления нужны в тех случаях, когда целью ставится *использовать* стандартную терминологию из словаря с большим число определений, значительная часть которых не нужна целевой аудитории потребителей информации. Например, в представлении словаря с терминами, относящимися к работе отдела маркетинга, не нужны термины, относящиеся к работе финансового отдела.

Представления словарей делают их более удобными в использовании, ограничивая содержание представлений лишь терминами, относящимися к сфере интересов пользователей. Представление словаря можно строить как посредством выбора желательных терминов вручную, так и посредством определения и применения бизнес-правил непосредственно к терминологическим данным или к метаданным. Во втором случае следует определить наборы бизнес-правил для включения терминов из словаря в каждую из тематических выборок-представлений.

Микроконтролируемый словарь (*micro-controlled vocabulary*) является представлением словаря с добавлением узкоспециализированных терминов, не отображаемых в общем словаре. Пример контролируемого на микроуровне словаря — медицинский словарь с определенными подмножествами терминов, относящимися к различным профилям и специализациям. Все узкоспециализированные термины иерархически подчиняются статьям общего контролируемого медицинского словаря, но отображаются только в представлениях для пользователей из числа медицинских специалистов соответствующего профиля. Важно, чтобы в каждом словаре, контролируемом на микроуровне, сохранялась внутренняя согласованность и связность терминологии, обусловленная преемственностью определений и отношений между терминами, заданными в общем словаре.

Контролируемые на микроуровне словари бывают необходимы в тех случаях, когда ставится целью использовать преимущества стандартного словаря, но словарного запаса в нем недостаточно для адекватного описания контента, а потому требуются дополнения и/или расширения для специфических групп потребителей информации. Словарь, контролируемый на микроуровне, строится по тому же алгоритму, что и представление словаря, но включает дополнительные шаги по добавлению или привязке дополнительных предпочтительных терминов, структурно обособленных от ранее существовавших предпочтительных терминов за счет указания в них ссылки на иной, нежели в основных словарных статьях, источник.

1.3.2.3 ТЕРМИНЫ И СПИСКИ ВЫБОРА

Списки терминов ничего лишнего, кроме собственно списков, не содержат. Связей или отношений между терминами они не определяют и не описывают. По форме это могут быть списки выбора (pick lists) опций, раскрывающиеся списки веб-приложений, списки выбора допустимых текстовых значений в меню настроек информационных систем и т. п. Все подобные средства ничего не объясняют и не дают пользователю с точки зрения понимания терминологии, но помогают

избегать двусмысленности и неопределенности, строго ограничивая область определения терминологических значений.

Списки выбора терминов часто запрятаны глубоко в недрах приложений. Специализированное программное обеспечение по управлению контентом позволяет извлекать эти списки, равно как и контролируемые словари, на поверхность, преобразуя их в прозрачные списки выбора, доступные для поиска с домашней страницы. В таком виде эти списки выбора превращаются во встроенные таксономии, управляемые в среде программного обеспечения.

1.3.2.4 УПРАВЛЕНИЕ ТЕРМИНАМИ

Согласно определению стандарта ANSI/NISO Z39.19-2005, «*термин* — слово или словосочетание, используемое для обозначения понятия». Как и словари, отдельные термины (terms) также нуждаются в управлении. Управление терминами включает функции первоначального определения, классификации и регулирования порядка использования термина различными системами. Управление терминами осуществляется в рамках руководства данными. При этом распорядителям данных иногда необходимо вставать на позицию третейских судей в плане разрешения разногласий и обеспечения учета мнений всех заинтересованных сторон при внесении изменений в терминологию. Далее стандарт Z39.19 определяет понятие *предпочтительный термин* (*preferred term*) как выбранный для включения в контролируемый словарь из числа двух или более синонимических терминов.

Управление терминами включает также определение соотношений между терминами в рамках контролируемого словаря. Соотношения бывают трех типов.

- **Эквивалентность** — соотношения, допускающие замену одних терминов контролируемого словаря другими или сочетаниями других при переходе по перекрестным ссылкам. Чаще всего такие соотношения определяются в рамках картирования соответствий между различными ИТ-системами, указывая, каким терминам или значениям в другой системе соответствует термин или значение в исходной системе, без чего невозможна эффективная интеграция и стандартизация.
- **Иерархия** — соотношения между двумя и более терминами контролируемого словаря, описывающие ситуацию, когда один из терминов отражает частный случай или более узкое понятие, чем обобщенный термин.
- **Родство** — ассоциативная, но иерархически не закрепленная связь между терминами контролируемого словаря.

1.3.2.5 КРУГИ СИНОНИМОВ И НОРМАТИВНЫЕ ПЕРЕЧНИ

Кругом синонимов (*synonym ring*) называют группу терминов, которые в грубом приближении можно считать эквивалентными. Круг синонимов позволяет выдавать пользователям по поисковому запросу, включающему один из синонимов, ссылки на контент, который относится также и к остальным синонимам. Разработанные вручную кольца синонимов используются только для

поиска; для индексирования они непригодны. Полные и близкие синонимы при таком подходе никак не различаются, и используются круги синонимов лишь для поиска в индексируемых средах с неконтролируемыми списками словаря или в неиндексируемых средах. Круги синонимов обязательно определяются в информационно-поисковых системах и очень часто применительно ко всевозможным реестрам метаданных (см. главу 13). А вот привязка словарей синонимов к пользовательским интерфейсам — задача весьма затруднительная.

Нормативный перечень или *список* — контролируемый словарь описательных терминов, предназначенный для упрощения и ускорения обработки поисковых запросов в узкоспециализированных информационных системах. В данном случае синонимы и близкородственные термины эквивалентными не считаются (в отличие от круга синонимов); напротив, в нормативном перечне один термин указывается в качестве предпочтительного, а остальные включаются в список в качестве иерархически подчиненных вариантов с перекрестными ссылками на предпочтительный термин. Таким образом, по запросу синонима или варианта пользователь отсылается к предпочтительному термину. Нормативный перечень может быть дополнен определениями. Нормативные перечни ведутся уполномоченными администраторами. Также они могут иметь многоуровневую структуру. Пример: предметный рубрикатор Библиотеки Конгресса США (см. также раздел 1.3.2.1).

1.3.2.6 ТАКСОНОМИИ

Таксономия (*taxonomy*) — обобщенное наименование любой классификации или контролируемого словаря. Общеизвестным примером таксономии является система классификации растительного и животного мира, разработанная шведским биологом Карлом Линнеем.

В управлении контентом под таксономией понимают структуру наименований, включающую контролируемый словарь, которая используется для схематического определения и разграничения тематик с целью обеспечения функциональных возможностей навигации и поиска. Таксономии помогают устранять двусмысленности и контролировать синонимы. Иерархическая таксономия может включать различные типы материнско-дочерних отношений между категориями, предназначенных для использования в целях как индексации, так и поиска. Также подобные таксономии используются и для создания многоуровневых интерфейсов с развертывающимися списками.

По структуре таксономии подразделяются на следующие типы.

- **Плоская** или **горизонтальная таксономия** не предусматривает связей внутри множества контролируемых категорий. Все категории равноправны. Простейший пример: список объектов с одинаковым статусом, например независимых государств.
- **Иерархическая таксономия** имеет древовидную структуру, в которой связи между узлами определяются правилами. Иерархия имеет не менее двух уровней и допускает двустороннее движение по древовидной структуре. При переходе на уровень выше предыдущего категория расширяется (обобщается), на уровень ниже — сужается (уточняется). Пример из географии: карта мира ↔ карта страны ↔ … ↔ карта города ↔ план застройки квартала.

- **Полииерархия** — древовидная таксономическая структура, в которой одному и тому же узлу может соответствовать более одной родственной связи, то есть дочерние узлы могут иметь множественных родителей, а каждый из родительских узлов — множественных прародителей следующего уровня, в том числе и общих. Таким образом, пути обхода родственных связей могут быть весьма хитросплетенными и требуют особо тщательной проработки структуры связей во избежание появления взаимоисключающих путей: например, если шаг вверх по одной ветви приводит к родительскому узлу, то путь, приводящий к тому же узлу за несколько шагов, должен быть запрещен. В результате крайне усложняется и структура правил. Поэтому для определения связей в сложных полиноминальных структурах лучше использовать не древовидную таксономию, а звездообразную схему связей, которой соответствует фасетная классификация.
- **Фасетная таксономия** хорошо представляется в форме звезды из узлов, от каждого из которых имеется единственный путь к центральному узлу. Фасеты[1] — это атрибуты центрального объекта. Пример: структура метаданных библиотеки электронных документов, где каждый атрибут (автор, название, права доступа, ключевые слова, версия и т. д.) выступает фасетом объекта контента.
- **Сетевая таксономия** является гибридом иерархической и фасетной структур. Два любых узла таксономической сети связаны через последовательность ассоциаций. Примеры: 1) рекомендательная система («…заказываете обувь, закажите и средство по уходу за обувью этого класса…»); 2) тезаурус с перекрестными ссылками «*см. также*».

С учетом колоссальных объемов генерируемых в наши дни данных никакие, даже самым тщательным и детальным образом определенные таксономии не избавляют от необходимости дополнять их правилами автоматизированной трассировки, корректировки и маршрутизации запросов. Без должного сопровождения и обновления таксономии быстро утрачивают актуальность и перестают пользоваться спросом или, хуже того, начинают приводить к выдаче некорректных результатов. Последнее, в свою очередь, создает риск невольных нарушений каких-либо законов, правил или регламентов сотрудниками, полагающимися на неверную информацию, почерпнутую через поисковые системы с устаревшими или некорректными таксономиями. Например, если в финансово-юридической таксономии предпочтительным термином выбрано словосочетание «лицо пенсионного возраста», следует проследить, чтобы ему соответствовал адекватный круг синонимов, включая все грамматические формы склонения в единственном и множественном числе самого словосочетания, а также, безусловно, всех словоформ, включающих корень «пенсионер», а возможно, и любые слова с другими корнями, явно или с большой вероятностью указывающими на принадлежность к этой категории. В таких случаях, например, американская версия общепринятых принципов бухгалтерского учета в явном виде предписывает использовать круги синонимов (US GAAP, 2008).

[1] Facet — грань (*англ.*). — *Примеч. пер.*

Организации разрабатывают собственные таксономии также и с целью формализации результатов коллективного осмысления различных аспектов и специфических особенностей своей деятельности. Особенно важны и полезны таксономии с точки зрения обеспечения адекватного представления информации об организации в выдачах поисковых систем, ведь многие из них ищут на веб-сайтах лишь точные совпадения или только тегированные элементы, а иногда требуют еще и точного совпадения порядка слов.

1.3.2.7 СХЕМЫ КЛАССИФИКАЦИИ И ТЕГИ

Схемы классификации являются закодированными представлениями контролируемых словарей. Эти схемы часто имеют иерархическую структуру и могут включать описания классов и подклассов, как, например, в десятичной классификации Дьюи или классификации Библиотеки Конгресса США. Таксономии на основе цифровых кодов, такие как классификация Дьюи, удобны тем, что не имеют лингвистической привязки, поскольку код любого класса или подкласса с равным успехом «переводится» на любой язык.

Фолксономии, то есть «народные таксономии», представляют собой схемы классификации онлайн-контента посредством так называемого социального тегирования. Индивидуальные пользователи и социально-сетевые группы сами аннотируют и классифицируют цифровой контент по жанрам или иным категориям, выбирая для него подходящие теги. Народные классификации обычно не имеют ни иерархической структуры, ни словарей с предпочтительными терминами, а складываются стихийно. Соответственно, их не принято относить к нормативным классификациям и использовать для индексирования документов, поскольку их вырабатывали не эксперты. Зато, будучи прямым отражением преобладающего среди пользователей лексикона, фолксономические теги способствуют повышению вероятности нахождения нужной информации. В то же время народные термины из фолксономии вполне можно привязать посредством ссылок к структурированным контролируемым словарям.

1.3.2.8 ТЕЗАУРУСЫ

Тезаурусом (*thesaurus*) называют разновидность контролируемого словаря, используемого для нахождения контента. Структурно тезаурус сочетает в себе характеристики списков синонимов и таксономии, поскольку предоставляет как самостоятельную информацию о каждом термине, так и отсылки к другим терминам. Связи между терминами в тезаурусе могут быть как иерархическими (общее/частное или шире/уже), так и ассоциативными («см. также») или эквивалентными (синонимы или «используется вместо / заменяет»). Допустимыми синонимами считаются только те, которые приемлемым образом могут использоваться вместо исходного слова в любом сценарном контексте. Тезаурус может также включать определения, ссылки на первоисточники и т. п.

Тезаурусы можно использовать для организации неструктурированного контента, раскрытия смысловых взаимосвязей между контентом различных медиафайлов и ресурсов, улучшения навигации по веб-сайту и оптимизации поиска. По запросу термина пользователями система может

использовать скрытый (от них) тезаурус для автоматического перенаправления поискового запроса на синонимичный или близкий по смыслу термин. Или же, как альтернативное решение, система может в явном виде подсказывать пользователям связанные по смыслу термины для продолжения поиска.

Общие принципы создания тезаурусов регламентируются стандартами ISO 25964 и ANSI/NISO Z39.19 (10.2.2.1.5 Ontologies).

1.3.2.9 ОНТОЛОГИЯ

Онтологией (*ontology*) в информатике называют тип таксономии, представляющий набор понятий и связей между ними, используемый в пределах некой области знаний. Онтологии служат основным средством представления знаний в семантической паутине (Semantic Web) и используются для обмена информацией между ее узлами и веб-приложениями[1].

Онтологические языки, такие как RDFS[2], используются для разработки онтологий посредством кодирования информации, наработанной в различных областях знания. Также они могут включать правила построения логических выводов, позволяющие обрабатывать знания. Язык OWL[3], являющийся расширением RDFS, задает формальные синтаксические правила определения онтологий.

Онтологии описывают классы (понятия), индивидов (экземпляры), атрибуты, отношения и события. Онтология может строиться как собрание таксономий и тезаурусов, что обеспечивает общность лексикона и семантических правил представления информации в описываемой области и обеспечивает возможность обмена ею без риска неверной интерпретации. Часто онтология соотносится с таксономической иерархией классов и определений с помощью родовидовых отношений, что позволяет, например, подвергать разумное поведение декомпозиции на простейшие модули поведения, а те — на слои.

Два ключевых различия между таксономией (как моделью данных) и онтологией:

- Таксономия предлагает смысловую классификацию данных или контента по признаку их отнесения к определенным концептуальным областям, а модель данных выводит в ответ на запрос объекты, к которым логически может относиться фигурирующий в запросе атрибут. В онтологии же объекты, атрибуты и концептуальные области могут смешиваться произвольным образом, а различия между ними идентифицируются через метаданные или иным образом заданные отношения.

---

[1] *Семантическая паутина*, известная также как *Всемирная сеть связанных данных* или Web 3.0, представляет собой надстройку над существующей *Всемирной паутиной*, в которой поддерживается машинная обработка смыслового значения (то есть семантики) данных. Чем больше машина (компьютерная система) понимает, тем проще ей находить, распространять, анализировать и обобщать данные и/или информацию.

[2] RDFS (*сокр.* от *англ.* Resource Description Framework Schema) — *досл.* схема рамочной модели описания ресурсов. — *Примеч. пер.*

[3] OWL (*сокр.* от *англ.* Web Ontology Language) — *досл.* язык веб-онтологии. — *Примеч. пер.*

- В рамках таксономии или модели данных об объекте известно только то, что содержится в его определении, — и ни битом информации больше. В информатике это называется постулатом замкнутого мира: всё неизвестное считается несуществующим. В онтологии же допускаются вероятные отношения, выводимые из природы доказанным образом существующих отношений. То есть нечто гипотетическое и в явном виде не заявленное считается истинным с некой долей вероятности, — и это так называемый постулат открытого мира.

Таксономии изначально разрабатывались в узких рамках библиотечного дела, но на сегодняшний день управление таксономиями и онтологиями вышло на передний край развития науки о семантическом управлении информацией (см. главу 10).

Поскольку процесс моделирования онтологий отчасти субъективен, важно помнить о «капканах», «граблях» и «ложных тропах» на этом пути, из которых наиболее распространенными являются следующие ошибки, чреватые двусмысленностями, неопределенностями и путаницей:

- смешение экземпляров и подклассов объектов при моделировании отношений;
- моделирование событий как связей;
- нечеткие, повторные или плохо разграниченные определения терминов;
- моделирование ролей как классов;
- моделирование ранее смоделированных структур и отношений вместо повторного использования;
- смешение семантики языка моделирования и языка концептуального представления.

Избежать вышеперечисленных и иных ошибок поможет простой прием валидации онтологии с помощью платформенно-независимых веб-приложений, позволяющих оперативно выявлять допущенные недоработки выдачей убедительных сообщений типа: «OOPS! По вашему запросу ничего не найдено».

1.3.3 Документы и записи

Документами (*documents*) называются электронные или бумажные (как печатные, так и рукописные) материалы с инструкциями, руководствами, требованиями, распоряжениями и т. п., которые касаются выполнения различных задач или функций, подтверждениями выполнения вышеназванного, протоколами собраний и решений и т. п. Также документы могут использоваться для распространения информации или обмена знаниями и опытом. Примеры распространенных типов документов — акты, методики, правила, протоколы, процедуры руководства, спецификации, стандарты, технические задания и т. д. и т. п.

Записи (*records*) — подмножество документов определенного вида. То есть не всякий документ классифицируется как запись, но всякая запись относится к категории документов. Записи свидетельствуют, что действия, сведения о которых в них зафиксированы, были действительно

произведены и сделано это было в установленном нормативными документами порядке; соответственно, записи могут использоваться для представления, например, в надзорные органы в качестве доказательства соблюдения организацией в ходе осуществления текущей деятельности установленных всевозможными регламентами требований. В современных условиях записи создаются или ведутся не только людьми, но и автоматическими средствами мониторинга и регистрации.

1.3.3.1 УПРАВЛЕНИЕ ДОКУМЕНТАМИ

Управление документами (*document management*) — понятие, которое описывает весь спектр процессов, приемов и технологий распоряжения документами и записями на протяжении всего их жизненного цикла, включая хранение, учет и контроль как электронных, так и бумажных документов. В наши дни свыше 90% документов создаются в электронной форме. Однако на фоне повсеместного распространения безбумажного документооборота архивы всего мира по-прежнему заполнены бумажными документами за прошлые периоды.

В целом управление документами занимается их формой, а не содержанием, — иными словами, файлами и папками, а не контентом. Информационное наполнение может служить лишь подсказкой, как лучше этим файлом распоряжаться, но в рамках управления документами практическое обращение с этим файлом как с документом будет всё так же строиться на основе его рассмотрения как единого и неделимого целого.

И рыночные, и юридические соображения заставляют тщательно планировать графики хранения, размещения, транспортировки и утилизации записей. Например, не допускается трансграничная передача некоторых категорий личных данных.

Нормативно-правовые и регламентирующие акты наподобие закона Сарбейнса — Оксли и поправок об электронном раскрытии информации к «Федеральному гражданскому процессуальному кодексу» в США или Билля 198 в Канаде доставляют массу хлопот службам информационной безопасности, отвечающим за обеспечение соблюдения всё более ужесточающихся стандартов управления документами и записями организаций (см. главу 7). Управление жизненным циклом документов и записей включает следующие виды работ.

- **Учет** — идентификация и инвентаризация существующих и вновь создаваемых документов и записей.
- **Политика** — разработка, утверждение и обеспечение соблюдения политик ведения, оборота, хранения и уничтожения документов и записей.
- **Классификация** документов и записей.
- **Хранение** физических и электронных документов и записей (текущее и архивное).
- **Получение и распространение** — регулирование доступа к документам и записям, их тиражирования и распространения в соответствии с установленными политиками и правилами, стандартами информационной безопасности и защиты данных, распоряжениями руководства организации и нормативно-правовыми требованиями.

- **Сохранение и уничтожение** — своевременное архивирование и уничтожение документов и записей в соответствии с нуждами организации, законами, нормами и правилами.

Профессионалы в области управления данными сами заинтересованы в адекватной классификации и надлежащем хранении документов, ведь они отвечают за согласованность базовых структурированных данных со специфическими неструктурированными данными. Например, если завершенный годовой отчет считается достаточным документальным подтверждением исторических данных, все включенные в него данные можно вычистить из онлайн-систем и операционных баз данных, существенно их разгрузив, а сам отчет сохранить в архивном хранилище.

Документы часто разрабатываются в рамках иерархии по уровням детализации. Рисунок 72, основанный на текстах «Введения» из ISO 9000 и Руководства по требованиям к документации ИСО 9001 (Guidance on the Documentation Requirements of ISO 9001) (раздел 4.2), отражает ориентированную на документацию парадигму, которой надлежит придерживаться в государственных и силовых структурах. В самом же стандарте ISO 9001 описаны лишь минимальные компоненты базовой системы управления качеством. В коммерческих структурах и иерархия документов, и потоки документооборота от представленной модели могут существенно отличаться, поскольку призваны обеспечить максимальную эффективность бизнес-процессов с учетом специфики отрасли и рынка.

1.3.3.2 УПРАВЛЕНИЕ ЗАПИСЯМИ

Управление записями (*records management*) — важнейший компонент управления документами. К управлению записями предъявляются особые требования[1]. Управление записями также ведется на протяжении всего их жизненного цикла — начиная с создания или получения, включая обработку, использование, распространение, упорядочение и выдачу по запросам, и заканчивая утилизацией. Записи бывают весьма разнородные, включая: физические (документы, записки, договора, отчеты, квитанции, письма, микрофильмы и т. д. и т. п.); электронные (письма и вложенные файлы e-mail, SMS, сообщения по мессенджерам и т. д.); контент веб-сайтов; документы на любых носителях и в любых аппаратных средах; записи в базах данных любого рода и типа. Наконец, не следует забывать и о гибридных записях, таких как, например, диапозитивы в рамках с печатными пояснениями, апертурные перфокарты (со вставкой кадра с деталями описываемого объекта на фотопленке) и всевозможные другие комбинированные форматы. Особый статус имеют так называемые *жизненно важные записи* (*vital records*), необходимые для скорейшего возобновления работы организации в случае возникновения чрезвычайных ситуаций.

[1] Стандарт ISO 15489 определяет управление записями как «область управления, отвечающую за эффективный и систематический контроль создания, получения, ведения, использования и ликвидации записей, включая процессы фиксации и сохранения подтверждающих сведений и информации о деловых операциях и транзакциях в форме записей» (http://bit.ly/2sVG8EW).

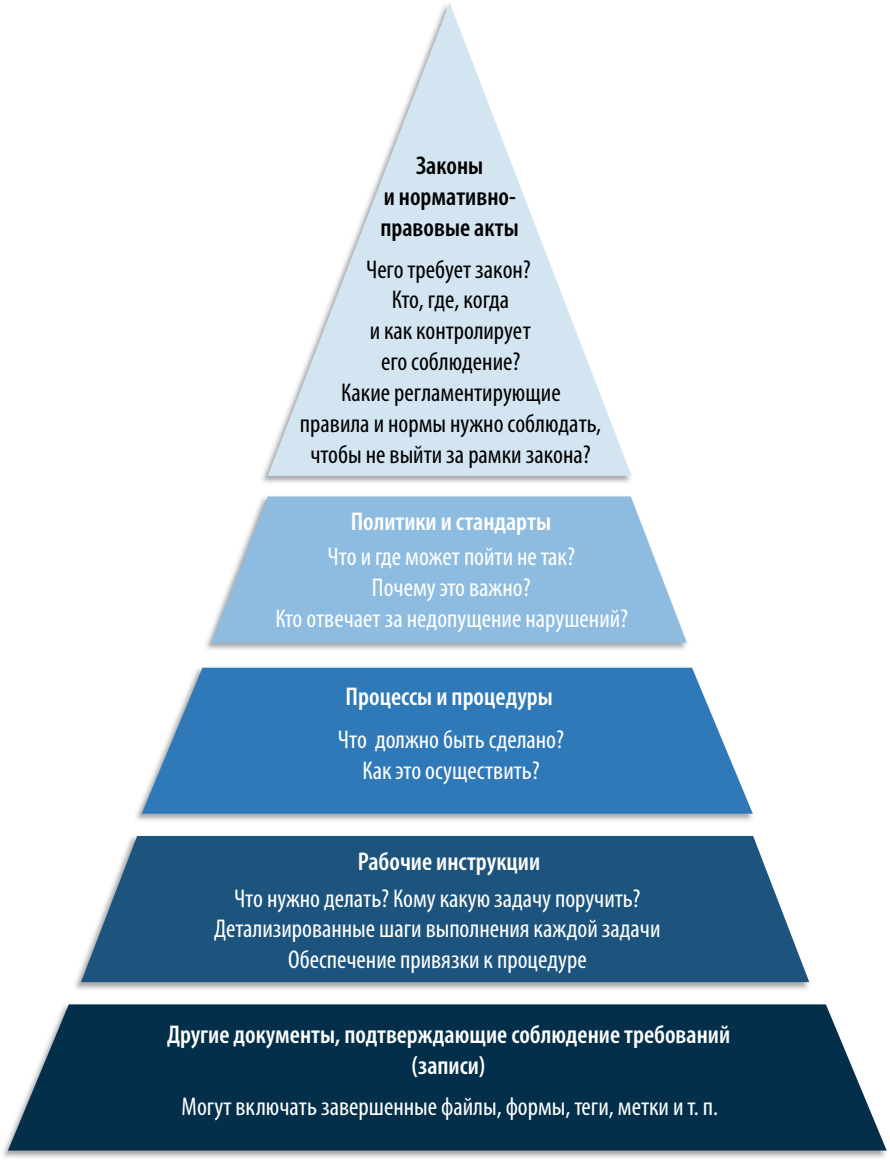


Рисунок 72. Иерархия документов, составленная на основе Руководства по требованиям к документации ИСО 9001 (п. 4.2)

Заслуживающие доверия записи — не самоцель делопроизводства, а средство обеспечения нормального функционирования организации и подтверждения ее законопослушности. Наличие на всех записях действительных подписей (физических или электронных) ответственных лиц — важный элемент подтверждения их достоверности. Другие меры по подтверждению достоверности записей включают заверение подлинности событий (например, свидетельскими показаниями в режиме реального времени) и двойной контроль достоверности ретроспективной информации о событиях.

О качестве подготовки записей можно судить по следующим характеристикам.

- **Контент**: информация должна быть точной, полной и правдивой.
- **Контекст**: метаданные (сведения о записи) должны включать указание автора/создателя записи, дату создания, а при необходимости также данные о связях с другими записями, причем собираться и сопоставляться все контекстные сведения должны непосредственно при создании записи.
- **Своевременность**: запись должна создаваться сразу же после события, действия или решения, которое в ней фиксируется.
- **Неизменность**: после регистрации записей в качестве таковых они не подлежат изменению до истечения установленного законом срока их хранения.
- **Структура**: внешнее представление и порядок элементов информационного наполнения записи должны обеспечивать ее четкость и ясность. Для этого запись должна выполняться с использованием установленных форм или шаблонов. Контент записи должен быть читаемым, а используемая терминология — последовательной и непротиворечивой.

Многие записи существуют одновременно в электронной форме и на бумаге. Лица, отвечающие за управление записями организации, должны четко знать, какой именно экземпляр каждой записи — электронный или бумажный — является официальным «экземпляром записи» с точки зрения учета. После того как официальный экземпляр определен и его сохранность обеспечена, копию записи на альтернативном носителе можно спокойно уничтожить.

1.3.3.3 УПРАВЛЕНИЕ ЦИФРОВЫМИ АКТИВАМИ

Управление цифровыми активами (*Digital Asset Management, DAM*), как процесс, организуется аналогично управлению документами и призван обеспечить надлежащее хранение, отслеживание и использование мультимедийных документов, таких как видео, рисунки, фотографии, логотипы и т. д.

1.3.4 Карта данных

Карта данных (*data map*) — исчерпывающая опись всей имеющейся в распоряжении организации информации, сохраняемой в электронном виде (ESI) с указанием источников данных, приложений и ИТ-сред, в которых они находятся или используются, включая владельцев приложений, ответственных за хранение, фактические географические местонахождения хранилищ и типы данных.

1.3.5 Электронное раскрытие информации (e-discovery)

Досудебное обоюдное истребование и раскрытие сторонами процесса информации с целью нахождения относящихся к делу фактов и проверки аргументированности встречных претензий — стандартная процедура в американской судебной практике. Федеральный гражданский процессуальный кодекс США (US FRCP) регулирует процедуру обнаружения и раскрытия доказательств в рамках гражданских судебных разбирательств с 1938 года. Десятилетиями правила раскрытия

бумажных доказательств в равной мере применялись и к электронным документам и записям. В 2006 году специальные поправки к FRCP официально распространили требования по раскрытию на всю информацию ESI участников судебно-процессуальных действий.

В глобальном масштабе во многих юрисдикциях действуют похожие регламенты, позволяющие организациям представлять суду доказательства в электронной форме. Примеры включают «Закон о взяточничестве» (UK Bribery Act, 2010) в Великобритании; законы Додда — Франка о налоговой отчетности по зарубежным счетам (FATCA) и о коррупции за рубежом (FCPA) в США; Общий регламент по защите данных (GDPR) в ЕС; антимонопольные законы, отраслевые регламенты и процессуальные нормы судов низших инстанций во всем мире.

Электронные документы обычно маркированы метаданными (в отличие от многих бумажных, которые могут как иметь, так и не иметь данных об их происхождении), которые позволяют использовать их в качестве важной части доказательной базы. Ключевые процессуальные требования в рамках судопроизводства предусматривают, помимо электронного раскрытия информации и ранее упомянутых обязательных сроков хранения архивных данных и записей, обязанность выполнения судебных предписаний о продлении срока хранения данных, которые могут потребоваться для судебных нужд, и представления доказательств реальной утилизации данных и уничтожения документов, предельный срок хранения которых истек. По получении предписания о сохранении данных для судебных нужд организация обязана немедленно принять меры по идентификации запрошенной информации и полному закрытию доступа к найденным данным или документам во избежание их изменения, удаления или уничтожения, после чего уведомить все имеющие доступ к документу или данным стороны в организации о факте их сохранения и закрытия по судебному предписанию.

Рисунок 73 наглядно представляет схему эталонной модели электронного раскрытия (Electronic Discovery Reference Model), разработанной EDRM, неформальной организацией специалистов, заинтересованных в выработке единых стандартов и методологий в этой области. Модель предлагает удобный и универсальный подход к поиску и раскрытию требуемых электронных документов (ЭД), позволяя оперативно отыскивать места их хранения внутри организации, определять применимые правила сроков хранения, выявлять недостающие или затерявшиеся ЭД и средства для их скорейшего розыска или восстановления.

Модель исходит из того, что в организации реализована функция руководства данными или информацией. Далее следуют восемь шагов или фаз процесса e-discovery, которые могут носить итерационный характер. По мере продвижения к цели раскрытия всей требуемой электронной информации объем обрабатываемых данных снижается, а степень их актуальности повышается.

Первая фаза — идентификация — на самом деле включает параллельно ведущиеся работы по двум направлениям: предварительной экспертизы обстоятельств и предварительной экспертизы данных (на диаграмме не отображены). Предварительная экспертиза обстоятельств судебного дела позволяет выявить характеристики относящейся к делу информации, по которой следует искать требующиеся документы по совпадению слов в названиях и описаниях (в случае бумажных документов) или метаданных (ключевые слова, диапазон дат создания и т. д.).

Предварительная экспертиза данных позволяет определить типы и потенциальные местонахождения относящихся к делу данных. В процессе выявления необходимые электронные записи и документы (ESI) должны помечаться как не подлежащие уничтожению по истечении нормативного срока хранения. Недостающую информацию следует уточнять у персонала, отвечающего за ведение записей, ответственных за хранение данных или владельцев данных, а также администраторов и операторов баз данных и иных ИТ-систем. Кроме того, персонал, имеющий доступ к данным, идентифицированным как подлежащие раскрытию, должен быть введен в курс дела, осведомлен о судебном уведомлении о необходимости обеспечения сохранности относящихся к делу данных, а также получить разъяснения относительно роли самих сотрудников в ходе рассмотрения судебного спора.

Рисунок 73. Эталонная модель электронного раскрытия (EDRM)[1]

Следующие фазы модели — обеспечение сохранности и сбор недостающих ЭД. Первая заключается в помещении потенциально относящихся к делу данных в надежное хранилище документов особой юридической важности во избежание их уничтожения. Сбор недостающих данных выполняется по запросу юридического отдела или адвокатов компании, которым они передаются под нотариально заверенную расписку.

На фазе обработки ЭД уничтожаются не имеющие юридической силы дубликаты и копии оригиналов, а также производится предварительный отсев не относящихся к делу записей и документов из числа передаваемых на ревизию ЭД. На фазе ревизии определяются записи и документы, строго соответствующие критериям запроса на раскрытие ЭД/ESI. Из числа раскрываемых на стадии ревизии изымаются данные (записи и документы), классифицированные

[1] *Источник*: EDRM, по лицензии Creative Commons Attribution 3.0.

в качестве конфиденциальных. Применительно к электронным документам обычно достаточно отбора/отсева по метаданным. Как правило, процедура *ревизии* предшествует обработке, поскольку обработка более трудоемка в силу необходимости вникать в содержание и разбираться в обстоятельствах, фактах и потенциальных доказательствах, имеющих силу для суда или следствия, после чего результаты в порядке обратной связи передаются на фазы поиска и ревизии ЭД.

Обработка и ревизия ЭД сами по себе являются процессами аналитического характера, однако анализ ЭД выделен в особую фазу, поскольку речь идет об анализе содержания (контент-анализе). Цель контент-анализа — выяснить содержащиеся в ЭД обстоятельства, факты и потенциальные доказательства, имеющие силу для суда или следствия, с целью выработки стратегии ответных действий по мере развития юридической ситуации.

На фазе производства материалов данные и информация выборочно передаются юристам противоположной стороны после их приведения в соответствие с согласованными спецификациями. Первоисточниками информации могут служить файлы, электронные таблицы, e-mail, базы данных, чертежи, схемы, рисунки, фотографии, данные в форматах различных коммерческих программных продуктов, контент веб-сайтов, записи голосовой почты и т. д. ЭД/ESI могут собираться, обрабатываться и выдаваться в самых разнообразных форматах. Для выдачи материалов *в исходном формате* (*native production*) файлов дополнительной обработки не требуется. Производство материалов *в формате, близком к исходному* (*near-native production*), заключается в выборочном извлечении и конвертировании относящихся к делу ЭД. Записи ESI могут передаваться в виде фотокопий или сканов, то есть *в формате, близком к бумажному*. Наконец, *поля данных* представляют собой выборки метаданных и других элементов данных из исходных ЭД, представленные в компактных файлах форматов *.csv, *.txt, *.xml и т. п. Важно также фиксировать информацию о происхождении (lineage) материалов, получаемых на фазе производства, с целью подтверждения подлинности содержащихся в них данных во избежание возможных обвинений в фальсификации или искажении информации, чреватых неприятными судебно-юридическими последствиями.

Представление материалов, произведенных на основе ESI, в ходе предварительного следствия и судебных слушаний или процессов — завершающая фаза электронного раскрытия. Вещественные доказательства категории ESI могут представляться в любом формате — от бумажного до исходного электронного — в качестве подтверждающих или опровергающих аргументов по существу рассматриваемого дела. Также они могут использоваться для инициирования запросов на получение дальнейшей информации или установление дополнительных обстоятельств, проверки достоверности фактов или предположений и просто для убедительного воздействия на аудиторию.

1.3.6 Информационная архитектура

Проектированием *информационной архитектуры* (ИА) называют процесс создания структуры для размещения корпуса знаний, информации или контента. ИА включает следующие компоненты:

- контролируемые словари и тезаурусы;
- таксономии и онтологии;
- навигационные карты;
- карты метаданных;
- спецификации поискового функционала;
- сценарии использования;
- модель потоков обработки пользовательских запросов.

В комплексе со стратегическим планированием контента ИА описывает объекты управления в рамках информационной системы: «чем», собственно, предполагается управлять. А «как» этими предметами управлять — определяется уже на фазе проектирования.

В случае системы управления документами и контентом ИА выявляет или устанавливает связи и отношения между ними, конкретизирует требования к документам и их атрибуты, определяет структуру информационного наполнения документа или системы управления контентом. Продуманная информационная архитектура — основа основ разработки эффективного веб-сайта. Пошаговый сценарный план — проектный эскиз веб-сайта. Он задает общий подход к архитектурному проекту, определяет необходимые элементы каждой веб-страницы и отражает навигационные переходы и информационные потоки между страницами, позволяя составить общее представление о проектируемом веб-сайте во всей его полноте и динамике. Именно исходя из архитектурного проекта разработчики затем создают модели навигации, меню и другие компоненты, необходимые для эффективного управления сайтом и его продуктивного использования.

1.3.7 Поисковые системы

Поисковой системой (*search engine*) в самом широком понимании называется любое программное обеспечение, поддерживающее функциональность поиска в сети веб-сайтов с контентом, включающим термины, введенные в поля поискового запроса. Общеизвестные примеры — Google и другие глобальные поисковые системы. Для реализации поисковой функциональности требуется интеграция нескольких компонентов, включая: собственно поисковое ПО («поисковый движок»); программа-обходчик («паук»), обшаривающая все узлы Всемирной паутины и собирающая URL-адреса найденных сайтов и страниц с контентом, удовлетворяющих заданным условиям поиска; подсистема или программа индексирования найденных ключевых слов и текста; правила ранжирования найденных результатов.

1.3.8 Семантическая модель

Семантическое моделирование (*semantic modeling*) заключается в схематическом описании структурной модели знаний в виде сети узлов, соответствующих концептам (понятиям, идеям, предметам или темам), и связей или отношений между ними. Будучи интегрированными

в информационные системы, семантические сетевые модели позволяют пользователям запрашивать интересующую их логически связную и структурированную информацию, не вдаваясь технические детали. Например, семантическая модель может использоваться для перевода таблиц и представлений базы данных на понятный бизнес-пользователям язык.

Семантические модели содержат объекты и связки. Семантические объекты являются представлением предметов моделирования. Они могут иметь атрибуты различной мощности, доменные классификаторы и идентификаторы. По структуре семантический объект может быть простым, составным, смешанным, гибридным, ассоциацией, подтипом родительского объекта или версией архетипа. Связки представляют собой ассоциации или классы ассоциаций в переводе на общепринятый язык UML. Такие модели помогают выявлять структурные закономерности (семантические паттерны) и тенденции, а через них обнаруживать взаимосвязанность фрагментов информации, которые по всем иным признакам выглядят совершенно разрозненными. Именно это уникальное свойство делает семантические модели незаменимым подспорьем в деле интеграции данных из разных предметных областей и отраслей знания. Однако семантическое моделирование становится возможным лишь при наличии тщательнейшим образом проработанных онтологий и контролируемых словарей по всем рассматриваемым дисциплинам.

В интеграции данных онтологии используются несколькими различными способами. Отдельно взятая онтология может послужить основой для эталонной модели. При наличии множественных источников данных каждый источник может включаться в модель с собственной онтологией, после чего составляются перекрестные карты соответствий между всеми онтологиями. Гибридный подход заключается в использовании множественных онтологий с последующей их интеграцией в единый общий словарь.

1.3.9 Семантический поиск

Семантический поиск заключается в нахождении смысловых и контекстных совпадений, а не заданных ключевых слов. В семантической поисковой системе могут использоваться сложные алгоритмы, вплоть до искусственного интеллекта, позволяющие выявлять соответствия критериям поиска по словам и контексту. Поисковые машины такого рода в идеале должны быть способны анализировать географические привязки, намерения, нюансы словоупотребления, синонимы, понятийные соответствия и много чего еще.

Требования к системе семантического поиска включают необходимость наличия надежных алгоритмов вычисления, что именно ищут пользователи, — то есть, по сути, умения читать мысли по словам. Но если пользователям хочется, чтобы поисковые системы понимали обиходный язык, то, вероятно, они и от веб-контента ожидают подобной же «свойской» семантики. Вот где встает трудная задача перед специалистами по маркетингу и рекламе: как связать воедино ассоциации и ключевые слова, близкие пользователям, с описанием «непревзойденных качеств» продаваемых и рекламируемых брендов?

Оптимизация веб-контента под семантический поиск предусматривает использование ключевых слов из живого языка в самом контенте вместо искусственной вставки жестко заданных ключевых слов в теги. Среди типов семантических слов-ключей стоит отметить: центральные ключевые слова с вариациями (ядро семантики контента); тематические ключевые слова для обозначения концептуально связанных терминов; примыкающие ключевые слова, предвосхищающие вероятные запросы пользователей. Дальнейшая оптимизация контента возможна посредством мониторинга релевантности и цитируемости, а также использования средств продвижения контента веб-сайта в соцсетях.

Пользователи всевозможных аналитических программ и ресурсов, в частности бизнес-аналитики, часто бывают весьма требовательны к поддержке этими средствами функционала семантического поиска. Поэтому инструменты сбора информации для нужд бизнес-аналитики должны быть достаточно гибкими, чтобы бизнес-пользователи всегда могли найти информацию, необходимую им для анализа, отчетности и вывода на всевозможные панели мониторинга текущих показателей. Наконец, и потребителям больших данных требуется некое общее понимание смысла контента, поступающего по всевозможным каналам и в никак не согласованных друг с другом форматах.

1.3.10 Неструктурированные данные

По последним оценкам, не менее 80% от общего объема накопленных в мире данных хранятся вне реляционных баз данных. Строго говоря, все эти данные относятся к неструктурированным, поскольку не подчиняются строгой модели, которая дала бы пользователям возможность однозначно понимать происхождение содержания, порядок организации и систематизации этих данных. Зачастую эти неструктурированные данные никак не маркированы и не классифицированы, не говоря уже об отсутствии строгой табличной разбивки по строкам и столбцам. При этом сам термин *неструктурированные данные* нельзя признать удачным, поскольку он вводит в заблуждение, так как сами документы, формально относящиеся к категории неструктурированных данных, могут быть идеально структурированы с точки зрения иерархии заголовков, глав, перекрестных ссылок, форматов и графиков. Иногда данные, хранящиеся вне реляционных баз данных, называют *не табличными* или *полуструктурированными*. Но ни одно из этих определений не описывает со всей адекватностью ситуацию, сложившуюся в современном общемировом пространстве электронной информации, где сосуществуют колоссальные объемы разнородных по формату и структуре организации их хранения электронных данных.

К неструктурированным данным можно отнести (как минимум) весь контент: файлы в форматах любых текстовых редакторов, электронную почту, социальные сети, чаты, бесструктурные файлы, электронные таблицы, XML-файлы, транзакционные сообщения, отчеты, графики, оцифрованные изображения, фото, видео- и аудиозаписи. Остается прибавить к этому списку колоссальный (и не поддающийся строгой оценке) поток бумажных документов и записей.

Фундаментальные принципы управления данными тем не менее в равной мере применимы и к неструктурированным данным. Неструктурированные данные должны рассматриваться в качестве ценного актива организации, которая ими располагает, наравне со структурированными. Соответственно, на неструктурированные данные распространяются те же принципы и правила хранения, доступа, безопасности, контроля качества и эффективности использования, что и на структурированные. То есть неструктурированные данные — наравне со структурированными — требуют высокоуровневого руководства распоряжения, проектирования архитектуры, защиты, метаданных — и далее по списку.

Интерес к неструктурированным и полуструктурированным данным возрастает по мере развития крупных хранилищ данных и методов бизнес-анализа. В рамках этих подходов модели данных могут предусматривать индексирование больших массивов неструктурированных данных с целью их анализа и, в конечном итоге, структуризации. Некоторые СУБД включают функционал обработки URL-адресов неструктурированных данных с последующей выдачей гиперссылок на их извлечение в таблицы БД в структурированном виде. Подробнее о неструктурированных данных и «озерах данных» рассказано в главе 14.

1.3.11 Поток работ

Наработку контента нельзя пускать на самотек; должен быть организован четкий *поток работ* (*workflow*) и план-график создания и утверждения контента к публикации. Компоненты рабочего процесса могут включать создание, обработку, маршрутизацию, правила; элементы администрирования и надзора, обеспечения безопасности, включая управление электронными подписями; крайние сроки сдачи работ различных типов, порядок вынесения проблем (в случае возникновения) на рассмотрение высшего руководства, форматы, сроки и периодичность предоставления отчетности и технической документации. Поток работ должен быть предельно автоматизирован, для чего можно использовать стандартную или собственной разработки систему управления контентом (Content Management System, CMS). Ручное управление информационным наполнением серьезных веб-ресурсов при современных объемах и потоках разнородных данных — занятие избыточно трудоемкое и в целом контрпродуктивное.

К тому же любое приложение класса CMS поддерживает столь полезную для администраторов и редакторов веб-сайтов функциональность, как контроль версий. Любой электронный материал, будучи обработан CMS, получает метки даты / времени проверки / публикации, номера версии и имени пользователя, внесшего изменения.

Поток работ нужно организовать циклическим образом, причем желательно, чтобы этапы подготовки, обработки, публикации, архивирования, хранения и утилизации контента различных типов, категорий и форматов были максимально унифицированы. Возможно, понадобится разработать наборы схем рабочих процессов и шаблонов для контента каждого типа, если приходится иметь дело со слишком разнородными материалами. Важно также обеспечивать согласованность точек выдачи распространяемого контента с фактическим местопребыванием его ключевых потребителей. Крайние сроки выдачи готовых материалов подлежат уточнению по

мере фактического выполнения работ, а организация рабочих процессов должна неуклонно совершенствоваться, иначе вы рискуете столкнуться с хроническими срывами сроков публикаций и неразберихой в части определения персонально ответственных за подготовку каждого отдельно взятого материала.

2. ПРОВОДИМЫЕ РАБОТЫ

2.1 Планирование управления жизненным циклом

Практика управления документами предусматривает планирование жизненного цикла документов — от создания или получения вплоть до ликвидации, включая тиражирование и распространение, хранение и архивацию, контроль доступа и защиту от уничтожения. Планирование предусматривает классификацию и систематизацию документов посредством создания индексированных каталогов и таксономий для упорядоченного хранения и быстрого извлечения документов из библиотек или хранилищ. Важный момент: отдельного планирования требует управление жизненным циклом записей (см. раздел 1.3.3.2).

Прежде всего нужно определить подразделение организации, несущее ответственность за управление документами и записями. Именно это подразделение будет координировать доступ к записям и документам и их распространение внутри и вне организации; перенимать и включать в свою работу передовой опыт и практики работы других подразделений и обеспечивать единый комплексный подход к документообороту и ведению записей в масштабах организации; осуществлять стратегическое планирование управления документами, включая выработку плана обеспечения бесперебойного доступа к жизненно важным для продолжения бизнес-процессов документам и записям в чрезвычайных ситуациях. Подразделение обеспечивает: соблюдение правил и сроков хранения документов и записей в соответствии со стандартами организации и внешними нормативно-правовыми требованиями; надлежащее ведение архивов записей, подлежащих длительному хранению, и уничтожение по истечении установленного срока годности или хранения всех прочих записей в соответствии с правилами организации и требованиями действующего законодательства и регламентов.

2.1.1 Планирование управления записями

Управление записями начинается с четкого определения состава и структуры записей. Команда по определению структуры записей на уровне функциональной области должна включать профильных экспертов в этой области наряду со специалистами, разбирающимися в системах управления записями соответствующих типов.

Управление электронными записями требует принятия решений относительно мест хранения текущих (активных) записей и устаревших (архивных) записей и порядка списания активных записей в архив. Вопреки иллюзии повсеместного и всеобщего перехода на электронное делопроизводство, бумажные записи по-прежнему ведутся и в ближайшей перспективе никуда не

исчезнут. Поэтому подход к управлению записями должен предусматривать учет и определять порядок хранения и утилизации бумажных записей, а также неструктурированных электронных данных наряду со структурированными электронными записями.

2.1.2 Разработка стратегии управления контентом

Стратегическое планирование управления контентом должно быть прямым образом ориентировано на реализацию общеорганизационного подхода к выдаче пользователям релевантной и максимально полезной информации оперативно и в полном объеме. План должен вырабатываться с учетом стимулов и мотивов, побуждающих пользователей запрашивать контент, и предусматривать соответствующие этим стимулам и мотивам задачи по наработке и выдаче контента. Требования к содержанию информационных и мультимедийных материалов (контента) должны затем в полной мере учитываться при принятии технологических решений и даже служить основными критериями выбора в пользу того или иного решения — например, в пользу конкретной системы управления контентом.

Стратегическое планирование контента начинается с инвентаризации имеющегося, оценки текущего состояния и выявления недостатков (пробелов). Выработанная стратегия должна определять порядок приоритизации и организации контента и доступа к нему. Экспертиза часто позволяет выявить резервы оптимизации производства, обработки и утверждения создаваемого контента. Единая стратегия контента должна строиться с особым упором на его модульную структуру и производство шаблонных заготовок многоразового использования, а не штучных экземпляров.

Чтобы облегчить людям поиск разнообразного и разнородного контента, стратегия его планирования должна в обязательном порядке включать его классификацию на основе метаданных и поисковую оптимизацию (search engine optimization, SEO). Продукты стратегического планирования контента должны включать четкие рекомендации по созданию, публикации и администрированию контента. Поддержанию на стабильно высоком уровне и дальнейшему развитию общеорганизационной стратегии управления контентом весьма способствуют продуманные политики, стандарты и руководства.

2.1.3 Определение политик обращения с контентом

Политики обращения с контентом систематизируют требования посредством определения принципов, направлений и руководящих указаний по проведению работ. Они же способствуют выработке у сотрудников понимания смысла требований в сфере обращения с документами и записями, минимизируя риск непреднамеренных нарушений.

Большинство административных программ управления документами предусматривает политики, регулирующие следующие аспекты обращения с контентом:

- объемы и порядок проведения внутренних проверок соблюдения правил;
- порядок идентификации и защиты жизненно важных записей;
- требования по сохранению записей (так называемый порядок хранения — retention schedule);

- порядок действий при получении судебных предписаний (или иных распоряжений уполномоченных органов) о сохранении каких-либо категорий записей по истечении стандартных сроков хранения;
- правила хранения записей в учреждении и за его пределами;
- правила использования локальных и сетевых жестких дисков;
- управление электронной почтой на предмет обеспечения соблюдения требований ИБ в части препятствия утечкам чувствительной информации;
- использование адекватных методов уничтожения записей (например, с помощью специализированных организаций с получением справки об уничтожении).

2.1.3.1 ПОЛИТИКИ В ОБЛАСТИ СОЦИАЛЬНЫХ МЕДИА

Помимо вышеперечисленных стандартных тем многие организации сегодня разрабатывают еще и политики размещения контента в приобретших огромную популярность социальных сетях и медиаресурсах. Например, организации нужно четко определиться, являются ли посты ее сотрудников на страницах Facebook, Twitter, LinkedIn, чатов, блогов, вики-ресурсов или онлайн-форумов контентом, затрагивающим интересы организации и подлежащим регламентации (особенно в тех случаях, когда они размещаются в соцмедиа в рабочее время или с использованием учетных записей организации).

2.1.3.2 ПОЛИТИКИ ДОСТУПА С ПЕРСОНАЛЬНЫХ УСТРОЙСТВ

Поскольку маятник внимания пользователей необратимо качнулся от офисной оргтехники и ПО в сторону персональных мобильных устройств со встроенными ОС (от ноутбуков и планшетов вплоть до «умных» часов и брелоков) и платформенных решений, предлагающих доступ к системам и данным через клиентские мобильные приложения с функциями управления контентом и записями, приходится считаться с такими сценариями доступа и тщательно прорабатывать вопросы обеспечения требований нормативно-правового соответствия и безопасности данных применительно к ним.

Политики в этой области должны проводить четкое различие между неформальным контентом (например, наполнением Dropbox или Evernote) и формальным (например, контрактами и соглашениями) — и уделять первоочередное внимание контролю доступа к формальному контенту и допустимому порядку его использования. Но это не означает, что неформальный контент можно пускать на полный самотек. Политики должны быть сформулированы и для него.

2.1.3.3 ОБРАЩЕНИЕ С ЧУВСТВИТЕЛЬНЫМИ ДАННЫМИ

Организации по закону обязаны обеспечивать защиту персональных, конфиденциальных и установленных категорий чувствительных данных. Структуры ИБ и/или руководства данными обычно определяют классы конфиденциальности и схемы защиты данных, отнесенных к различным категориям информации ограниченного доступа. Лица, занятые производством, сборкой или

публикацией контента, обязаны соблюдать все правила конфиденциальности, определенные для этих классов. Документы, веб-страницы и прочие компоненты контента должны маркироваться соответствующими метками конфиденциальности или чувствительности согласно правилам организации и требованиям законодательства и надзорных органов. Выявленные конфиденциальные данные помимо маркировки могут при необходимости подлежать маскировке или удалению с сайтов и страниц (см. главу 7).

2.1.3.4 ГОТОВНОСТЬ К РОЛИ ОТВЕТЧИКА

Организациям следует обеспечивать готовность к представлению данных, запрашиваемых судебными или надзорными органами, проводя проактивные мероприятия в части электронного раскрытия информации (надеясь на лучшее, всегда стоит готовиться к худшему). Необходимо организовать ведение скрупулезного инвентарного учета всех источников данных и рисков, обусловленных их использованием. Выявляя источники потенциально чреватых юридическими осложнениями данных и фиксируя их, организация получает возможность оперативно реагировать на уведомления о необходимости сохранения данных для предстоящих разбирательств и избегать их невосполнимой утери. Излишне говорить, что процессы поиска требуемой к раскрытию электронной информации нужно максимально автоматизировать.

2.1.4 Определение информационной архитектуры контента

Многие информационные системы, включая семантическую паутину, поисковые системы, анализаторы контента соцсетей, системы проверки корректности записей и управления рисками, геоинформационные системы (geographic information systems, GIS), а также BI-приложения содержат структурированные и неструктурированные данные, документы, тексты, изображения и прочий контент. Пользователям приходится подбирать формулировки запросов таким образом, чтобы они были понятны механизмам поиска и извлечения информации, интегрированным в эти системы. Соответственно, каталоги, инвентарные перечни документов, а также файлы или источники структурированных и неструктурированных данных нужно описывать или индексировать в таком формате, чтобы поисковые алгоритмы имели возможность оперативно находить совпадения и выдавать пользователям ссылки на удовлетворяющие критериям запросов данные. С учетом несовершенства формулировок пользовательских запросов выдача в большинстве случаев так или иначе будет либо избыточной, либо неполной.

Поисковые системы просматривают либо результаты индексирования контента, либо метаданные. Следовательно, следует уделять внимание проработке структуры индексов, чтобы она соответствовала понятиям и отражала предпочтения целевой аудитории пользователей. Также поиск ведется с учетом определенных правил управления словарем и синтаксисом, и это важно учитывать в формулировках заголовков публикуемых материалов, выстраивая их в соответствии с аналогичными правилами.

Профессионалы в области управления данными могут привлекаться к разработке структуры и терминологии контролируемых словарей в процессе работы над справочными данными (см.

раздел 1.3.2.1) и метаданными, описывающими неструктурированные данные и контент (см. главу 12). В этом случае следует делать всё возможное для обеспечения согласованности по формулировкам, структуре и схемам классификации между контролируемыми словарями, индексами, используемыми для выдачи данных по поисковым запросам, терминологией модели данных и метаданными в рамках каждого проекта по управлению данными и разработке приложений.

2.2 Управление жизненным циклом документов и контента

2.2.1 Сбор записей и контента

Первыми шагами к управлению документами и контентом являются отслеживание их появления, регистрация местонахождения и сбор. Электронный контент обычно изначально представлен в формате, пригодном для сохранения в хранилищах цифровых данных. Что касается бумажных записей, то во избежание утери содержащейся в них информации (контента) их следует оперативно сканировать, загружать в корпоративную систему, индексировать и отправлять на хранение в соответствующий репозиторий, при необходимости заверяя цифровой подписью ответственного лица.

После того как контент оцифрован, его обязательно следует маркировать (индексировать) корректными метаданными, которые должны включать (как минимум) идентификатор исходного бумажного документа или изображения, дату регистрации/оцифровки, название и автора/авторов документа или записи. Без этих метаданных невозможно впоследствии ни обрабатывать запросы на извлечение документов/записей из хранилищ, ни интерпретировать смысл их содержания в привязке к контексту. Автоматизации процессов регистрации и включения подобных артефактов в базы данных цифровых документов и контента в значительной мере способствуют современные технологии сканирования, распознавания образов и текста, оцифровки аналоговых аудио- и видеосигналов, но их использование должно обязательно протоколироваться в журналах регистрации событий, чтобы при необходимости можно было подтвердить подлинность происхождения оцифрованных материалов.

Некоторые платформы соцмедиа предлагают функциональность автоматического сбора и регистрации записей. Сохранение этой информации в репозитории позволяет затем просматривать и анализировать сделанные записи и присваивать им метатеги с целью классификации и последующего управления контентом. Роботы-поисковики позволяют отслеживать и фиксировать текущие версии веб-сайтов. Веб-регистраторы, API-интерфейсы и RSS-ленты также можно использовать для захвата контента или импорта данных из соцмедиа. Регистрация записей, поступающих из социальных сетей, может вестись в ручном, полуавтоматическом или полностью автоматическом (согласно предопределенным установкам) режимах.

2.2.2 Управление версиями и контроль

Стандарт ANSI/IEEE 859, регламентирующий порядок действий в случае аварий в энергосетях, предусматривает три уровня контроля данных в зависимости от их критичности и потенциального ущерба от их повреждения или утери: формальный, по номеру версии и по усмотрению.

- **Формальный контроль (formal control)**: наиболее строгий уровень, предусматривающий наличие формальной процедуры инициации, согласования с заинтересованными сторонами и внесения изменений.
- **Контроль версий (revision control)**: менее формализованный контроль, предусматривающий уведомление заинтересованных сторон и обновление номера версии при внесении изменений в документ.
- **Контроль хранения (custody control)**: наименее формализованный контроль, требующий просто надежного хранения документов с возможностью их извлечения и восстановления в первоначальное состояние.

Таблица 15 содержит пример списка информационных активов и возможных уровней контроля.

Таблица 15. Уровни контроля документов согласно стандарту ANSI/IEEE 859

Информационные активы	Формальный контроль	Контроль версий	Контроль хранения
Перечни мероприятий		x	
Планы совещаний			x
Результаты проверок		x	x
Бюджеты	x		
Инструкции			x
Окончательные условия предложений			x
Финансовая отчетность	x	x	x
Данные о персонале		x	
Протоколы рабочих совещаний			x
Примечания к протоколам и списки участников рабочих совещаний		x	x
Планы проектов (в том числе управления данными и управления конфигурациями)	x		
Предложения (на стадии рассмотрения)		x	
Планы-графики работ	x		
Технические задания	x		
Отраслевые исследования		x	
Учебные материалы	x	x	
Рабочие материалы			x

ANSI 859 рекомендует учитывать следующие критерии при определении уровней контроля документов:

- затратность предоставления данных и проведения обновлений;
- влияние на проект при высокой стоимости или сложности проведения изменений;
- другие возможные существенные осложнения в реализации проекта или в деятельности организации;
- необходимость перехода на более раннюю версию информационного ресурса;
- ведение истории изменений версий (если этого требуют интересы организации и/или реализуемого проекта).

2.2.3 Резервное копирование и восстановление

Система управления документами, записями и/или контентом должна быть включена в общеорганизационный план мероприятий по резервному копированию и восстановлению данных, включая аварийное восстановление с целью скорейшего возобновления деятельности. Программа управления жизненно важными записями обеспечивает организации доступ к данным, необходимым для продолжения функционирования на фоне чрезвычайных ситуаций и возобновления работы в нормальном режиме после их ликвидации. Жизненно важные записи должны быть, во-первых, идентифицированы, а во-вторых — надлежащим образом защищены и продублированы в рамках плана резервного копирования и восстановления. Менеджер записей должен привлекаться к планированию мер по минимизации рисков и обеспечению бесперебойности бизнеса, чтобы не упустить из виду необходимые меры по обеспечению сохранности жизненно важных записей.

Чрезвычайные ситуации могут возникать по самым разным причинам, включая аварийные отключения энергоснабжения, человеческий фактор, отказы компьютерного или сетевого оборудования, фатальные сбои в работе программного обеспечения, кибератаки, теракты, техногенные катастрофы, стихийные бедствия. «План обеспечения непрерывности бизнеса» (или «План аварийного восстановления») включает письменные правила, процедуры и информацию, которые необходимы для смягчения последствий чрезвычайных ситуаций в плане угроз данным организации, включая документы и записи, и их скорейшего восстановления с минимальными потерями в случае реального наступления нежелательного события.

2.2.4 Обеспечение соблюдения сроков хранения и правил ликвидации

Эффективное управление документами или записями подразумевает наличие четких правил и процедур управления сроками хранения и ликвидации записей/документов с истекшим сроком хранения. Для каждой категории записей/документов должны быть четко определены временны́е рамки хранения в оперативном доступе и в архивах. Сроки архивного хранения определяются применимыми законодательными, юридическими или финансово-надзорными требованиями, но могут продлеваться для записей/документов, имеющих историческую ценность. Политика

хранения определяет, когда именно документ закрывается для активного доступа и переводится в архив (например, автономное резервное хранилище), или срок, на протяжении которого документ в активном доступе не был ни разу востребован, чтобы он подлежал списанию в архив. Также политика определяет и срок хранения архивных документов до их окончательного уничтожения, и графики, правила и методы их утилизации. Требования законодательства и регламентов — обязательный для принятия в расчет фактор при определении планов-графиков ликвидации устаревших записей и/или документов.

Менеджеры записей или владельцы информационных ресурсов обязаны обеспечивать надзор за соблюдением их командами требований в области защиты личной информации и чувствительных данных и принимать все необходимые меры по выявлению и пресечению возможных хищений и утечек.

Само по себе хранение документов/записей осуществляется с учетом соображений, обусловленных характером программного обеспечения, необходимого для доступа к их содержимому. В зависимости от формата документов или структуры записей, для доступа к ним могут требоваться конкретные версии определенных приложений и даже операционных систем. Изменения в ИТ-среде — даже столь простые, как, например, установка нового приложения, — могут сделать документы некоторых форматов нечитаемыми, а записи — недоступными.

Не представляющая более никакой ценности информация должна удаляться из хранилищ организации во избежание неэффективного использования физического и электронного пространства, а также с целью уменьшения расходов на ее сопровождение. Кроме того, хранение в активном доступе неактуальных просроченных записей после установленной законом даты чревато негативными юридическими последствиями. Если записи будут востребованы в качестве доказательств, их всегда можно отыскать в архиве.

Несмотря на вышесказанное, многие организации не включают в число приоритетов удаление утратившей всякую ценность информации, в частности по следующим причинам:

- недостаточная проработанность политики и правил;
- информация, не представляющая ценности для удаляющего ее, может оказаться ценной для кого-то другого;
- невозможность предугадать будущие информационные потребности, из-за чего утратившие ценность физические или электронные записи могут оказаться вновь востребованными и даже незаменимыми;
- отсутствие у кого бы то ни было заинтересованности в управлении записями;
- неспособность определить критерии оценки записей на предмет их ценности или ненужности;
- кажущаяся трудоемкость или затратность реализации процессов выявления и удаления утративших ценность физических и электронных записей;
- дешевизна электронной памяти: проще докупить носители информации, чем искать и удалять устаревшую и сделавшуюся ненужной запись.

2.2.5 Аудит документов/записей

Управление документами/записями требует проведения периодических аудиторских проверок, призванных установить, доходит ли нужная информация до нужных людей в установленные сроки, как это задумано и определено в целях информационной поддержки принятия решений или оперативного управления. Таблица 16 содержит примеры направлений таких аудиторских проверок.

Таблица 16. Примеры направлений аудита управления документами/записями

Компонент управления документами/записями	Пример предмета аудиторской проверки
Инвентаризационный учет	Каждому документу/записи в инвентаризационной описи соответствует уникальный идентификатор и место хранения
Хранение	Хранилища физических документов/записей не переполнены и имеют достаточные резервы свободных площадей для размещения новых поступлений на период до следующей плановой проверки
Достоверность и точность	Выборочные проверки документов/записей на предмет адекватности отображения в них той информации, которую они призваны были зафиксировать при ее создании или получении
Схемы классификации и индексирования	Проверка адекватности метаданных электронных документов/записей и планов размещения папок с бумажными документами
Доступ и извлечение	Конечные пользователи без проблем находят критически важную для них информацию
Процедуры обеспечения соблюдения сроков хранения	Календарный график сроков хранения и списания документов/записей структурирован логичным образом на уровне отделов, функциональных или административных подразделений организации
Методы уничтожения	Документы/записи утилизируются или уничтожаются в соответствии с утвержденными рекомендациями
Информационная безопасность и конфиденциальность	Случаи нарушений требований конфиденциальности, утери или утечки документов/записей регистрируются в качестве инцидентов в сфере ИБ и надлежащим образом расследуются
Понимание организацией смысла и содержания управления документами/записями	Методическая и разъяснительная работа с ключевыми лицами и сотрудниками относительно их ролей и обязанностей в сфере управления документами/записями проводится в достаточном объеме

Типичная аудиторская проверка включает следующие этапы:

- определение действующих в организации движущих сил и факторов, а также лиц, заинтересованных в адекватном управлении документами/записями, с целью выяснения полноты понимания ответа на вопрос «зачем?»;
- сбор данных о процессе управления документами/записями (получение ответа на вопрос «как?») после того, как определено, что именно следует проверять/измерять и какие средства для этого использовать (например, стандарты, сравнительные показатели, анкетирование);
- составление отчета о результатах проверки;
- разработка плана-графика дальнейших действий.

2.3 Публикация и доставка контента

2.3.1 Предоставление доступа для поиска и получения

После того как контент описан метаданными / тегирован ключевыми словами и классифицирован в рамках существующей архитектуры информационного наполнения веб-ресурса, он становится доступен для получения и использования. Технология порталов с поддержкой профилей пользователей помогает им отыскивать неструктурированные данные. Поисковые системы находят подходящий для выдачи контент по ключевым словам. В некоторых организациях имеются также профессиональные специалисты по поиску и получению требуемой информации с помощью внутренних инструментов поиска.

2.3.2 Доставка и выдача по всем возможным каналам

Происходит сдвиг в ожиданиях потребителей контента, которые всё больше хотят получать его на те устройства, которые сами выбирают. Многие организации, однако, по старинке создают контент в каком-нибудь редакторе документов вроде MS Word, а затем преобразуют документы в формат HTML или выдают контент, ориентированный на какую-либо единственную платформу, а то и строго определенное разрешение или размер экрана. Если требуется отправить такой контент на выдачу по какому-либо другому каналу (например, на печать), его приходится специальным образом обрабатывать дополнительно (например, пропускать через процедуру подготовки к печати). Потенциально возможны ситуации, при которых для внесения в контент каких-либо изменений его приходится возвращать в исходный формат, редактировать, а затем конвертировать в публикуемый формат заново.

При преобразовании в HTML структурированных данных (например, таблиц из СУБД) результат зачастую получается невразумительным, поскольку структура исходных данных адекватно не отображается. Увы, таковы издержки разделения формы и содержания, и извлечение данных из особым образом структурированного и отформатированного контейнера бывает задачей не из простых.

3. ИНСТРУМЕНТЫ

3.1 Системы управления корпоративным контентом

Система управления корпоративным контентом (ECM[1]) может представлять собой как единое платформенное решение, включающее все основные компоненты, так и набор приложений с различной степенью интеграции в единую систему (от полностью интегрированных до полностью самостоятельных). Компоненты или приложения могут находиться как по месту работы, так и в облачной среде.

Отчеты могут доставляться пользователям по различным каналам, включая распечатку, e-mail, веб-сайты, порталы и мессенджеры, а также через интерфейс системы управления документами. В зависимости от способа доступа пользователи могут искать нужную информацию посредством углубления в многоуровневые списки, просматривать, полностью или выборочно загружать и при необходимости выводить отчеты на печать или заказывать полиграфическим образом исполненные тиражи. Функции управления каталогами, включая добавление и удаление отчетов из папок и их перенос из папки в папку, существенно упрощают управление отчетами. Систему можно настроить таким образом, чтобы она автоматически контролировала сроки хранения отчетов и по их истечении автоматически удаляла их из открытого доступа, перемещая в архив: например, на выделенном под это дисковом накопителе или записывая на CD-ROM, DVD и т. п. Можно хранить отчеты и в облачном хранилище. Важно помнить о недопустимости хранения контента в не поддерживаемых или устаревших форматах, поскольку это представляет риск для организации, что не раз уже отмечалось (см. главы 6, 8 и раздел 3.1.8).

Границы между управлением документами и управлением контентом становятся всё более размытыми по мере взаимопроникновения и тесного переплетения между собой различных бизнес-процессов и ролей на фоне непрекращающихся усилий поставщиков программных продуктов осваивать всё новые и новые рынки.

3.1.1 Управление документами

Система управления документами — это прикладное программное обеспечение, используемое для отслеживания и хранения электронных документов и электронных образов бумажных документов. Примерами специализированных систем управления документами являются библиотеки документов, электронные почтовые программы и системы управления изображениями. Системы управления документами часто включают функции управления каталогами, сравнения версий, ИБ и защиты данных, управления метаданными, индексирования контента и поиска. Расширенные возможности некоторых систем могут включать, например, предварительный просмотр документов в режиме сравнения метаданных.

[1] *сокр.* от *англ.* Enterprise Content Management [system]. — *Примеч. пер.*

Документы могут создаваться непосредственно в системе управления документами или посредством сканирования с возможной обработкой программами распознавания символов (OCR) с целью преобразования в структурированные текстовые документы. Все электронные документы, включая отсканированные копии бумажных, должны в обязательном порядке индексироваться по ключевым словам или словосочетаниям, чтобы впоследствии их можно было находить. Как правило, любой документ автоматически сохраняется вместе с метаданными, такими как имя создателя и даты создания, последнего изменения и сохранения. С целью упрощения поиска документы могут распределяться по категориям с присвоением уникального идентификатора и/или внесением в *метаданные* дополнительных слов-ключей, ожидаемых в запросах поиска. *Метаданные* могут извлекаться из документа автоматически и редактироваться либо добавляться пользователем к имеющимся вручную. Библиографические записи о документах представляют собой структурированные описательные данные, как правило, в формате машиночитаемых каталожных записей стандарта MARC (Machine-Readable Cataloging), хранящихся в локальных базах данных библиотек и доступных через обмен каталогами пользователям во всем мире (если позволяют правила приватности и разрешения).

Некоторые системы оснащены дополнительно функционалами поддержки работы со сложносоставными документами и тиражирования контента. Современные программные средства обработки текстов позволяют внедрять в формально текстовые документы всевозможный контент, включая электронные таблицы, графику, аудио, видео и прочие мультимедийные объекты. Кроме того, составной документ может включать и набор элементов пользовательского интерфейса, позволяющих обеспечить единое представление интегрированного контента.

Системы управления хранилищами документов обычно поддерживают ряд функций работы непосредственно с документами, таких как включение/исключение документов, контроль версий, открытие для совместного доступа, сравнение версий, архивирование, управление кодами статусов, перенос из одной среды хранения в другую и безвозвратное удаление. Также может поддерживаться доступ и некоторые функции управления свойствами документов во внешних хранилищах (например, в файлообменной или облачной среде).

Некоторые системы управления документами включают модули поддержки рабочих процессов различных типов, например:

- ручная обработка с отслеживанием адресов отправки документа пользователем;
- обработка документов согласно заданным правилам маршрутизации потоков документов внутри организации;
- применение динамических правил, позволяющих дифференцировать порядок обработки документов в зависимости от их содержания.

Системы управления документами обязательно имеют модуль управления правами, через который администратор определяет правила открытия доступа к документам в зависимости от их типа и допусков пользователей. Организации могут дополнительно определять типы

документов, требующие дополнительной защиты или контроля доступа. Ограничения доступа по соображениям безопасности, включая защиту конфиденциальных данных, применяются на всех стадиях создания, управления и выдачи документов. Электронная подпись, к слову, — лучшее подтверждение идентичности личности отправителя документа и подлинности сообщения.

Некоторые системы в большей степени ориентированы на защиту данных и информации от утечки или несанкционированного доступа, нежели на обеспечение удобства доступа к ним с целью использования или извлечения. Такая ситуация характерна для спецслужб, вооруженных сил, отраслевых научно-исследовательских институтов и лабораторий. В отраслях с высоким уровнем конкуренции или жестким регулированием, таких как фармацевтическая промышленность и финансовый сектор, также преобладают системы с жестким контролем доступа и обширными мерами по обеспечению ИБ.

3.1.1.1 УПРАВЛЕНИЕ ЦИФРОВЫМИ АКТИВАМИ

Поскольку требуемый функционал, по сути, не отличается, многие системы управления документами успешно используются и для управления файлами с цифровым медиаконтентом, включая аудио- и видеозаписи, фотографии и т. п. Задачи включают каталогизацию, хранение и извлечение цифровых активов.

3.1.1.2 ОБРАБОТКА ИЗОБРАЖЕНИЙ

Системы обработки изображений позволяют фиксировать и редактировать образы бумажных и электронных документов, а также управлять полученным контентом. При этом используются такие технологии, как сканирование, оптическое/интеллектуальное распознавание символов и обработка форм. Пользователи могут индексировать полученные файлы, снабжая их системными метаданными перед сохранением.

Наряду с программами, основанными на использовании традиционного алгоритма OCR, которые позволяют конвертировать в электронный формат отсканированные печатные документы, также появились продвинутые технологии интеллектуального распознавания символов (ICR), способные работать как с печатными документами, так и с рукописными записями. Обе технологии крайне важны для перевода в форматы, поддерживаемые системой управления контентом (CMS) больших объемов неструктурированных данных.

Обработка форм заключается в сканировании заполненных табличных анкет с последующим распознаванием символов и оцифровкой данных в полях ввода. Для обработки отсканированной заполненной стандартной формы (например, загруженной через веб-сайт) системе нужно знать разметку верстки, структуру, логический формат и тип содержания полей.

Помимо образов (сканов) документов, в репозиториях могут храниться и другие оцифрованные изображения — фотографии, инфографика, спутниковые и географические карты и т. п. Некоторые системы управления электронным контентом (ECM) способны «переваривать» множество разнообразнейших форматов оцифрованных документов и мультимедиа, включая контент

лазерных дисков, аудиофайлы (*.wav, *.wmv, *.mp3 и др.), документы XML, медицинскую информацию в формате HL7 и другие отраслевые форматы обмена сообщениями, и включать всё это в интегрированное хранилище.

Изображения часто изначально создаются с помощью различных компьютерных приложений или снимаются на цифровые камеры и не требуют сканирования и оцифровки. В зависимости от формата файлы изображений могут содержать векторную или растровую (она же пиксельная, побитовая или точечная) графику, а также документы — например, в формате MS Word *.doc или *.docx — с внедренным изображением. Векторная графика основана на использовании математических формул для описания геометрии рисунка и обеспечивает возможность создания рисунков, допускающих масштабирование без ущерба для качества изображения, что выгодно отличает векторную графику от пиксельной. Форматы файлов векторных рисунков включают .AI, .EPS, .PDF, .SVG, .WMF и др. Растровые изображения содержат фиксированное число цветных пикселей (точек), из которых складывается картинка. Соответственно, увеличить их размер без ущерба для разрешения невозможно. Пример форматов — .JPEG, .GIF, .PNG, .TIFF.

3.1.1.3 СИСТЕМА УПРАВЛЕНИЕ ЗАПИСЯМИ

Система управления записями может предлагать функциональность автоматизации хранения и удаления, поддержку поиска запрашиваемой для раскрытия электронной информации и долговременного архивного хранения в соответствии с установленными нормативно-правовыми требованиями. Также она должна поддерживать работу программы управления жизненно важными записями, необходимыми для оперативного восстановления критических для ведения бизнеса записей после сбоев. Система управления записями может быть интегрирована с системой управления документами.

3.1.2 Система управления контентом

Система управления контентом (CMS) используется для сбора, упорядочения, индексирования, добавления и выдачи контента. При этом могут поддерживаться как сохранение и извлечение документов целиком, так и управление отдельными компонентами контента документов без нарушения целостности документов и внутренних связей между компонентами. CMS может также обеспечивать средства изменения (редактирования) контента документов, который находится под их управлением, но при этом система управления контентом, в целом, остается независимой от систем управления хранилищами документов.

CMS управляют контентом на протяжении всего его жизненного цикла. В частности, любая система управления контентом веб-сайта контролирует его с помощью инструментария авторской разработки, совместной работы и управления контентом основного хранилища. Дополнительно могут поддерживаться надстройки для удобства создания, обработки и изменения контента, а также функции дифференцированного развертывания контента для интранета, интернета и внешних клиентских приложений (расширенной интрасети). Функции доставки

контента подписчикам могут включать гибкий дизайн в зависимости от пользовательских настроек и поддержку широкого спектра типов клиентских устройств. Дополнительные функциональные модули или компоненты могут поддерживать поиск, компоновку документов, работу с электронными подписями, семантический анализ контента и доступ к CMS через мобильные приложения.

3.1.3 Потоки работ по обработке контента и документов

Средства поддержки потоков работ (workflow) настраиваются в соответствии с бизнес-процессами и позволяют маршрутизировать потоки контента и документов, распределять задачи между сотрудниками, отслеживать статус их выполнения, создавать и вести журналы аудита. Должна быть предусмотрена возможность определения процедур обязательного рецензирования и утверждения контента перед публикацией.

3.2 Инструменты поддержки совместной работы

Средства поддержки совместной работы позволяют членам команды собирать, сохранять, обрабатывать и организовывать документы и контент, необходимый для реализации текущих проектов. Взаимодействуя через социальные сети, отдельные участники и команды делятся документами и контентом внутри своих групп и доносят его до сведения внешних (целевых) групп через блоги, вики-ресурсы, RSS и теги.

3.3 Инструменты управления контролируемыми словарями и метаданными

Средства, помогающие разрабатывать контролируемые словари и метаданные или управлять ими, варьируются в широком спектре и могут включаться в состав самых разнообразных пакетов программных продуктов, в том числе офисное ПО, системы управления репозиториями метаданных, средства бизнес-аналитики и системы управления документами и контентом. Примеры включают:

- модели данных, используемые в качестве справочных руководств по классификации и структурированию данных организации;
- системы управления документами и пакеты офисных приложений;
- репозитории метаданных, глоссарии, справочные каталоги и т. п.;
- таксономии и схемы перекрестных ссылок между таксономиями;
- индексированные предметные указатели по категориям (например, по продукту, рынку или конфигурации), файловым системам, опросам, архивам, локализациям или офлайновым ресурсам;
- информационно-поисковые системы;
- средства бизнес-аналитики, поддерживающие обработку неструктурированных данных;
- тезаурусы предприятия и подразделений;
- библиотеки опубликованных отчетов, аннотации, оглавления, библиографии и каталоги.

3.4 Стандартные форматы разметки и обмена

Компьютерные приложения не могут работать с неструктурированными данными или контентом напрямую. Для обеспечения совместимости различных информационных систем по данным и возможности обмена ими через интернет используются стандартные языки и форматы разметки и обмена данными.

3.4.1 XML

Расширяемый язык разметки (XML) позволяет представлять как структурированные, так и неструктурированные данные и информацию. XML использует метаданные для описания содержания, структуры и бизнес-правил любого документа или базы данных.

Перевод данных в формат XML-документа обеспечивает возможность обмена ими между системами. В XML элементы данных маркируются таким образом, чтобы однозначно идентифицировался смысл данных. Отношения между элементами данных задаются посредством простых вложений и ссылок.

Пространства имен XML позволяют избежать конфликтов имен между различными документами, включающими идентичные имена элементов. Имеются и по-прежнему широко используются и более старые языки разметки: достаточно упомянуть HTML и SGML. Однако потребность именно в предлагаемых XML функциональных возможностях по управлению контентом назрела по ряду серьезных причин.

- XML позволяет включать неструктурированные данные наряду со структурированными в реляционные модели в качестве BLOB (больших двоичных объектов) или XML-файлов и управлять ими с помощью стандартных реляционных СУБД.
- XML позволяет интегрировать структурированные данные в неструктурированные документы, отчеты, e-mail, изображения, графики, аудио и видео (и об этом важно помнить проектировщикам моделей данных, чтобы не забывать включать эти продукты в процессы учета и исправления ошибок обработки, резервного копирования и архивирования).
- XML позволяет строить корпоративные порталы классов B2B (бизнес для бизнеса) и B2C (бизнес для клиента) для предоставления их пользователям единой точки доступа к разнообразному контенту.
- XML обеспечивает идентификацию и маркировку неструктурированных данных и/или контента понятным для компьютерных приложений образом. Благодаря этому приложения получают возможность обрабатывать неструктурированный контент и сопоставлять его со структурированными данными. Стандарт обмена метаданными на XML (XMI) определяет правила генерирования XML-документа по актуальным метаданным, задающим его структуру.

3.4.2 JSON

JSON (JavaScript Object Notation) — стандарт предельно облегченного формата обмена данными. Будучи формально независимым от языков и легко читаемым, синтаксически он всё-таки тяготеет

к семейству C-языков программирования. Формат JSON включает два типа структур: неупорядоченный набор объектов формата имя: значение и упорядоченный массив значений. В последнее время формат JSON всё чаще используется в веб-ориентированных базах данных NoSQL.

Формат JSON может использоваться вместо XML и в обмене данными между сервером и веб-приложением. Структурно форматы JSON и XML похожи, но JSON компактнее и проще в прочтении и интерпретации. При использовании архитектурных решений на базе REST-технологий выдается контент, отформатированный и в XML, и в JSON.

3.4.3 Модель RDF и стандарты W3C

Разработанная Консорциумом Всемирной паутины (W3C) «Модель описания ресурсов» (Resource Description Framework, RDF) позволяет стандартизировать не только описания веб-ресурсов, но и обмен данными между ними в глобальных масштабах. Данные о включенных в RDF ресурсах сохраняются в единой базе данных триплетов, используемой для выдачи ссылок на ресурсы в ответ на семантические запросы на языке SPARQL.

В модели RDF любой ресурс описывается в рамках множества утверждений семантической структуры «субъект — предикат — объект», где субъект — имя описываемой сущности (ресурса), предикат — имя свойства, характеристики или отношения, а объект — значение свойства предиката. Обычно каждый из элементов триплета «субъект — предикат — объект» описывается адресом URI (Uniform Resource Identifier), однако субъект или объект (но не оба одновременно) может быть представлен так называемым незаполненным узлом (blank node), который называется также анонимным ресурсом. Кроме того, допускается использование в качестве объекта литерала (безадресной текстовой строки). Неопределенные или пустые предикаты недопустимы, поскольку именно предикат URI определяет смысловую связь между двумя ресурсами. Самая распространенная и общеизвестная разновидность URI — адрес URL (Uniform Resource Locator). Именно через URL-адреса всевозможные приложения получают совместный доступ к структурированным и частично структурированным данным.

Семантическая паутина требует доступа не только к данным как таковым, но и к определениям связей между различными множествами данных, описывающих отношения между элементами. Полный набор взаимосвязанных множеств данных называют связанными данными. Семантика URI позволяет однозначно идентифицировать любую уникальную сущность. Язык HTML служит средством структурирования и привязки веб-документов. Модель RDF описывает все данные, имеющиеся во Всемирной паутине, как единую структуру графов с узлами (объектами/субъектами) в вершинах и предикатами (отношениями) на ребрах — и тем самым увязывает между собой данные обо всем сущем[1].

Синтаксически модель RDF использует XML в качестве языка определения кодов. Метаданные рассматриваются как обычные элементы данных (автор, дата создания и т. д.). Совокупность вышеописанных свойств модели RDF позволяет использовать ее для придания семантических

[1] Под «всем сущим» здесь имеется в виду всё сущее во вселенной Всемирной семантической паутины W3C. — *Примеч. пер.*

смыслов сетевым ресурсам. Схемой RDF (сокращенно RDFS) называют структурированный словарь логической RDF-модели данных, являющийся расширением базового словаря концептуальной модели RDF.

SKOS (Simple Knowledge Organization System)[1] представляет собой семантическую модель структуры тезаурусов RDF (иными словами, проекцию модели данных RDF на иерархическую структуру терминологических понятий). Модель SKOS настолько универсальна, что в ее рамках можно представить любую классификацию, таксономию или тезаурус, не говоря уже о простых словарях.

OWL (W3C Web Ontology Language)[2] является семантическим расширением модели RDF и служит языком разметки публикуемых и распространяемых через семантическую паутину OWL-документов (онтологий). В отличие от базовой модели RDF, описывающей ресурсы, ориентированные на конечных пользователей, язык онтологий OWL используется, как правило, для генерирования документов, предназначенных для машинной обработки приложениями, а не людьми. И RDF, и OWL утверждены W3C в качестве стандартов семантической паутины и образуют единую рамочную модель обмена данными и их многократного использования, обеспечивающую интеграцию, согласованность и совместимость данных с различными веб-приложениями в масштабах глобальной семантической паутины.

RDF также помогает справляться с избыточной разнородностью характеристик больших данных. Если данные из различных источников доступны через триплеты RDF-модели, они допускают слияние в единый пул («озеро»), а обращенные к ним запросы на языке SPARQL позволяют затем отыскивать связи и закономерности без задания какой-либо предопределенной схемы. Вот как это описывается консорциумом W3C: «RDF обладает функциональностью, которая позволяет с легкостью производить слияние данных даже в тех случаях, когда они структурированы по принципиально различным схемам, и даже в явном виде поддерживает эволюцию схем во времени без отключения всех потребителей изменяемых данных»[3]. Также RDF поддерживает интеграцию разрозненных данных в различных форматах из множества рассогласованных источников с последующим редуцированием (удалением избыточных данных) или заменой исходных наборов данных на семантически согласованные и оптимизированные. То есть речь идет, по сути, о синтезе данных (см. главу 14).

3.4.4 Семантическая разметка Schema.org

Унификация семантической разметки контента — в частности, по схеме, предложенной в рамках открытого проекта Schema.org, — упрощает информационно-поисковым системам задачу индексирования содержимого веб-страниц, а поисковым роботам-обходчикам (так называемым «веб-паукам») в составе этих систем — сопоставление контента с поисковым запросом.

[1] «Простая система организации знаний» (*англ.*). — *Примеч. пер.*

[2] «Язык веб-онтологии» (*англ.*). — *Примеч. пер.*

[3] W3C, «Resource Description Framework (RDF)», http://bit.ly/1k9btZQ.

Schema.org предлагает собрание иерархически организованных словарей общеупотребимых терминов и типов данных, помогающих унифицировать схемы разметки веб-страниц для упрощения их понимания ведущими поисковыми системами. Основное внимание уделяется значению слов на веб-страницах, а также определению терминов и ключевых слов.

Текст, выдаваемый поисковой системой под каждым найденным результатом, — так называемый сниппет[1] — обычно содержит пример найденного контекстного совпадения. Под избранными результатами появляются расширенные сниппеты (rich snippets) с дополнительными деталями информации о содержимом найденной страницы (включая иногда также золотые звезды рейтинга) под гиперссылкой. Чтобы поисковые системы выдавали расширенные сниппеты, контент веб-страниц должен быть надлежащим образом отформатирован с использованием структурированных данных типа Microdata (набор стандартных тегов, введенный в версии HTML5) и общеупотребимых словарей, опубликованных на Schema.org.

Словари Schema.org можно использовать также и для обеспечения совместимости структурированных данных (например, с JSON).

3.5 Технологии e-discovery

Чтобы найти вдруг потребовавшуюся архивную запись, часто приходится поднимать целые горы документов. Технологии e-discovery предлагают множество полезных функций и приемов, включая раннее выявление информации, которая может потребоваться по заведенному делу, сбор, идентификацию, сохранение, обработку, оцифровку, в том числе с распознаванием символов (OCR), сортировку, отбраковку, выявление и сопоставление дублирующих друг друга записей, анализ цепочек переписки по электронной почте и многое другое. Ревизия с помощью технологий (Technology-Assisted Review, TAR) должна носить характер тщательно структурированного потока работ, в рамках которого команда проводит сплошную проверку ряда типичных документов, предварительно отобранных как потенциально содержащие искомую информацию, и маркирует каждый из них как относящийся или не относящийся к делу. Эти результаты служат входными материалами для программирования алгоритма прогностического поиска и сортировки по релевантности остальных документов. Может также поддерживаться руководство информацией.

4. МЕТОДЫ

4.1 Сценарий подготовки электронной доказательной базы

К розыску релевантных электронных документов (выполнению процедуры e-discovery) приступают сразу же, как становится известно о подаче иска или возбуждении дела. Впрочем, организация может начать готовиться к защите своих интересов в суде или предъявлению подтверждений

[1] Сниппет (калька с *англ.* snippet) — *досл.* обрывок, лоскут; *вчт.* фрагмент кода или текста. — *Примеч. пер.*

своей законопослушности надзорным органам загодя, разработав сценарный план реакции на возбуждение дела, в котором будут расписаны задачи, измеримые показатели и ответственные, чтобы к любому крупному проекту по поднятию архивов в поисках требуемых документов и записей быть готовой подойти ко всеоружии.

Типичный сценарий определяет целевую среду e-discovery, требования к ней и предусматривает оценку соответствия текущего состояния среды этим требованиям. В нем документируются бизнес-процессы, относящиеся к жизненному циклу разыскиваемых электронных данных, определяются направления работ, роли и ответственные. Сценарий может также включать общеорганизационные мероприятия по выявлению рисков и активному профилактическому предупреждению ситуаций, чреватых неприятными юридическими последствиями.

Для составления сценария следует:

- инвентаризировать политики, правила и процедуры, установленные для каждого из затрагиваемых отделов (юридического, делопроизводства, архивного, ИТ и т. п.);
- составить проекты политик по таким предметам регулирования, как действия в случае получения судебного предписания о бессрочном сохранении подтверждающих записей, хранение, архивирование и резервное копирование документов;
- оценить функциональные возможности инструментальных средств индексирования результатов розыска электронных записей, поиска, сбора, сортировки и защиты данных, а также источников и/или систем сбора неструктурированной информации электронного хранения (ESI);
- определить процедуру всестороннего анализа юридических аспектов дела;
- разработать план информационно-разъяснительной работы и дополнительной подготовки сотрудников к возможным событиям подобного рода;
- разработать порядок выявления материалов, которые можно подготовить заранее, не дожидаясь начала формального рассмотрения дела компетентными органами;
- проанализировать возможных сторонних поставщиков услуг на случаи, когда такие услуги могут потребоваться;
- разработать правила и процедуры действий при получении предписаний, а также периодического пересмотра и обновления версии самого сценарного плана.

4.2 Карта данных, которые могут быть найдены и представлены

Часто на раскрытие и представление электронных данных отводится весьма ограниченное время (например, 90 календарных дней). Чем скорее юристы получат хотя бы общую карту данных, имеющихся в информационных системах и архивах ESI вашей организации, тем эффективнее они сумеют выстроить линию защиты ее интересов. Карта данных представляет собой полный каталог информационных систем и их контента с детальным описанием назначения каждой системы, содержащейся в ней информации, правил хранения и прочих характеристик. В таких каталогах часто указываются все системы записи, приложения-источники, архивы, резервные копии

данных на случай аварийного восстановления и т. п. с указанием среды или носителя, где они хранятся. Важно, чтобы карта данных была исчерпывающей и охватывала все без исключения системы. Поскольку в процессуальных действиях часто фигурирует в качестве улик или доказательств переписка по электронной почте, карта данных должна описывать и порядок хранения, обработки использования и уничтожения e-mail-переписки. Сопоставление бизнес-процессов со списком систем и документирование распределения ролей и привилегий также является необходимым этапом документального отображения информационных потоков.

Процесс картирования данных позволяет еще раз убедиться в незаменимости метаданных как ценнейшего ресурса управления документами. Без метаданных отыскать что-либо крайне проблематично. Именно они позволяют осуществлять контекстный поиск электронных документов, относящихся к конкретным событиям или содержащих протоколы, стенограммы, подтверждения и т. п., которые можно приобщить к делу.

В карте данных, представляемой по юридическому запросу электронных документов в рамках процедуры e-discovery, должно быть указано, какие записи готовы для предъявления по первому требованию, а какие в настоящее время в электронной форме недоступны. К двум этим категориям затем применяются различные правила раскрытия электронных данных. В перечне выявленных недоступных данных должны быть детально зафиксированы причины невозможности получить доступ к ним. Для надлежащего ответа в рамках судопроизводства организация обязана представить полный инвентарный перечень записей, физически находящихся на хранении за ее пределами, включая опись данных в облачных хранилищах.

Что касается систем, то их инвентарные описи зачастую уже имеются, например, у специалистов по проектированию архитектуры данных, управлению метаданными или управлению ИТ-ресурсами. В этом случае юридическому и/или архивному подразделению достаточно определить, не требуют ли они дополнения какими-либо еще записями и документами с целью обеспечения полноты раскрытия затребованных электронных данных.

5. РЕКОМЕНДАЦИИ ПО ВНЕДРЕНИЮ

Реализация управления корпоративным контентом (ECM) требует долгосрочных усилий и может показаться слишком дорогостоящим начинанием. Как и всякое направление деятельности в масштабах всей организации, ECM требует заинтересованности и усилий от широкого круга лиц, а также адекватного финансирования из средств, выделяемых высшим руководством. В случае крупного проекта всегда есть риск срыва его реализации из-за урезания финансирования, резких изменений в политике или руководстве, да и просто в силу инерционности организации. Для минимизации рисков следите за тем, чтобы решения по реализации ECM фокусировались прежде всего на содержательной, а не технологической стороне управления данными и контентом. И выстраивайте рабочий процесс вокруг организационных нужд, чтобы всегда была видна его ценность.

5.1 Оценка готовности / Оценка рисков

Оценка готовности организации к внедрению ECM нужна для выявления слабых мест в управлении контентом и определения мер по изменению и совершенствованию управления записями и документами, необходимых для удовлетворения выявленных нужд. Для проведения такой оценки вполне подходит модель оценки зрелости управления данными (см. главу 15).

Ряд ключевых факторов успеха ECM совпадают с аналогичными факторами успеха ИТ-проектов (поддержка высшим руководством, заинтересованность и участие пользователей, переподготовка, управление изменениями, в том числе корпоративной культуры, обеспечение двусторонней связи и взаимопонимания). Специфичные для ECM критические факторы успеха включают учет, аудит и классификацию имеющегося контента, разработку архитектурных решений информационных систем управления контентом на протяжении его жизненного цикла, определение тегов метаданных, обеспечение возможности гибкой настройки функций ECM-решения. По причине технической сложности ECM-решений и реализованных с их помощью рабочих процессов организации нужно выделять достаточные ресурсы на поддержку реализации подобного проекта.

Риски при внедрении ECM могут быть обусловлены масштабностью проекта, сложностью интеграции с другими прикладными системами, административно-процедурными и организационными проблемами, сложностью и трудоемкостью переноса контента в новую электронную среду. Недостаточная подготовленность ключевых членов проектной группы и персонала, в целом, могут привести к несогласованности действий и трактовок правил использования контента. Кроме того, всегда имеются риски неудачной или недостаточной проработки политик, правил, процессов и процедур, а также неспособности достигнуть взаимопонимания со всеми заинтересованными лицами.

5.1.1 Оценка зрелости управления записями

Общепринятые принципы ведения записей (GARP), разработанные ассоциацией ARMA International (см. раздел 1.2), помогают организации оценивать собственную политику и практику управления записями. Помимо GARP у ARMA International имеется Модель зрелости руководства информацией[1], помогающая организации производить экспертную оценку своей программы и практик ведения и хранения записей. Модель описывает характерные признаки пяти уровней зрелости организационной среды руководства распоряжения информацией и ведения записей с учетом всех восьми принципов GARP.

- **Уровень 1 (не соответствующий стандартам)**: вопросы руководства информацией и учета записей вовсе не регулируются или решаются в минимальных объемах и/или спорадически.
- **Уровень 2 (в разработке)**: формируется понимание важности для организации высокоуровневого руководства информацией и упорядоченного ведения записей.

[1] ARMA International, Information Governance Maturity Model.

- **Уровень 3 (базовый)**: обеспечено соблюдение минимальных требований, установленных законодательством и/или надзорными органами.
- **Уровень 4 (опережающее развитие)**: реализована программа руководства информацией, ориентированная на непрерывное совершенствование процессов.
- **Уровень 5 (трансформационный)**: руководство информацией интегрировано в корпоративную инфраструктуру и бизнес-процессы.

Для технической экспертизы соответствия систем и приложений, используемых для управления записями, можно использовать различные отраслевые стандарты, спецификации или регламенты, например:

- DoD 5015.2, Electronic Records Management Software Applications Design Criteria Standard — «Стандарт критериев разработки ПО для управления электронными записями» МО США;
- ISO 16175, Principles and Functional Requirements for Records in Electronic Office Environments — «Принципы и функциональные требования к записям в электронной офисной среде»;
- The Model Requirements for the Management of Electronic Records — «Типовые требования к автоматизированным системам электронного документооборота» Еврокомиссии (MoReq2);
- Records Management Services (RMS) — «Службы управления записями», спецификации, разработанные Object Management Group® (OMG)[1].

Выявленные по результатам оценки готовности организации недостатки и риски в сфере управления записями подлежат тщательному анализу на предмет оценки проистекающих от них потенциальных угроз благополучию организации. Несоблюдение установленных законами требований по надлежащему и безопасному хранению, ведению и уничтожению записей из-за отсутствия надлежащих систем их учета — само по себе представляет для организации серьезный риск, поскольку никто в ней не может с определенностью утверждать, что какие-то важные записи не были похищены или утеряны. Кроме того, при отсутствии реально функционирующей системы учета записей поиск нужных записей может обернуться неоправданно высокими затратами времени и средств. Несоблюдение законодательства и требований надзорных органов чревато крупными штрафами, а неспособность выявлять жизненно важные записи и обеспечивать их защиту — выбыванием компании из бизнеса.

5.1.2 Оценка программы электронному раскрытию информации

В рамках оценки готовности должна проводиться комплексная экспертиза имеющейся у организации программы поиска документов и записей, запрашиваемых судебными, арбитражными или надзорными инстанциями. В зрелой программе должны быть четко расписаны роли

[1] OMG® — международная некоммерческая организация по «разработке единых технологических стандартов для вертикально выстроенных отраслей» (см.: omg.org). — *Примеч. пер.*

и обязанности, протоколы сбора, хранения и раскрытия информации электронного хранения (ESI). И сама программа, и предусмотренные ею процессы и процедуры должны быть задокументированными, юридически весомыми и проверяемыми.

В основе программы должно лежать понимание жизненного цикла информации в рамках организации. Обязательным ее компонентом является тщательно проработанная карта соотнесения всех категорий ESI с источниками данных (см. раздел 2.1.3.4). Поскольку вся ответственность за обеспечение сохранности данных, обязательных к раскрытию, по закону возложена на организацию, политики ИБ должны регулярно и активно пересматриваться и обновляться с целью обеспечения постоянной готовности к раскрытию ESI любого рода по требованию компетентных органов. Также должен иметься четкий план оперативного взаимодействия со специалистами по ИТ на случай поступления судебного предписания о сохранении каких-либо данных ESI сверх установленного срока.

Риски, проистекающие от отсутствия заранее проработанных процедур реагирования на судебные запросы и предписания, подлежат выявлению и количественной оценке. Иногда организации озабочиваются подобными вопросами лишь по факту получения запроса или в преддверии неизбежного судебного или арбитражного разбирательства, и в таких случаях поиск и изучение релевантных документов и записей зачастую выливается в сущий переполох и неразбериху, что свидетельствует, как минимум, об одном из двух: либо в организации ведутся архивы явно избыточных данных (то есть хранится всё подряд), либо отсутствуют механизмы обеспечения соблюдения правил и сроков хранения архивных данных и их уничтожения по истечении срока хранения. Отсутствие утвержденного плана-графика хранения и ликвидации данных и информации по категориям к тому же чревато привлечением организации к юридической ответственности как в случае выявления у нее сохраненных данных, которые по закону должны были быть уничтожены, так и за неспособность предоставить данные, которые должны были быть в сохранности.

5.2 Организационные и культурные изменения

С людьми нередко возникает больше сложностей, чем с технологиями. Возможны проблемы и с адаптацией к новым правилам управления текущей работой, и с выработкой привычки к электронному документообороту и управлению контентом (ECM). В некоторых случаях переход на ECM приводит к росту нагрузки и появлению дополнительных задач — например, по сканированию и оцифровке бумажных документов или определению обязательных метаданных.

Организации часто управляют информацией, включая электронные записи, на уровне отдельных подразделений, что приводит к размежеванию и несогласованности, появлению изолированных друг от друга «информационных бункеров» (information silos), мешающих надлежащему обмену и управлению данными. Целостный подход к управлению контентом и записями помогает избавить пользователей от ложного представления о том, что они обязаны сохранять у себя копии всех без исключения документов или записей, которые через них проходят. Идеальным решением является единое хранилище данных под централизованным управлением, надежно

защищенное средствами обеспечения ИБ и регулируемое четкими политиками и процедурами, контроль соблюдения которых ведется единообразно в масштабах всей организации. Учебно-методическая и информационно-разъяснительная работа с целью усвоения всеми сотрудниками утвержденных процессов и правил, а также навыков обращения с техническими средствами и инструментами — критически важный фактор успешного управления записями или реализации программы ECM.

Защита персональных и личных данных, конфиденциальной информации и интеллектуальной собственности, правила шифрования и этичного обращения с данными, гарантии их подлинности и аутентичности, идентификация и доступ пользователей, — все эти важнейшие вопросы управления документами и контентом должны решаться профессионалами в области управления данными комплексно, в тесном взаимодействии с другими сотрудниками и представителями администрации и по согласованию с надзорными органами. Централизованным организациям часто приходится заботиться о совершенствовании процессов доступа к информации и контроля всевозрастающих объемов документов на материальных носителях, загромождающих офисное пространство, а также о снижении операционных издержек, минимизации судебно-юридических рисков, надежной защите жизненно важных данных и информационном обеспечении процессов принятия решений.

Управление как контентом, так и записями нужно поднять на качественно новый организационный уровень, ведь в современных условиях их никак нельзя относить к чисто техническим и низкоприоритетным функциям. В сильно зарегулированных отраслях функция управления записями и информацией (Records and Information Management, RIM) должна быть реализована в тесном согласовании с корпоративной юридической службой и дополнена функцией раскрытия электронных документов и записей по запросам компетентных органов (e-discovery). Если организация действительно нацелена на повышение эффективности и производительности за счет совершенствования оперативного управления данными, RIM должно также согласовываться с функциями маркетинга и сбыта и/или оперативной поддержки. Если организация относит RIM к области ИТ-обеспечения, непосредственное руководство RIM должно осуществляться CIO или CDO. Распространенный альтернативный вариант — включение функции RIM в корпоративную программу управления контентом (ECM) или информацией (EIM).

6. РУКОВОДСТВО УПРАВЛЕНИЕМ ДОКУМЕНТАМИ И КОНТЕНТОМ

6.1 Рамочные структуры руководства информацией

Хранение документов, записей и прочей неструктурированной информации (контента) сопряжено с риском для организации, которая ими располагает. Для управления этим риском, как и для извлечения максимальной пользы из неструктурированной информации, в организации должна иметься единая система руководства. Стимулом к ее созданию выступает, в частности, необходимость обеспечивать:

- соблюдение требований законодательства и регламентирующих документов;
- гарантированное и юридически доказуемое уничтожение записей по истечении срока хранения;
- превентивное обеспечение готовности к изъятию или раскрытию электронных данных (процедуре e-discovery);
- защиту чувствительной информации;
- управление ИБ в областях повышенного риска, таких как электронная почта и большие данные.

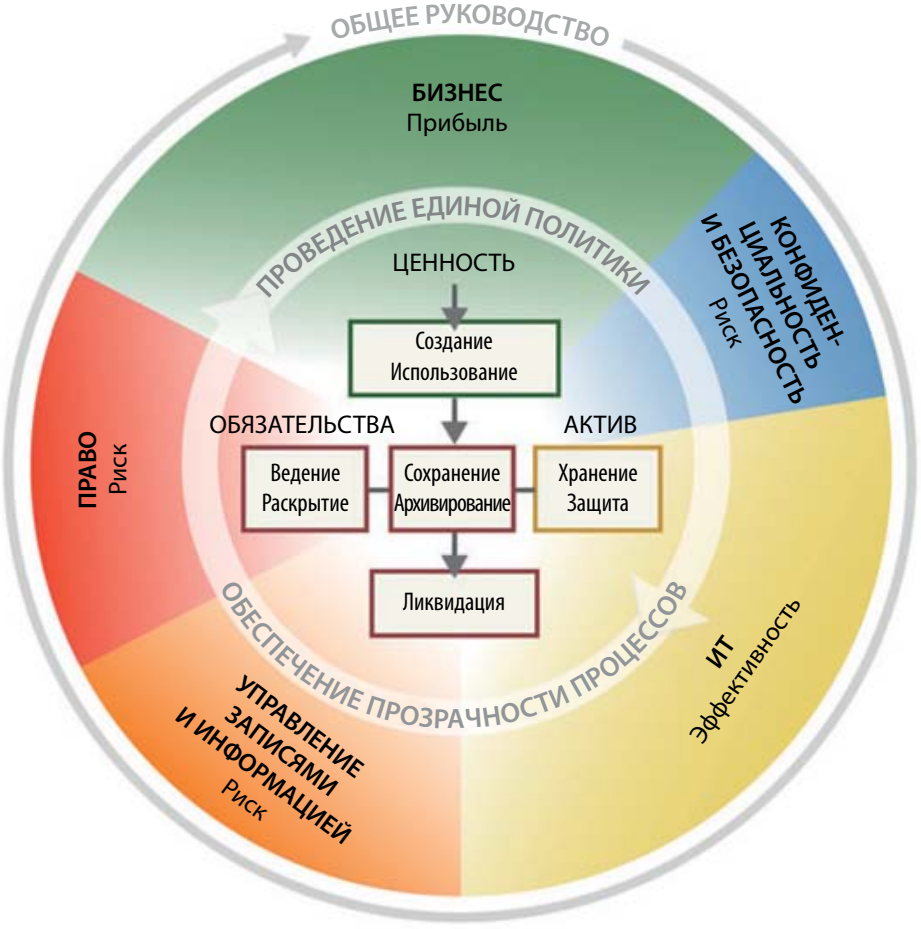
Проработка принципов успешного руководства информацией начала вестись совсем недавно. Один из первых наборов принципов такого рода — ARMA GARP® (см. раздел 1.2). Кроме того, можно порекомендовать придерживаться следующих принципов:

- назначить куратора программы руководства информацией из числа высших руководителей;
- провести разъяснительную работу с сотрудниками относительно их обязанностей по обеспечению ответственного руководства информацией;
- классифицировать информацию согласно корректно определенному кодификатору записей или таксономическому классификатору;
- обеспечивать аутентичность и целостность информации;
- установить правило, согласно которому все официальные записи должны вестись в электронном формате, если иное не оговорено особо;
- разработать политику и правила согласования собственных бизнес-систем со стандартами руководства информацией третьих сторон;
- контролировать надлежащее хранение, управление и доступность информации, включая мониторинг и аудит утвержденных репозиториев и систем управления записями и контентом предприятия;
- обеспечивать защиту конфиденциальных, персональных и личных данных;
- не допускать накопления избыточной информации;
- обеспечивать уничтожение записей по истечении установленного срока хранения;
- обеспечивать раскрытие информации (по предписаниям, ордерам, повесткам и т. п.);
- непрерывно совершенствовать все вышеперечисленные процессы.

Эталонная модель руководства информацией (Information Governance Reference Model, IGRM) наглядно описывает связи между руководством информацией и другими организационными функциями (см. рис. 74). Внешнее кольцо определяет стороны, обеспечивающие наличие необходимых для управления информацией политик, стандартов, процессов, процедур, инструментальных средств и инфраструктуры. В центре показана схема жизненного цикла, компоненты которого соотнесены со сторонами, отвечающими за их исполнение и/или реализацию посредством цветовой маркировки рамок. Модель IGRM служит дополнением к модели ARMA GARP®.

Эталонная модель руководства информацией (IGRM)

выполнение обязательств + ценность информационного актива = эффективность



Обязательства: Юридические обязательства в отношении конкретной информации

Ценность: Необходимость или полезность конкретной информации для бизнеса

Актив: Объект (контейнер), содержащий информацию

Information Governance Reference Model / © 2012 / v3.0 / edrm.net

Рисунок 74. Эталонная модель руководства информацией (IGRM)[1]

[1] *Источник*: EDRM, по лицензии Creative Commons Attribution 3.0.

Патронаж со стороны кого-то из высших руководителей или вхожих в «ближний круг» — необходимое условие инициирования, создания и обеспечения долговременного устойчивого функционирования программы руководства информацией. На высшем уровне управления организацией должен быть сформирован кросс-функциональный Совет или Управляющий комитет по руководству информацией, а рабочие заседания этого органа должны проводиться на регулярной основе. Совет или Комитет вырабатывает корпоративную стратегию руководства информацией, определяет порядок работы, дает общие указания относительно выбора технологий и стандартов, организации информационно-разъяснительной работы и профессиональной подготовки, определяет порядок мониторинга и объемы финансирования. Политики руководства информацией разрабатываются применительно к каждой области, описываемой моделью IGRM, и (в идеале) реализуются с использованием технологических средств контроля их соблюдения.

6.2 Рост объемов информации

У неструктурированных данных есть свойство множиться и накапливаться значительно более высокими темпами, чем структурированные, и это дополнительно осложняет управление документами и контентом. К тому же неструктурированные записи зачастую не поддаются строгому отнесению к какой-либо бизнес-функции или подразделению. Трудно бывает определить и персонально ответственного за ту или иную единицу хранения информации. Сложно навскидку выявить и классифицировать содержание неструктурированных данных, поскольку их назначение или смысл сохранения с точки зрения ведения бизнеса в информационных системах зачастую никак не описан. Неуправляемые неструктурированные данные, не сопровождаемые требуемыми метаданными, представляют собой серьезный риск. Во-первых, в отрыве от контекста они могут быть превратно интерпретированы, а во-вторых, из-за неосведомленности об их истинном содержании кто-то из имеющих доступ к таким данным сотрудников может непредумышленно допустить утечку или разглашение персональных или конфиденциальных данных (см. главу 14).

6.3 Управление качеством контента

Управление неструктурированными данными требует эффективного взаимодействия между распорядителями данных, другими специалистами по управлению данными и менеджерами контента. Например, распорядители бизнес-данных могут оказывать помощь веб-разработчикам в плане определения требуемой структуры порталов, корпоративных таксономий, индексов для поисковых систем и общих решений по управлению контентом.

Функция руководства документами и контентом должна уделять особое внимание определению политик в области хранения, электронных подписей, форматов и распространения отчетности. Политики формулируют явные или подразумеваемые ожидания относительно качества контента. Точность, полнота и актуальность информации важны для грамотного принятия решений. Обладание высококачественной информацией дает бизнесу преимущество перед конкурентами и служит залогом эффективности работы организации. Для определения критериев качества

контента требуется понимание контекста его производства и использования. В частности, желательно располагать следующими сведениями о каждом документе, файле или записи.

- **Авторы/продюсеры/производители**: кто и зачем создал и/или опубликовал этот контент?
- **Потребители**: кто и зачем использует эту информацию?
- **Сроки и частота использования**: когда бывает или будет востребована эта информация? Как часто она запрашивается или используется? Нуждается ли в обновлении? Если да, то с какой периодичностью?
- **Формат**: соответствует ли формат контента запросам потребителей и целям, в которых предполагается его использовать? Нет ли среди контента документов, файлов или записей в недопустимых форматах?
- **Доставка/выдача**: каким образом эта информация будет выдаваться или доставляться подписчикам или пользователям? Каков порядок доступа потребителей к контенту? Как будет обеспечиваться его защита от несанкционированного доступа?

6.4 Метрики

Ключевые показатели эффективности (Key Performance Indicators, KPI) организации используют как количественные, так и качественные критерии оценки соответствия ее реальной работы поставленным целям. KPI могут разрабатываться как на стратегическом, так и на операционном уровнях, включая отдельные показатели, применимые на обоих уровнях, особенно в тех случаях, когда речь идет об оценке функций управления жизненным циклом или рисками.

6.4.1 Управление записями

На стратегическом уровне можно разработать KPI, отражающие положение дел в таких областях, как соблюдение нормативно-правовых требований и отраслевых регламентов управления записями (в частности, по срокам раскрытия и хранения) или внутриорганизационных административных требований (в частности, соблюдения установленных политик и правил). На операционном уровне KPI могут быть определены в таких областях, как ресурсоемкость управления записями (например, операционные издержки и капитальные затраты), подготовка персонала (число занятий и прошедших дополнительное обучение сотрудников по уровням и подразделениям), соблюдение графиков ежедневной публикации или выдачи информации службами управления записями (процент случаев несоблюдения условий соглашений об уровне обслуживания), показатели интеграции функций управления записями с другими бизнес-системами (процент интегрированных функций).

Измеримыми показателями успеха внедрения управления записями могут служить:

- доля (%) выявленных корпоративных записей в общем количестве создаваемых пользователями документов и e-mail;
- доля (%) выявленных корпоративных записей, формально объявленных таковыми и поставленных на учет;

- доля (%) записей (в общем количестве хранящихся записей), для которых имеются и к которым должным образом применяются строго определенные правила и сроки хранения.

Полученные процентные показатели можно сравнивать с показателями за предыдущие периоды и публикуемыми данными с целью определения оценок качества работы подразделений.

Иногда материально ощутимым свидетельством успеха реализации программы управления записями становится просто достигнутая экономия бюджетных средств. Простой финансовой расчет показывает тот переломный момент, после которого расходы на реализацию систем электронного документооборота и управления записями становятся ниже затрат на расширение площадей, требующихся для хранения бумажных архивов.

Также систему KPI можно построить в соответствии с Общепринятыми принципами ведения записей (GARP) и Моделью зрелости руководства информацией ARMA. Кроме того, ассоциация ARMA предлагает платформенное решение Information Governance Assessment для проведения экспертизы организации управления записями, которая позволяет выявить риски и разработать метрики соблюдения действующих нормативно-правовых требований и оценки зрелости программы руководства информацией в таких областях, как ведение электронных записей и готовность к раскрытию требуемой (например, в судебных целях) электронной информации.

6.4.2 Электронное раскрытие информации (e-discovery)

Одним из универсальных показателей эффективности e-discovery является снижение издержек. Другой стандартный KPI — повышение эффективности сбора требуемой информации за счет его начала еще *до* поступления официального запроса (выражающееся, например, в снижении среднего числа календарных дней, которые требуются на удовлетворение требования о раскрытии). Третий стандартный показатель — время, требующееся организации на исполнение судебного предписания о сохранении записей или документации сверх стандартного срока в юридических целях.

Регистрация скорости раскрытия требуемых электронных записей критически важна, поскольку от этого показателя напрямую зависит вероятность благоприятного для организации исхода рассмотрения дел в запрашивающих данные судебных, арбитражных или иных инстанциях. Эталонная модель электронного раскрытия (EDRM) позволяет разрабатывать системы KPI, руководствуясь описаниями того, что требуется от организации на каждой фазе поиска и подготовки данных к раскрытию. Кроме того, проект EDRM публикует и регулярно обновляет также и Модель метрик[1] для оценки готовности организации к раскрытию требуемой электронной информации. К основным показателям отнесены объем, время и стоимость, а вокруг них сгруппированы семь основных аспектов работы по выявлению и раскрытию электронной информации, которые эти итоговые показатели определяют, а именно — мероприятия, ответственные за сохранность, системы, среды хранения, статус, формат и обеспечение качества.

[1] *англ.* EDRM Metrics Model, см.: http://bit.ly/2rURq7R

6.4.3 Управление корпоративным контентом

Должны быть разработаны KPI, отражающие как материальные, так и нематериальные выгоды от ECM. К первым относятся повышение производительности, снижение затрат, повышение качества информации и производственной дисциплины. Нематериальные достижения включают улучшение сотрудничества, упрощение рабочих процедур и совершенствование организации рабочего процесса.

По мере формирования сложившейся практики ECM и выработки у сотрудников привычки к соблюдению правил и процедур управления контентом в масштабах организации KPI переориентируются на мониторинг показателей результативности программы и эффективности операционной деятельности. Метрики программы включают число реализованных проектов ECM и показатели степени реализации принципов управления контентом и удовлетворенности пользователей. Операционные метрики включают стандартные для информационных систем KPI (отказоустойчивость, число пользователей, и т. п.).

Специфические для ECM метрики, такие как загруженность хранилищ (например, текущий объем данных в системе ECM в процентном отношении к объему документов/контента в информационных хранилищах предприятия перед внедрением ECM), производительность и эффективность поисковых функционалов и/или скорость обработки запросов на выдачу контента, также могут быть включены в число KPI. Отнюдь не лишним будет и мониторинг качества и полноты выдаваемой в ответ на поисковые запросы информации. Качество определяется прежде всего процентом релевантных документов в выдаче, а полнота — процентом имеющихся в хранилищах релевантных документов, которые попадают в выдачу по запросу.

Со временем можно разработать KPI, отражающие ценность различных решений и компонентов системы ECM для бизнеса, включая:

- показатели финансовой отдачи от внедрения ECM за счет снижения затрат на хранение физических архивов и сокращения операционных издержек;
- показатели качества обслуживания клиентов/потребителей, включая число жалоб и процент проблем, устраненных или разрешенных по результатам первого же обращения в службу поддержки;
- KPI, отражающие повышение эффективности и производительности внутренних бизнес-процессов, включая процентные показатели снижения объемов бумажного документооборота, ошибок при обработке и выдаче документов, а также процентный показатель автоматизации;
- показатели подготовленности сотрудников (могут включать количество часов дополнительного обучения и результаты проверки усвоения полученных знаний, в том числе, например, раздельно по уровням ответственности и подразделениям);
- KPI в области смягчения рисков, включая снижение затрат на поиск затребованных документов, записей и материалов, а также число контрольных журналов аудита с отслеживанием всех деталей обработки запросов на получение информации электронного хранения.

7. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Boiko, Bob. *Content Management Bible*. 2nd ed. Wiley, 2004. Print.

Diamond, David. *Metadata for Content Management: Designing taxonomy, metadata, policy and workflow to make digital content systems better for users*. CreateSpace, 2016. Print.

Hedden, Heather. *The Accidental Taxonomist*. Information Today, Inc., 2010. Print.

Lambe, Patrick. *Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness*. Chandos Publishing, 2007. Print. Chandos Knowledge Management.

Liu, Bing. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. 2nd ed. Springer, 2011. Print. Data-Centric Systems and Applications.

Nichols, Kevin. *Enterprise Content Strategy: A Project Guide*. XML Press, 2015. Print.

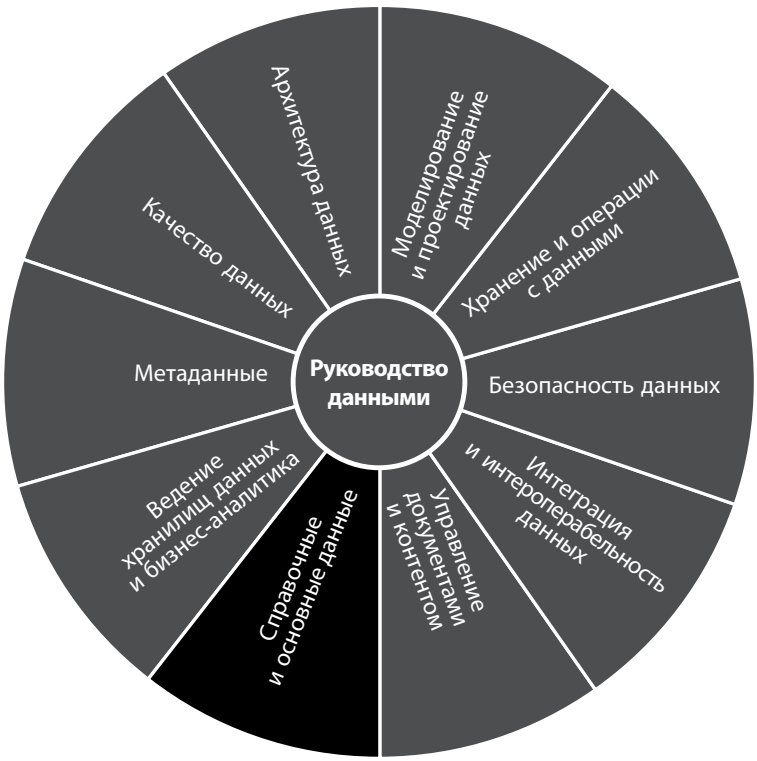
Read, Judith and Mary Lea Ginn. *Records Management*. 9th ed. Cengage Learning, 2015. Print. Advanced Office Systems and Procedures.

Rockley, Ann and Charles Cooper. *Managing Enterprise Content: A Unified Content Strategy*. 2nd ed. New Riders, 2012. Print. Voices That Matter.

Smallwood, Robert F. *Information Governance: Concepts, Strategies, and Best Practices*. Wiley, 2014. Print. Wiley CIO.

US GAAP Financial Statement Taxonomy Project. *XBRL US GAAP Taxonomies*. v 1.0 Technical Guide Document Number: SECOFM-USGAAPT-TechnicalGuide. Version 1.0. April 28, 2008, http://bit.ly/2rRauZt

Справочные и основные данные



DAMA-DMBOK2 Data Management Framework

1. ВВЕДЕНИЕ

В любой организации имеются часто используемые данные, без которых невозможно нормальное функционирование многих бизнес-процессов или систем. И организация, и ее клиенты только выигрывают, когда подобные данные централизованы и открыты для общего доступа всем нуждающимся в них сотрудникам. Данные совместного доступа могут включать списки клиентов, коды географических месторасположений, списки бизнес-подразделений, варианты поставочной комплектации, номенклатуры комплектующих, коды центров затрат, коды и ставки налогов и сборов, а также любые иные данные, необходимые для ведения бизнеса. Пользователи данных, как правило, рассчитывают на их согласованность в пределах организации, и любое столкновение с противоречащими друг другу данными становится для них пренеприятным сюрпризом.

Рисунок 75.
Контекстная диаграмма: справочные и основные данные

В большинстве организаций информационные системы и модели данных развиваются более стихийными путями, чем хотелось бы специалистам по управлению данными. Особенно часто так происходит в крупных организациях, где реализуется множество разнообразных проектов и инициатив, проводятся слияния и поглощения, закрываются старые и открываются новые направления деятельности, в связи с чем возникает множество независимых и никак не связанных между собой информационных систем, дублирующих друг друга функционально. В результате в подобных условиях неизбежно возрастает рассогласованность данных, а любые расхождения чреваты серьезными издержками и рисками, и для их снижения следует осуществлять управление *основными данными и справочными данными* (*Master Data and Reference Data*).

1.1 Бизнес-драйверы

Самыми распространенными драйверами инициирования программы управления основными данными являются следующие.

- **Выполнение требований организации к данным.** В различных областях работы организации требуются одни и те же наборы данных — и нужна уверенность в их полноте, актуальности и согласованности. Основные данные часто служат фундаментом при определении таких наборов данных (например, для планомерного и полного учета всех клиентов в аналитических выкладках необходимо четкое и последовательно применяемое определение клиента).
- **Управление качеством данных.** Противоречивые, некачественные или неполные данные приводят к неверным решениям и упущенным возможностям; управление основными данными позволяет снизить подобные риски за счет обеспечения полного и согласованного представления всех важных для организации сущностей.
- **Управление затратами на интеграцию данных.** Стоимость интеграции данных из новых источников в и без того сложную информационную среду только повышается при отсутствии качественных основных данных, необходимых для минимизации разночтений в определениях критически важных сущностей.
- **Снижение риска.** Основные данные позволяют упрощать архитектуру обмена и совместного использования данных, снижая за счет этого издержки и риски, обусловленные избыточной сложностью ИТ-среды.

Драйверы управления справочными данными, в целом, аналогичны. Централизация управления справочными данными позволяет организациям:

- наиболее полно удовлетворять информационные потребности различных проектов и инициатив за счет использования согласованных справочных данных;
- управлять качеством справочных данных.

В то время как управляемые на основе данных (data-driven) инициативы организации обычно фокусируются на транзакционных данных (поскольку ставят целью увеличение продаж или рыночной доли, снижение издержек или подтверждение соблюдения установленных требований), способность повысить отдачу от транзакционных данных сильно зависит от наличия и качества справочных и основных данных. Доступность и качество справочных и основных данных приводят к резкому повышению доверия со стороны бизнеса к данным в целом. Это приносит организации дополнительную пользу, включая упрощение ИТ-ландшафта, повышение эффективности и производительности систем и приложений, а в конечном счете — и показателей удовлетворенности клиентов.

1.2 Цели и принципы

Цели и принципы программы управления справочными и основными данными включают:

- обеспечение наличия в организации полных, согласованных, актуальных и достоверных основных и справочных данных по всему спектру процессов;
- обеспечение возможности совместного использования основных и справочных данных в рамках всех функций и приложений организации;
- снижение стоимости и сложности использования и интеграции данных за счет применения стандартов, общих моделей данных и шаблонов интеграции.

Управление справочными и основными данными должно соответствовать следующим руководящим принципам.

- **Совместное использование.** Управление справочными и основными данными должно осуществляться таким образом, чтобы обеспечивалась возможность их совместного использования в рамках всей организации.
- **Владение.** Справочные и основные данные принадлежат всей организации, а не конкретным приложениям или подразделениям. Поскольку речь идет о данных, имеющих столь широкое распространение, распоряжение ими должно осуществляться на как можно более высоком уровне.
- **Качество.** Требуется непрерывный мониторинг качества справочных и основных данных и руководство его обеспечением.
- **Распоряжение.** За контроль и обеспечение качества справочных данных отвечают распорядители бизнес-данных.
- **Контролируемые изменения**
  - В каждый момент времени основные данные должны соответствовать как можно более точному текущему представлению организации о положении дел. Правила соответствия, предполагающие внесение изменений в основные данные, должны применяться с максимальной осторожностью и под строгим контролем. Всякое слияние или разделение идентификаторов должно быть обратимым.

- Процедуры внесения изменений в справочные данные должны выполняться в рамках строго определенного процесса. Все вносимые изменения подлежат предварительному согласованию и утверждению.

- **Авторитетность источника.** Значения основных данных должны тиражироваться только с помощью единой системы записи (system of record)[1]. Для обеспечения совместного использования основных данных в масштабах организации может потребоваться применение эталонной справочной системы (system of reference).

1.3 Основные понятия и концепции

1.3.1 Различия между основными и справочными данными

Данные различных категорий играют в организации специфические роли. Соответственно, и требования к управлению ими предъявляются разные. Обычно принято проводить четкое разграничение между транзакционными и основными данными, а также между основными и справочными данными. В предложенной Малкольмом Чисхолмом шестиуровневой таксономии данных выделены следующие категории: метаданные, справочные данные, данные о структуре организации, данные о структуре транзакций (transaction structure data), данные о деятельности в рамках транзакций (transaction activity data) и данные аудита транзакций (transaction audit data) (Chisholm, 2008; Talburt and Zhou, 2015). В рамках этой таксономии основные данные определяются как объединение справочных данных, данных о структуре организации и данных о структуре транзакций.

- **Справочные данные** — например, таблицы кодов и определений; нужны исключительно для описания характеристик других данных, используемых в организации, или для соотнесения данных внутри организации с информацией за ее пределами.
- **Данные о структуре организации** — например, план счетов; позволяют организовать отчетность о бизнес-деятельности в разрезе направлений и сфер ответственности.
- **Данные о структуре транзакций** — например, идентификаторы клиентов; описывают все элементы, необходимые для проведения транзакции, — продукты, клиентов, продавцов и т. п.

Согласно определению Чисхолма, основные данные отличаются как от данных о деятельности в рамках транзакций, отражающих детали проводимых операций, так и от данных аудита транзакций, отражающих состояние транзакций, а также от метаданных, описывающих другие данные (Chisholm, 2008). В этом плане определение Чисхолма практически идентично определению, данному в Словаре DAMA (DAMA Dictionary): «Основные данные — данные, предоставляющие контекст для сведений о бизнес-деятельности, представленные в форме относящихся к этой

[1] В русскоязычных публикациях можно встретить два варианта перевода термина «system of record»: «система записи» и «система записей». В данном издании использован первый вариант, поскольку существует еще термин «system of records», который наиболее точно переводится как «система записей». — *Примеч. науч. ред.*

деятельности общепринятых абстрактных понятий. Они включают описания (определения и идентификаторы) деталей внутренних и внешних объектов, вовлеченных в бизнес-процессы, таких как клиенты, продукты, сотрудники, продавцы и контролируемые области (значения кодов)» (DAMA, 2009).

Многие склонны относить к основным данным данные о структуре транзакций и данные о структуре организации. Определение Дэвида Лошина из этого же ряда. Он описывает объекты основных данных как ключевые бизнес-объекты, используемые в различных приложениях в рамках организации вместе с соответствующими им метаданными, атрибутами, определениями, ролями, связями и таксономиями. Объекты основных данных представляют самые значимые для организации «вещи» — те, которые отслеживаются в рамках транзакций, отражаются в отчетности, оцениваются и анализируются (Loshin, 2008).

Основные данные требуют выявления и/или выработки достоверной версии правды (trusted version of truth) для каждого экземпляра концептуальных сущностей, таких как продукт, место, счет, физическое лицо или организация, и поддержания этой версии в актуальном состоянии. Главная трудность при управлении основными данными связана с разрешением сущностей (entity resolution) (или управлением идентификационными данными — identity management[1]) — процессом определения различий и управления связями между данными различных систем и процессов. Экземпляры объектов, описываемых строками таблицы основных данных, в отдельных системах организации обычно представлены по-разному. В рамках управления основными данными должны быть отработаны механизмы разрешения этих рассогласованностей, иначе не получится однозначно и непротиворечиво идентифицировать одни и те же экземпляры каждой сущности (будь то клиенты, продукты и т. п.) в различных контекстах. Этим процессом необходимо управлять постоянно, чтобы не допустить рассогласования идентификаторов экземпляров сущностей основных данных на протяжении всего времени их использования[2].

Концептуально справочные и основные данные близки по своему назначению: и те и другие нужны для описания контекста транзакций, без которого невозможно создание и использование транзакционных данных (справочные данные при этом еще и определяют контекст для основных данных). Вместе они обеспечивают адекватное понимание данных. Важно иметь в виду, что и справочные, и основные данные — ресурсы совместного использования, управление которыми должно вестись исключительно на корпоративном уровне, а не на уровне отдельных систем.

---

[1] Термин «identity management» наиболее часто связывают с идентификацией персонала (физических лиц). Однако в данном случае имеется в виду управление идентификационными данными также и любых других сущностей. — *Примеч. науч. ред.*

[2] Дж. Талбот и И. Чжоу (John Talburt and Yinle Zhou, 2015) описывают двухэтапный процесс разрешения сущностей: 1) определить, относятся ли две записи к одной и той же сущности; 2) если да, произвести слияние и согласование данных из двух записей путем создания единой новой записи основных данных взамен двух рассогласованных. Под управлением идентификационной информацией сущностей (Entity Identity Information Management, *сокр*. EIIM) при этом понимается процесс гарантированного обеспечения ситуации, при которой «любой экземпляр сущности под управлением системы MDM безальтернативно фигурирует под одним и тем же уникальным идентификатором при его передаче из процесса в процесс».

Наличие нескольких экземпляров сущностей одних и тех же справочных данных — явление недопустимое, поскольку неизбежно ведет к рассогласованию этих экземпляров. Рассогласование же влечет неопределенность, что становится источником риска для организации. Отметим, что успешная программа управления справочными данными или основными данными должна охватывать весь спектр функций управления данными (руководство данными, обеспечение качества данных, управление метаданными, интеграцию данных и т. д.).

Справочные данные, однако, выделяются из общего ряда основных данных наличием у них только им присущих характеристик, которые отсутствуют, например, у данных о структуре организации или транзакций. Во-первых, справочные данные менее изменчивы. Во-вторых, они обычно проще по структуре и менее объемны, чем наборы транзакционных или основных данных, то есть таблицы справочных данных содержат меньше столбцов и меньше строк. И, в-третьих, никаких трудностей с разрешением сущностей при управлении справочными данными не возникает.

При управлении справочными данными и основными данными основное внимание фокусируется на разных вещах.

- **Управление основными данными (MDM)** подразумевает контроль значений и идентификаторов, обеспечивающий их согласованность во всех системах и наиболее точное отражение актуальных сведений об основных бизнес-сущностях. Цели MDM включают обеспечение доступности точных текущих значений основных данных и минимизацию риска, связанного с их неоднозначной идентификацией (то есть с появлением в системах идентификаторов, относящихся к нескольким экземплярам одной и той же сущности, или соответствующих двум или более сущностям).
- **Управление справочными данными (RDM)** подразумевает контроль допустимых множеств значений данных и их определений. Цель RDM — обеспечить организации доступ к полному набору точных и актуальных текущих значений всех представляемых справочными данными понятий.

Одна из главных трудностей в управлении справочными данными — правильно определить их владельца, то есть лицо, отвечающее за их определение и ведение. Часть справочных данных может поступать в организацию из внешних источников; другая часть — быть разбросанной по различным подразделениям и не иметь формального владельца; еще какие-то справочные данные могут генерироваться и учитываться в одном подразделении, а полученные значения использоваться в других подразделениях. Поэтому определение ответственных за сбор и обновление данных — важная функция RDM. Безответственность в сфере RDM порождает риск, поскольку разночтения в справочных данных влекут за собой неправильное понимание контекста данных (например, когда два бизнес-подразделения по-разному классифицируют одно и то же понятие).

Поскольку и справочные, и основные данные предоставляют контекст для транзакций, они оформляют и приводят в порядок транзакционные данные, вводимые подразделениями организации при выполнении операций (например, в системах CRM и ERP). Кроме того, они задают рамки анализа транзакционных данных.

1.3.2 Справочные данные

Как уже отмечалось, *справочные данные* — это любые данные, используемые для определения характеристик или классификации других данных, или же для соотнесения данных внутри организации с внешней информацией (Chisholm, 2001). В основном справочные данные состоят из кодов и их описаний, но могут иметь и более сложную структуру, в том числе включать отображения и иерархии. Справочные данные имеются практически в любом хранилище. Классификации и категории могут включать статусы или типы (например, статус заказа: новый, обрабатывается, закрыт, отменен). Внешние данные могут включать информацию о географическом местонахождении или применимых стандартах (например, код страны: DE, US, RU).

Справочные данные могут храниться по-разному в зависимости от их функционального назначения: например, для целей интеграции (в частности, описания мэппинга или проверок качества данных) или для обеспечения функций поиска (в частности, круги синонимов). Кроме того, могут предусматриваться различные варианты пользовательского интерфейса для различных устройств и приложений (например, возможность выбора языка). Стандартные методы хранения справочных данных включают:

- таблицы кодов в реляционных базах данных, связанные с другими таблицами с помощью внешних ключей с целью обеспечения ссылочной целостности;
- системы управления справочными данными, которые позволяют осуществлять ведение бизнес-сущностей, наборов допустимых/недопустимых значений, а также правил отображения терминов с целью поддержки широкого спектра приложений или обеспечения интеграции данных;
- метаданные, описывающие атрибуты объекта и позволяющие определять области допустимых значений при доступе к данным через API или пользовательский интерфейс.

Управление справочными данными влечет за собой необходимость контроля и обновления заданных областей значений, определений и связей как между значениями внутри каждой области, так и между значениями в различных областях. Цель управления справочными данными — обеспечить их согласованность и актуальность в рамках выполнения различных функций, а также их доступность для всех приложений и подразделений организации. Как и любые другие данные, справочным данным требуются метаданные. Важнейшим атрибутом метаданных для справочных данных является их источник. Например, в случае стандартного отраслевого справочника это будет руководящий орган, который его утвердил.

1.3.2.1 СТРУКТУРА СПРАВОЧНЫХ ДАННЫХ

В зависимости от уровня детализации описания и сложности описываемых объектов справочные данные могут быть отструктурированы в виде простого списка, таблицы перекрестных ссылок или таксономии. При выборе для хранения справочных данных обычной базы данных или системы управления справочными данными следует учитывать требования по ведению данных и их использованию.

1.3.2.1.1 Списки

Простейшая форма справочных данных — пары код/описание (табл. 17). Значение кода служит первичным идентификатором и краткой формой отображения значения элемента справочных данных в других контекстах. Описание исчерпывающим образом отражает сущность того, что кроется за кодом. Описание может отображаться вместо кода в экранных представлениях, формах, на страницах, в списках, отчетах и т. п. Обратите внимание, что в приведенном примере в качестве кодов названий стран взяты двухбуквенные коды стандарта ISO 3166-1 (например, GB), а не распространенные сокращения, фигурирующие в большинстве коммуникационных контекстов (например, UK). В целом, при определении требований к справочным данным следует крайне взвешенно и обдуманно подходить к поиску разумного компромисса между соблюдением стандартов и удобочитаемостью.

Таблица 17. Справочные данные в формате простого списка

Код	Описание
US	Соединенные Штаты Америки
GB	Соединенное Королевство (Великобритания)

В зависимости от содержания и сложности справочных данных для полного описания смысла кодов могут требоваться дополнительные атрибуты. Определения дают дополнительную информацию, отсутствующую в описании. В отчетах и раскрывающихся списках определения, как правило, не отображаются по причине их избыточной длины. В основном они фигурируют в таких местах, как функции справки приложений, где разъясняется порядок использования кодов в различных контекстах.

Списки, как и справочные данные в любых иных форматах, должны соответствовать информационным потребностям пользователей, что подразумевает достаточную степень детализации данных. Если список значений рассчитан на классификацию данных пользователями, среди которых могут оказаться люди случайные, небрежные и попросту некомпетентные, слишком детализированные определения с большой вероятностью повлекут проблемы с качеством данных в части их распределения по категориям. С другой стороны, скудный выбор из искусственно ограниченного числа слишком обобщенных категорий также не обеспечит надлежащего качества, поскольку не позволит компетентным работникам классифицировать данные с должным уровнем детализации. В таких ситуациях лучше предусматривать два различных по степени детализации, но обязательно взаимосвязанных списка кодов справочных данных с описаниями и определениями, нежели пытаться вести единый стандартный список для пользователей всех категорий и уровней компетентности. Таблица 18 содержит пример кодов статуса заявок, зарегистрированных службой технической поддержки. Без развернутых пояснений, которые даны в столбце «Определения», разобраться в тонкостях различий между статусами человеку, не знакомому со спецификой системы,

было бы проблематично. Подобная дифференциация особенно важна при классификации, например, рабочих показателей и в целом данных, используемых в бизнес-аналитике.

Таблица 18. Справочные данные в формате расширенного списка

Код	Описание	Определение
1	Новая	Заявка только поступила, ресурс обработки не выбран
2	Передана	Заявка передана на обработку с использованием выбранного ресурса
3	Обрабатывается	Выбранный ресурс приступил к обработке заявки
4	Разрешена	Выбранный ресурс отчитался о разрешении проблемы, указанной в заявке
5	Отмена	Заявка отменена запрашивающей стороной
6	Отложена	Заявка не может быть обработана без получения дополнительных сведений
7	Выполнена	Выполнение заявки подтверждено запрашивающей стороной

1.3.2.1.2 Таблица перекрестных ссылок

Разные приложения могут использовать различные наборы кодов для представления одних и тех же понятий. Наборы при этом могут как отличаться, так и не отличаться по степени детализации. Таблица перекрестных ссылок позволяет переводить значения из одних кодировок в другие. Таблица 19 содержит пример перекрестных ссылок между различными системами обозначений американских штатов (то есть без изменения уровня детализации): USPS — стандартные двухбуквенные почтовые коды штатов для внутренних отправлений; ISO — то же для международных отправлений (добавляется префикс с кодом страны); FIPS — код штата в почтовом индексе.

Таблица 19. Список перекрестных ссылок в табличном формате

USPS	ISO	FIPS	Стандартное сокращение	Имя штата	Полное официальное наименование штата
CA	US-CA	06	Calif.	California	State of California
KY	US-KY	21	Ky.	Kentucky	Commonwealth of Kentucky
WI	US-WI	55	Wis.	Wisconsin	State of Wisconsin

Особый случай — мультиязычные табличные списки справочных данных. Если каким-либо приложениям требуется мультиязычная поддержка, коды остаются представленными в стандартном машиночитаемом формате, а вот названия или описания на различных языках берутся из языковых глоссариев. Таблица 20 содержит пример мультиязычной расшифровки кодов стран из стандарта ISO 3166. Порядок работы с подобными таблицами зависит от числа языков

и кодировок. Впрочем, ситуация упрощается за счет того, что подобные таблицы не требуют нормализации, а в денормализованном представлении связи видны и понятны.

Таблица 20. Мультиязычный список справочных данных

Код страны по ISO 3166-1	Английское название	Местное название в английской транскрипции	Местное название в местной транскрипции	Русское название	…
BY	Belarus	Belarus	Беларусь	Белоруссия	
CN	China	Zhong Guo	中国/中國	Китай/КНР	

1.3.2.1.3 Таксономии

Таксономические структуры справочных данных позволяют фиксировать информацию на различных уровнях детализации. Например, почтовые индексы США сами по себе полезны и могут выделяться в специальную осмысленную категорию и на общенациональном уровне, и при классификации данных на уровне штата, округа или города. Эти отношения могут быть отражены в справочной таблице и использованы для анализа данных с разбивкой по почтовым индексам на различных уровнях детализации.

Таксономии обеспечивают возможность многоуровневой классификации контента и многомерной навигации, — например, в целях бизнес-анализа. Таксономические справочные данные могут храниться в таблицах, связанных с помощью рекурсивных внешних ключей. Средства управления таксономиями позволяют также выстраивать иерархии. Таблица 21 и таблица 22 отражают примеры построения иерархических таксономий. В обоих случаях в иерархии фигурирует код, описание и ссылка на код родительской категории. В первом примере Цветочные культуры (10161600) — общий код, под которым проходят дочерние коды культивируемых цветочных растений (розы, орхидеи и т. п.).

Таблица 21. Фрагмент UNSPSC (Универсальная стандартная классификация продуктов и услуг ООН)[1]

Код	Описание	Код родителя
10161600	Цветочные культуры	10160000
10161601	Розы	10161600
10161602	Молочайные	10161600
10161603	Орхидеи	10161600
10161700	Срезанные цветы	10160000
10161705	Срезанные розы	10161700

[1] http://bit.ly/2sAMU06

Во втором примере к Розничной торговле (440000) в качестве дочерней категории отнесена Розничная торговля продовольственными товарами (445000), а к последней — Специализированные продуктовые магазины и рынки (445200). Уровнем ниже начинается их кодификация по специализациям.

Таблица 22. Фрагмент NAICS (Североамериканская система классификации отраслей)[1]

Код	Описание	Код родителя
440000	Розничная торговля	440000
445000	Розничная торговля продовольственными товарами	440000
445200	Специализированные продуктовые магазины и рынки	445000
445210	Мясные магазины и рынки	445200
445220	Рыбные магазины и рынки	445200
445290	Прочие специализированные продуктовые магазины и рынки	445200
445291	Булочные	445290
445292	Кондитерские	445290

1.3.2.1.4 Онтологии

Некоторые организации включают в справочные данные онтологии, которые обычно используются для управления веб-контентом. Такой подход полезен, когда требуется дополнительно предусмотреть возможность описания характеристик других данных или соотнести данные организации с информацией из внешних источников. Онтологии также можно рассматривать в качестве особой разновидности метаданных. Онтологии и другие таксономии сложной структуры должны управляться с особой тщательностью, но в целом принципы и способы управления ими остаются всё теми же, что и при управлении менее сложными справочными данными. Важно обеспечивать их непротиворечивость и полноту, актуальность и четкость определений. Оптимальные практики ведения онтологий также совпадают с теми, которые описаны в настоящем разделе применительно к управлению справочными данными в целом. Поскольку основным назначением онтологий по-прежнему остается управление контентом, подробное их описание дано в главе 9.

1.3.2.2 СОБСТВЕННЫЕ ИЛИ ВНУТРЕННИЕ СПРАВОЧНЫЕ ДАННЫЕ

Многие организации создают справочные данные для внутренних систем, процессов и приложений самостоятельно. Со временем объемы таких служебных справочных данных естественным образом разрастаются. Поэтому одним из компонентов RDM является управление наборами

[1] http://bit.ly/1mWACqg

справочных данных организации, а в идеале еще и обеспечение их полной согласованности между собой в масштабах организации, поскольку рассогласованность чревата серьезными рисками. Например, если в отдельных бизнес-подразделениях используются разные терминологии и коды для описания статусов клиентских счетов, трудно будет найти в такой организации человека, способного со всей ответственностью однозначно назвать точное число клиентов, обслуживаемых организацией в данный момент. Помогая управлять наборами внутренних справочных данных, распорядители данных должны сбалансированно учитывать как необходимость общей для всех подразделений терминологии и классификации информации, так и потребность различных подразделений в ее гибкой адаптации к специфике различных по содержанию бизнес-процессов.

1.3.2.3 ОТРАСЛЕВЫЕ СПРАВОЧНЫЕ ДАННЫЕ

Отраслевые справочные данные (*Industry Reference Data*) — предельно обобщенный термин, описывающий любые наборы данных, создаваемые и регулярно обновляемые отраслевыми ассоциациями или регулирующими органами с целью обеспечения единообразия стандартов кодификации важных понятий. Только кодификация способна обеспечить взаимопонимание и единую трактовку данных в отрасли, без чего невозможны ни корректный обмен, ни обеспечение совместимости систем по данным. Например, в здравоохранении Международная классификация болезней (International Classification of Diseases, ICD) используется для присвоения общеупотребимых кодов заболеваниям (диагнозам) и методам их лечения (вмешательствам и процедурам), и без ICD в глобальном масштабе не было бы возможности вырабатывать единообразный подход к методам лечения пациентов и оценке их результативности. Если бы каждый врач и каждое медучреждение использовали собственные наборы кодов заболеваний, ведение какой бы то ни было эпидемиологической или клинической статистики было бы невозможным.

Отраслевые справочные данные создаются и ведутся за пределами организаций, которые, однако, обязаны их использовать в своей работе с целью обеспечения ее прозрачности и, в целом, возможности взаимодействия между организациями отрасли. В таких условиях могут понадобиться особые меры в области управления качеством получаемых извне справочных данных (например, контроль достоверности данных), бизнес-расчетов (например, на основе экспортируемых таблиц валютных курсов) или дополнительных данных (например, о рыночной конъюнктуре). Характер и состав таких наборов данных извне, подлежащих контролю качества, могут варьироваться в широчайшем диапазоне в зависимости от отрасли и конкретного набора кодов (см. главу 13).

1.3.2.4 ГЕОГРАФИЧЕСКИЕ ИЛИ ГЕОСТАТИСТИЧЕСКИЕ ДАННЫЕ

Справочные данные в привязке к географическим координатам или объектам называют географическими или геостатистическими. Например, бюро переписи населения ведет учет и публикует данные о плотности населения и демографических показателях на уровне единиц административно-территориального деления, и эта демографическая геостатистика затем используется в планировании и прикладных научных исследованиях. История метеорологических наблюдений

в привязке к географическим координатам может использоваться, например, в планировании ресурсов коммунальных служб или календарных графиков промомероприятий.

1.3.2.5 СПРАВОЧНЫЕ ДАННЫЕ ДЛЯ РАСЧЕТОВ

Многие бизнес-приложения требуют оперативного доступа к текущим справочным данным, требующимся для расчета различных параметров. Например, расчеты показателей внешнеэкономической деятельности немыслимы без доступа к надежно управляемым, достоверным и имеющим метки времени таблицам обменных курсов валют. Справочные данные для расчетов резко выделяются из общего ряда справочных данных крайне высокой частотой обновления. Большинство организаций предпочитают получать их на платной основе из коммерческих источников, гарантирующих полноту и точность подобных данных. Попытки вести таблицы текущих расчетных справочных данных собственными силами чаще всего оборачиваются несвоевременным обновлением.

1.3.2.6 МЕТАДАННЫЕ СТАНДАРТНОГО НАБОРА СПРАВОЧНЫХ ДАННЫХ

Справочные данные, как и любые другие, могут периодически изменяться. Учитывая их особую значимость для любой организации, ведение ключевых метаданных, описывающих наборы справочных данных и позволяющих контролировать их происхождение и актуальность, является обязательным. Таблица 23 содержит пример таких метаданных.

Таблица 23. Ключевые атрибуты метаданных набора справочных данных

Ключевая информация о наборе справочных данных	Описание
Официальное название	Обязательный атрибут в случае набора справочных данных из внешнего источника. Пример: ISO 3166-1:2013 Country codes
Название для внутреннего пользования	Название, под которым набор справочных данных используется в организации. Пример: Коды стран по ISO
Поставщик данных	Сторона, отвечающая за ведение и предоставление набора справочных данных. Поставщик может быть внешним (например, ISO), внутренним (например, специальный отдел) или смешанного типа (данные из внешнего источника, переработанные или дополненные внутри организации)
Источник данных поставщика	Описание фактического источника данных, которые используются поставщиком каждого набора справочных данных. Обычно здесь указывается универсальный идентификатор ресурса (URI), который может находиться как в корпоративной сети, так и за ее пределами
№ версии набора данных от внешнего поставщика	Указывается в тех случаях, когда внешним поставщиком предусмотрено сравнение и обновление версий наборов данных. Позволяет своевременно добавлять новые или удалять устаревшие данные из версии, имеющейся в организации
Дата публикации версии набора данных от внешнего поставщика	Указывается в тех случаях, когда внешним поставщиком предусмотрено сравнение и обновление версий наборов данных

Ключевая информация о наборе справочных данных	Описание
№ версии собственного набора справочных данных	Номер версии текущего набора или последнего обновления набора справочных данных, который ведется внутри организации
Дата последней синхронизации собственного набора справочных данных	Дата последнего обновления собственного набора справочных данных путем их согласования с соответствующими данными из внешних источников
Дата последнего обновления собственного набора справочных данных	Дата последнего изменения собственного набора справочных данных без сверки с внешними источниками

1.3.3 Основные данные

Основные данные описывают ключевые бизнес-сущности (например, сотрудников, клиентов, продукты, финансовые структуры, ресурсы, адреса и т. д. и т. п.), определяющие контекст для бизнес-транзакций и их анализа. Сущность (entity) — это какой-либо объект реального мира (человек, организация, место или предмет). Сущности представлены своими экземплярами (entity instances), которые могут быть описаны в форме строк табличных данных или записей. Основные данные должны давать неоспоримо достоверное и точное представление об описываемых бизнес-сущностях, а это требует отлаженного управления.

Формат и области допустимых значений основных данных обычно определяются через бизнес-правила. Наиболее распространенными для любой организации примерами основных данных являются:

- **стороны**, к числу которых могут относиться физические и юридические лица во всевозможных ипостасях и ролях, например: клиенты, покупатели, граждане, пациенты, продавцы, поставщики, агенты, бизнес-партнеры, конкуренты, сотрудники или студенты;
- **продукты и услуги**, как предлагаемые, так и приобретаемые;
- **финансовые структуры**: контракты, приходные и расходные статьи, центры учета затрат и поступлений и т. п.;
- **места**, определяемые адресами или GPS-координатами.

1.3.3.1 СИСТЕМЫ ЗАПИСИ И ЭТАЛОННЫЕ СПРАВОЧНЫЕ СИСТЕМЫ

При наличии потенциально различных версий «правды» (versions of 'the truth') следует как-то определяться, которую из версий считать истинной, а для этого нужно знать происхождение данных, а также точное назначение и порядок подготовки каждой категории данных к использованию. *Система записи* (*System of Record, SOR*) — официально утвержденная система создания, сбора или регистрации данных и их последующего ведения согласно установленным правилам. Например, система ERP может по совместительству являться и системой записи для учета продаж. *Эталонная справочная система* (*System of Reference*) — это официально признанная система, через которую потребители данных могут получать надежные данные для текущей работы и анализа, даже

если данные не создаются в этой системе. Часто в роли эталонной справочной системы выступают MDM-приложения, хабы для совместного использования данных или хранилища данных.

1.3.3.2 ДОВЕРЕННЫЕ ИСТОЧНИКИ И «ЗОЛОТЫЕ ЗАПИСИ»

Доверенным источником (*Trusted Source*) признается тот, данные из которого представляют «лучшую версию правды» (best version of the truth), за счет автоматизированного применения правил проверки и ручного обслуживания контента распорядителями данных. Достоверный источник в зависимости от контекста может обозначаться различными терминами-эвфемизмами: например, «единое представление» (Single View) или «круговой обзор» (360° View). Любая MDM-система должна поддерживаться в таком состоянии, чтобы на нее можно было полагаться как на доверенный источник. Записи, хранящиеся в доверенном источнике, с наиболее точной информацией об экземплярах сущностей принято называть «*золотыми записями*» (*Golden Records*).

Недавно появившееся в MDM понятие *золотая запись* определяется весьма расплывчато. Справочный раздел портала TechTarget гласит: «Золотая запись — единственная четко определенная версия всей совокупности элементов данных, существующих в организационной экосистеме». Следом поясняется: «Золотую запись иногда называют „единственной версией правды“», где под «правдой» понимаются сведения из эталонного справочника, обратившись к которому пользователи данных могут удостовериться в том, что располагают новейшей корректной версией информации. Золотая запись охватывает все без исключения данные во всех системах записи (SOR), используемых в организации»[1].

Однако две части вышеприведенного определения плохо согласуются между собой, что ставит под вопрос корректность всей концепции золотой записи, поскольку данные в различных системах многопрофильной организации далеко не всегда укладываются в картину «единственной версии правды».

Сколько бы усилий по согласованию между собой *основных данных* из множественных источников ни предпринималось, невозможно гарантировать на 100% и полноту, и точность получаемой «золотой записи» с отображением всех сущностей организации, особенно если исходные данные поступают из разных SOR. Не всё то золото, что блестит, и преподносить потребителям данных под видом «золотых» скрижалей данные, достоверность которых не гарантирована, — прямой путь к подрыву доверия к себе.

Именно поэтому для описания «наилучшей из имеющихся» версии основных данных предпочтительнее термин «доверенный источник». За счет этого фокус внимания естественным образом переключается на обеспечение качества определения данных и управления ими с целью получения наилучшей возможной версии. Кроме того, такая трактовка позволяет потребителям свободно вычленять из «единственной версии» основных данных лишь те компоненты данных, которые их интересуют, не задаваясь вопросом о сохранении целостности и связности полумифической золотой записи при выборочном запросе данных. Особо отметим, что «единственная

[1] http://bit.ly/2rRJI3b

версия» основных данных из достоверного источника не универсальна, а зависит от профиля потребителя данных: финансовый отдел получает один набор основных данных, для статистического учета используется другой набор, для маркетинга используется третий, потребителям предлагается четвертый, и т. д. Единый доверенный источник обеспечивает возможность взгляда на структуру бизнес-сущностей под различными углами и на различных уровнях детализации, которые определяются распорядителями данных.

1.3.3.3 УПРАВЛЕНИЕ ОСНОВНЫМИ ДАННЫМИ

Как уже отмечалось во введении к настоящей главе, управление основными данными (MDM) подразумевает контроль значений основных данных и идентификаторов, обеспечивающих возможность согласованного использования всеми системами самых точных и свежих данных о важнейших бизнес-сущностях. Цели MDM включают обеспечение доступности точных, актуальных значений и минимизацию риска неоднозначности идентификаторов.

Gartner, Inc. предлагает следующее определение понятия управление основными данными: «высокотехнологичная дисциплина на стыке бизнес-управления и ИТ, призванная обеспечить единообразие, точность, ответственность, семантическую согласованность и подотчетность распространяемых или открываемых для совместного доступа официальных основных данных, имеющихся в активе организации. Основные данные — последовательно используемый единообразный набор согласованных между собой идентификаторов и расширенных атрибутов, который описывает сущности организации, включая действующих и потенциальных клиентов, граждан, поставщиков, [офисные и производственные] площади, иерархии и коды счетов»[1].

Определение Gartner особо подчеркивает комплексный характер MDM как дисциплины, которая включает совокупность людей, процессов и технологий. MDM никак не завязано на использование конкретных прикладных решений. К сожалению, в английском языке сокращение MDM (от Master Data Management) очень часто используется применительно к прикладным ИТ-системам или программным продуктам, используемым для управления основными данными[2]. MDM-приложения способны упростить реализацию методов, а в некоторых случаях и повысить их эффективность, но само по себе использование MDM-приложения не служит гарантией соответствия основных данных нуждам организации.

Для грамотной оценки потребностей организации в области MDM нужно исследовать следующие вопросы и получить по возможности исчерпывающие ответы на них.

- Какие роли, организации, места и вещи регулярно фигурируют в описаниях бизнес-процессов?
- Какие данные используются для описания этих лиц, организаций, мест и предметов?
- Как и на каком уровне детализации даются определения этих данных?

[1] http://gtnr.it/2rQOT33

[2] Тут следует особо отметить, что в DAMA-DMBOK термин MDM (MDM) используется прежде всего для обозначения процесса управления основными данными как такового, а не программных средств, используемых для его реализации.

- Где создаются или откуда берутся данные? Где они хранятся? Как организованы публикация и контроль доступа к данным?
- Как изменяются данные при перемещении из системы в систему внутри организации?
- Кто использует данные? В каких целях или по какому назначению?
- Каковы критерии оценки качества и надежности данных и их источников?

Организация управления основными данными — задача не из легких. Фундаментальные сложности, присущие всякой осмысленной работе с данными, становятся очевидными: во-первых, люди имеют различные представления об одних и тех же понятиях, и выработать консенсус бывает архисложно; а во-вторых, информация имеет свойство с течением времени эволюционировать, и для систематического учета этих временны́х изменений требуются планирование, доскональное знание природы и структуры данных, а также незаурядные технические навыки. Таким образом, MDM — занятие крайне трудоемкое.

Любая организация, признавшая необходимость MDM, вероятно, уже успела столкнуться с массой сложностей, обусловленных наличием в ИТ-среде множества разнородных систем, которые получают вводные по различным каналам и сохраняют ссылки на сущности реального мира в различных форматах и в различных местах. По причине естественного роста накапливаемых со временем объемов разнородной информации, а также возможных слияний и поглощений системы и процессы, обеспечивающие MDM исходными данными, могут содержать различные определения одних и тех же сущностей, а также использовать различные критерии и стандарты качества данных. Из-за всех этих сложностей лучше подходить к внедрению единой системы MDM поэтапно, вводя ее поочередно в различных предметных областях. Начинать лучше с малого, взявшись за относительно простую область с небольшим числом сущностей и атрибутов, а затем продолжать выстраивать систему MDM методом расширения.

Планирование управления основными данными включает несколько базовых этапов. В каждой предметной области нужно:

- выявить потенциальные источники, данные из которых обеспечат создание комплексного всестороннего представления сущностей основных данных;
- разработать правила, обеспечивающие точность сравнения и корректность слияния экземпляров сущности, оказавшихся идентичными;
- определить подход к выявлению некорректно распознанных как идентичные и необоснованно слитых экземпляров, дополненный корректной процедурой восстановления исходных экземпляров сущности;
- определить подход к распространению прошедших тест на достоверность данных во все системы организации.

Вышеописанные процессы не столь просты по исполнению, как может показаться, поскольку не стоит забывать о том, что MDM осуществляется на протяжении всего жизненного цикла

основных данных. Критически важные направления работы по управлению жизненным циклом основных данных включают следующее.

- Определение контекста сущностей основных данных, включая определения атрибутов, связей и условий их применимости, — и это процесс из области высокоуровневого распоряжения данными.
- Выявление множественных экземпляров сущности в различных источниках; проверка корректности их идентификаторов в каждой системе и построение таблиц перекрестных ссылок с целью интеграции данных.
- Согласование и консолидация данных во всех источниках с целью получения ближайшей к истине версии основных записей. Консолидированные записи обеспечивают слитное представление об информации, имеющейся во всех системах, и помогают выявлять и устранять несоответствия в именах атрибутов и рассогласованные значения данных.
- Выявление некорректно идентифицированных как дублирующие друг друга и слитых воедино экземпляров с последующим обеспечением их корректного разделения с присвоением уникальных идентификаторов каждому восстановленному экземпляру.
- Обеспечение доступа приложений к достоверным данным либо посредством прямого считывания, либо через службы данных, либо методом потоковой репликации данных в хранилища транзакционных, статистических или аналитических данных.
- Обеспечение соблюдения допустимых диапазонов значений основных данных в масштабах всей организации, — а это, опять же, процесс из области руководства данными организации.

1.3.3.4 УПРАВЛЕНИЕ ОСНОВНЫМИ ДАННЫМИ: КЛЮЧЕВЫЕ ШАГИ ПРОЦЕССА

Ниже проиллюстрированы ключевые технологические этапы MDM (см. рис. 76). К таковым относятся: управление моделью данных; сбор и накопление данных; проверка, стандартизация и обогащение данных; разрешение конфликтов; регулирование совместного доступа и обслуживание данных.

Рисунок 76. Ключевые шаги процесса MDM

В комплексной среде MDM единая логическая модель данных будет физически реализована на всем множестве платформ. Именно логическая модель служит и руководством по внедрению решения MDM, и базисом для сервисов интеграции данных, и главным ориентиром при конфигурировании приложений под извлечение максимальной пользы из согласованных данных проверенного качества.

1.3.3.4.1 Управление моделью данных

Работа с основными данными ярко подчеркивает важность наличия четких и согласованных определений на уровне единой логической модели данных. Необходимость выработки такой модели так или иначе заставляет организацию преодолевать склонность к «системному косноязычию». Термины и определения, которые были использованы в отдельно взятой исходной системе — источнике данных, часто ограничены по применимости профильным контекстом этой системы и не могут быть логичным образом распространены на обобщенный уровень в масштабах организации. В логической же модели основных данных термины и определения автоматически подразумевают их применимость в масштабе всей организации и должны с равным успехом вписываться в контекст операций всех без исключения бизнес-подразделений организации, а кроме того, не зависеть от терминологии, используемой в низовых системах, служащих источниками значений различных данных.

Атрибуты объектов модели основных данных должны определяться с настолько тщательной детализацией, насколько это необходимо для обеспечения пригодности соответствующих им значений данных для использования в нуждах всех подразделений организации. В исходных системах — источниках данных могут иметься атрибуты с теми же именами, но в другом контексте и с иным смысловым наполнением, чем у одноименных атрибутов модели данных организации. Возможна и обратная ситуация, когда в исходных системах атрибуты и значения данных поименованы иначе, нежели в модели основных данных организации, но полностью совпадают с ними по смыслу в прикладном контексте. Иногда бывает целесообразно устранять излишнюю детализацию, объединяя на уровне организации в один объект логической модели основных данных атрибуты нескольких объектов одной и той же системы-источника. Возможно и усреднение или агрегирование однотипных близкородственных значений данных исходной модели на уровне организации.

1.3.3.4.2 Сбор и накопление данных

Описания экземпляров сущностей, поступающих даже из одной и той же системы-источника, не говоря о различных источниках, могут дублироваться из-за различий в форматах. Таблица 24 иллюстрирует подобный случай на примере появления в системе MDM двух записей об одном и том же физическом лице из-за различных форматов представления фамилии/имени и номера телефона. Подробнее этот пример будет разобран в следующем разделе настоящей главы.

Таблица 24. Исходные данные, полученные системой MDM

ID исходный	Имя	Адрес	Телефон
123	John Smith	123 Main, Dataland, SQ 98765	
234	J. Smith	123 Main, Dataland, DA	2345678900
345	Jane Smith	123 Main, Dataland, DA	234-567-8900

Планирование, оценка и подключение новых источников данных к системному решению по управлению основными данными должны осуществляться в рамках надежного, многократно воспроизводимого процесса. Работы по организации сбора и накопления данных включают:

- получение и рассмотрение заявок на подключение новых источников данных;
- проведение экспресс-анализа и, по мере надобности, углубленного и высокоуровневого сравнительного анализа данных с использованием инструментов очистки и профилирования данных;
- оценку сложности интеграции дополнительных данных и сообщение ее результатов лицам, запрашивающим их интеграцию, с тем чтобы последние могли сами оценить ее целесообразность с точки зрения окупаемости затрат;
- апробирование подключения нового источника данных и тестирование последствий подключения с точки зрения влияния новых данных на правила согласования;
- окончательное определение метрик качества данных из нового источника;
- назначение ответственных за мониторинг и обеспечение качества данных из нового источника;
- производство и приемку работ по интеграции новых данных в единую среду управления данными организации.

1.3.3.4.3 Проверка, стандартизация и обогащение данных

Для обеспечения разрешения сущностей данные требуется сделать как можно более согласованными и однородными по структуре. Это подразумевает, как минимум, устранение разнобоя в форматах и рассогласованных значений. Устранение противоречий необходимо для минимизации риска ошибок из-за рассогласования записей в системе MDM. Подготовка исходных данных к интеграции включает следующие процедуры.

- **Проверка («валидация»)**: выявляются и исправляются очевидным или доказуемым образом ошибочные, некорректные, бессмысленные или неинформативные данные (например, номера счетов в несуществующих банках или фейковые адреса электронной почты).
- **Стандартизация**: обеспечивается соответствие контента стандартным/допустимым согласно *справочным данным* значениям (например, кодов стран), форматам (например, номеров

телефонов) или полям (например, почтовых адресов); по мере возможности данные переформатируются, в противном случае отбраковываются.

- **Обогащение (дополнение)**: добавление атрибутов, способствующих повышению глубины разрешения экземпляров сущности (например, дополнительная кодификация компаний и бизнес-подразделений по D-U-N-S® или индивидуальных клиентов по ID потребителей в базах данных Acxiom или Experian).

Таблица 25 иллюстрирует результаты очистки и стандартизации данных на примере тех же ранее собранных строк вводных данных, которые были представлены в предыдущем примере (см. табл. 24). Почтовые адреса исправлены и, как и номера телефонов, приведены к единому стандартному формату.

Таблица 25. Стандартизированные и обогащенные вводные данные

ID исходный	Имя	Адрес (исправлен)	Телефон (исправлен)
123	John Smith	123 Main, Dataland, SQ 98765	
234	J. Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900
345	Jane Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900

1.3.3.4.4 Разрешение сущностей и управление идентификаторами

Разрешением сущностей называется процесс определения, относятся ли две различные ссылки на объекты реального мира к двум разным объектам или одному и тому же объекту (Talburt, 2011). Разрешение сущностей реализуется через процедуру принятия решения. Модели этой процедуры могут варьироваться в зависимости от выбранного подхода к определению тождественности или различия объектов, на которые указывают две ссылки. В то время как базовая процедура предусматривает попарное сравнение ссылок, процесс разрешения сущностей может быть систематическим образом расширен на множества данных, включающих любое конечное число экземпляров сущностей. Разрешение сущностей — критически важный этап MDM, поскольку без сравнения и слияния дублирующих друг друга записей, описывающих одни и те же объекты, невозможно построить корректный набор основных данных.

Разрешение сущностей включает несколько видов работ (извлечение ссылок, подготовка ссылок, разрешение ссылок, управление идентификацией, анализ отношений), которые в комплексе и позволяют идентифицировать уникальные экземпляры сущности, изучать связи между ними и управлять их хронологическими изменениями. В рамках процесса разрешения ссылок подлежит выявлению всякая пара различных ссылок на один и тот же объект, для чего используется процедура определения эквивалентности. Ссылки, признанные эквивалентными, затем можно разрешить путем связывания их через общее для обеих ссылок значение (глобальный идентификатор), который будет указывать на их эквивалентность (Talburt, 2011).

1.3.3.4.4.1 Сопоставление

Сопоставление (*matching*) или идентификация кандидатов (candidate identification) — процедура выявления двух или более записей, относящихся к одному и тому же экземпляру сущности. Этот процесс призван минимизировать два риска ошибочной идентификации.

- **Ложноположительное заключение о тождественности**: две ссылки, указывающие на реально отличные друг от друга экземпляры сущности, привязываются к одному и тому же идентификатору. В результате получаем идентификатор, соответствующий двум или более экземплярам сущности.
- **Ложноотрицательное заключение о тождественности**: две различные ссылки на один и тот же экземпляр сущности остаются не связанными единым идентификатором. В результате имеем единственный реально существующий экземпляр сущности, которому соответствует более одного идентификатора в рамках модели, требующей однозначной идентификации.

Обе ситуации необходимо исправлять, и делается это посредством обработки записей с помощью процесса *анализа сходства* (*similarity analysis*) или *сопоставления* (*matching*), позволяющего выявить все пары записей с показателем сходства выше порогового уровня. Сходство часто оценивается по средневзвешенному приближенному совпадению значений атрибутов двух экземпляров. При превышении допустимого порогового уровня совпадения две записи считаются относящимися к одному и тому же экземпляру сущности (то есть тождественными). Анализ сходства позволяет выявлять записи с пренебрежимо малыми расхождениями по значениям данных и консолидировать их. Два основных подхода, которые могут использоваться как по отдельности, так и в комплексе, — детерминированный и вероятностный.

- **Детерминированные алгоритмы**, такие как синтаксический анализ и стандартизация, полагаются на строго определенные структурные схемы и правила присвоения весов и баллов показателям степени сходства сравниваемых записей по различным атрибутам. Такой детерминизм хорош предсказуемостью результатов и единообразием применяемых схем: одни и те же правила будут неизменно приводить к одним и тем же результатам в идентичных ситуациях. Алгоритмы такого рода не требуют тонких настроек и отличаются весьма высокой производительностью и неплохой эффективностью, но корректность их работы всецело зависит от компетентности и предусмотрительности разработчиков правил.
- **Вероятностные алгоритмы** полагаются на статистическую оценку вероятности описания парами записей одного и того же экземпляра сущности. Корректность их работы зависит от качества выборок данных, использовавшихся для обучения алгоритмов распознаванию статистической нормы и отклонений от нее, определения ожидаемых результатов для различных подмножеств записей и тонкой настройки анализатора на самообучение и автоматическую корректировку критериев в процессе статистического анализа. Такие обнаружители совпадений от правил не зависят и результаты выдают недетерминированные. Поначалу это чревато

массой ложных срабатываний, но со временем, по мере накопления статистики и уточнения распределений вероятностей и критериев статистической значимости, вероятностные анализаторы начинают справляться с поставленными задачами всё безукоризненней.

1.3.3.4.4.2 Разрешение неоднозначности идентификации

Некоторые совпадения выявляются с высочайшей степенью достоверности по полному совпадению данных по множеству полей. Другие совпадения носят лишь предположительный характер, поскольку совпадение данных в части полей вступает в противоречие с явными расхождениями в других полях. Примеры возможных ситуаций:

- В двух записях фигурируют одни и те же фамилия, имя, дата рождения и SSN[1], но указаны разные почтовые адреса и телефоны. Насколько безопасно предположить, что речь идет об одном и том же физическом лице?
- В двух записях совпадают данные в полях SSN, почтового адреса и имени, но указаны разные фамилии. Насколько безопасно предположить, что речь идет об одном и том же физическом лице, сменившем фамилию? Зависит ли вероятность смены фамилии от пола и возраста?
- Как изменится ситуация в двух вышеописанных простых примерах, если в одной из двух сопоставляемых записей не указан SSN? Какими еще идентификационными данными можно воспользоваться для уточнения вероятности совпадения? Какую вероятность ошибки считать допустимой при признании двух записей относящимися к одному лицу с целью их слияния в одну?

Таблица 26 иллюстрирует завершающую стадию процесса выявления дублирующих друг друга записей с целью разрешения конфликта, две первые стадии которого представлены выше (табл. 24 и табл. 25). В данном случае второй и третий экземпляры объекта (ID исходный: 234 и 345) на этапе стандартизации определены как представляющие одно и то же физическое лицо (Jane Smith), а первая запись (ID исходный: 123) — как относящиеся к другому физическому лицу (John Smith), проживающему по тому же адресу.

Как видно из примера, при всем желании сопоставление может оказаться ошибочным в силу неоднозначности критериев и неполноты информации. Именно поэтому важно вести историю сравнений и изменений признанных идентичными и объединенных записей, с тем чтобы в случае выявления ошибки можно было отыскать и отменить ошибочные слияния. Статистические показатели числа выявленных совпадений позволяют организациям отслеживать последствия и эффективность применения логических правил сопоставления записей на предмет устранения

[1] SSN (*сокр.* от *англ.* Social Security number) — уникальный девятизначный номер карты социального страхования граждан и резидентов США в формате AAA-GG-SSSS, где изначально AAA означало номер региона, GG — номер группы (технический), SSSS — индивидуальный порядковый (серийный) номер. С 2011 г. географическая привязка ликвидирована, и присваиваемые гражданам и резидентам США SSN генерируются рандомизированным образом. SSN служит официальным удостоверением личности. — *Примеч. пер.*

дублирований. Многократные прогоны данных через процедуры обработки с уточненными правилами помогают выявлять всё новые записи — кандидаты на слияние по мере поступления уточненной вводной информации со стороны процесса разрешения неоднозначностей идентификации экземпляров объектов.

Таблица 26. Выявление ID-кандидатов на соотнесение им разрешаемых записей

ID исх.	Имя	Адрес (исправлен)	Телефон (исправлен)	ID-кандидат	ID лица
123	John Smith	123 Main, Dataland, SQ 98765		XYZ	1
234	J. Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900	XYZ, ABC	2
345	Jane Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900	ABC	2

1.3.3.4.4.3 Потоки работ по сопоставлению данных / Типы согласования

Можно выделить следующие типы согласования данных и соответствующие им потоки работ, зависящие от применяемого набора правил.

- **Правила сравнения на предмет выявления дубликатов** фокусируются на строго определенном множестве элементов данных, уникальным образом описывающих объект, и выявляют потенциально дублирующие друг друга экземпляры. Автоматического слияния при этом не производится. Записи — кандидаты на слияние рассматриваются распорядителями бизнес-данных, которые и принимают экспертное решение о целесообразности слияния в каждом конкретном случае.
- **Правила сравнения и связывания** выявляют записи, которые, судя по всему, являются дубликатами или клонами основных записей, и привязывают записи-клоны к основным записям перекрестными ссылками, не обновляя контента. Правила сравнения со связыванием проще реализовать, а результаты связывания можно безболезненно отменить.
- **Правила сравнения и слияния** выявляют множественные записи, описывающие одно и то же, и объединяют все данные об этом сущностном объекте в единую запись с обязательным согласованием, уточнением и дополнением недостающих значений. В случае применения таких правил к данным из нескольких источников полученная единственная, уникальная и полная запись должна сохраняться в каждом исходном хранилище данных. При таком подходе должен иметься как минимум один доверенный источник, данные из которого будут использоваться для восполнения недостающих и исправления неточных данных в других хранилищах.

Правила сравнения и слияния сложны по определению, и к тому же не существует каких-либо универсальных рецептов, гарантирующих получение искомого результата, а именно — полностью консолидированной и согласованной версии массива информации, распределенного по множеству записей, получаемых из различных источников. Главная сложность проистекает от

необходимости всякий раз определять, какому источнику доверять при согласовании данных в каждом отдельно взятом поле, поскольку это подразумевает последовательное применение серии правил, и далеко не всегда удается логичным образом выбрать хотя бы последовательность их применения. Кроме того, при добавлении любого нового источника все правила приходится перекраивать заново. Дополнительные трудности обусловлены сложностью технической реализации правил сверки и слияния: алгоритмы сравнения и согласования данных являются весьма ресурсоемкими, а затраты на откат ошибочных слияний — неоправданно высокими, поскольку для отмены любого слияния требуется, по сути, восстанавливать все данные во всех хранилищах из резервных копий.

Сравнение и связывание — более простой операционный алгоритм, поскольку изменения вносятся лишь в реестр перекрестных ссылок, а не в поля атрибутов объединяемых через него записей основных данных. Однако в этом случае могут возникать трудности с целостным представлением информации об объекте, данные о котором распределены по множеству записей в различных источниках.

Обязательным требованием является периодическая переоценка правил сравнения и слияния или связывания записей, хотя бы по причине изменения со временем уровней доверительных вероятностей. Многие подсистемы сравнения данных поддерживают расчет статистической корреляции, позволяя точно определять доверительные интервалы (см. главу 13).

1.3.3.4.4.4 Управление идентификаторами основных данных

Одним из компонентов управления основными данными является управление их идентификаторами, к которым в среде MDM относятся *глобальные ID* (*Global IDs*) и *перекрестные ссылки* (*Cross-Reference, X-Ref*).

Глобальные ID присваиваются и ведутся автоматизированной системой MDM и являются уникальными идентификаторами записей, прошедших процедуру согласования. Назначение глобальных ID — обеспечивать уникальную идентификационную маркировку каждого экземпляра каждого объекта основных данных. Возвращаясь к примеру (табл. 26): выявив множественные записи, относящиеся к одному и тому же экземпляру, система MDM определила, что вторая строка, вероятно, является клоном третьей (с промежуточным значением ID-кандидата ‘ABC’), после чего вторая и третья записи были объединены в системе MDM в одну запись под глобальным идентификатором Party ID = ‘2’.

Глобальные ID должны генерироваться единственным авторизованным программным решением, вне зависимости от технологии интеграции основных данных, во избежание всякого риска дублирования экземпляров, записей или значений. Глобальные ID могут иметь различные форматы, например числовой или 128-битный GUID (Global Unique Identifier); главное, чтобы он обеспечивал уникальность идентификации. Основная сложность, с которой нужно каким-то образом справляться при генерировании глобальных ID, — учет изменений ID при слиянии/разбиении записей (с целью надлежащего обновления идентификаторов данных ниже по потоку). *Управление перекрестными ссылками* (*X-Ref Management*) обеспечивает привязку ID данных

в системах-источниках к глобальным ID. Функциональность управления X-Ref должна включать ведение истории карт такого сопоставления, учет метрик степени совпадения и обработки обращений к данным по ссылкам поисковых служб с целью интеграции данных.

1.3.3.4.4.5 Управление принадлежностью

Управление принадлежностью (Affiliation Management) состоит в определении и ведении связей между записями основных данных, отражающими реально существующие связи и отношения. Возможны варианты иерархической принадлежности (пример: компания X является дочерней компанией корпорации Y) или ассоциированной принадлежности (пример: гражданин XYZ работает в компании X).

Проект архитектуры данных MDM-решения должен четко разрешать вопрос о характере принадлежности каждого объекта на уровне логической модели: является ли объект родительским, дочерним или ассоциированным по отношению к каждому из связанных с ним объектов?

- **Отношения ассоциации** обеспечивают максимальную гибкость с точки зрения логики программирования. Ими можно описывать как горизонтально-сетевые, так и вертикально-иерархические связи. Вот только ниже по потокам технологических процессов (например, в системах бухучета или управления учетными записями) крайне желательно иметь иерархические представления информации.
- **Иерархические отношения** просты по логике программирования навигационной структуры, которая выстраивается, по сути, автоматически вслед за структурой логической модели. Однако в случае каких-либо изменений в отношениях подчинения в реальной бизнес-модели отсутствие горизонтальных ассоциаций повлечет необходимость переработки всей архитектуры данных с последующим изменением логической и физической моделей во избежание потери качества данных и целых измерений бизнес-аналитики.

1.3.3.4.5 Совместное использование данных и распоряжение данными

Хотя значительная часть работы по управлению основными данными поддается автоматизации с использованием средств обработки больших массивов данных и записей, следить за их состоянием всё равно необходимо, как минимум для того, чтобы своевременно выявлять и разрешать случаи некорректного сопоставления данных. В идеале уроки, почерпнутые в процессе обслуживания основных данных, можно и нужно использовать для совершенствования алгоритмов и дальнейшей автоматизации процессов сравнения и согласования записей с целью минимизации ручной работы (см. главы 3 и 8).

1.3.3.5 ОСНОВНЫЕ ДАННЫЕ О КОНТРАГЕНТАХ

Основные данные о контрагентах (*Party Master Data*) включают данные о физических и юридических лицах применительно к ролям, которые они играют в бизнесе. В коммерческой среде *контрагенты* включают клиентов или покупателей, сотрудников, продавцов или дистрибьюторов,

партнеров и конкурентов. В государственном секторе в качестве *контрагентов* обычно рассматриваются граждане. В правоохранительных органах и судебной системе роль контрагентов выполняют подозреваемые, обвиняемые, свидетели и жертвы, истцы и ответчики. В некоммерческих организациях — члены и заинтересованные стороны. В здравоохранении — пациенты, медучреждения и медработники, медицинские страховые компании и программы медицинского обслуживания; в образовании — учащиеся/студенты, преподаватели и учебные заведения.

Системы управления взаимоотношениями с клиентами (CRM) предназначены специально для управления основными данными о клиентах. Цель CRM — сбор исчерпывающей и точной информации о каждом клиенте организации.

Обязательным компонентом CRM является выявление в различных системах повторяющихся, избыточных или противоречивых данных с последующим определением, относятся ли они к одному и тому же клиенту или к разным лицам. Система CRM должна уметь разрешать конфликты значений, восстанавливать согласованность данных и обеспечивать точность представления актуальных сведений о каждом клиенте. Для реализации этих процессов требуются надежные правила наряду со знанием структуры, уровней детализации, происхождения и градаций качества источников данных.

Специализированные системы MDM предлагают практически идентичные наборы функций управления данными о физических и юридических лицах, их ролях, сотрудниках и коммерческих партнерах. Вне зависимости от отрасли или профиля бизнеса управление основными данными сопряжено с рядом сложностей, обусловленных следующими факторами:

- сложная структура ролей и сценариев взаимоотношений между физическими и юридическими лицами;
- трудности с однозначной идентификацией сторон;
- множественность, рассогласованность и неоднородность источников данных;
- множественность каналов мобильной связи и веб-коммуникаций;
- высокая значимость и ценность данных;
- необходимость привлекать клиентов расчетливыми информационными ходами.

Особенно сложно управлять основными данными о сторонах, выступающих в разных ролях на разных участках деятельности организации (пример: сотрудники компании, являющиеся одновременно ее постоянными клиентами) или использующих различные и часто нестандартные каналы и методы взаимодействия с организацией (пример: обмен данными через мобильное приложение с привязкой к учетной записи в социальной сети).

1.3.3.6 ФИНАНСОВЫЕ ОСНОВНЫЕ ДАННЫЕ

Финансовые основные данные включают данные о бизнес-подразделениях, центрах учета затрат и прибылей, статьях бухгалтерских проводок, приходно-расходных статьях бюджета, прогнозируемых финансовых показателях и проектах. Обычно финансовые основные данные хранятся

на центральном сервере данных системы планирования ресурсов предприятия (ERP) в виде схемы группировки и кодирования счетов, а учет деталей проектов и транзакций осуществляется периферийными приложениями, то есть по звездообразной схеме управления данными. Особенно часто такая архитектура встречается в организациях с распределенными функциями внутреннего учета.

Решения в области управления финансовыми основными данными могут поддерживать не только создание, ведение и передачу информации, но и моделирование последствий потенциальных изменений существующих финансовых данных в проекции на итоговые показатели. Функции параметрического моделирования финансовых основных данных часто включаются в прикладные модули бизнес-аналитики, отчетности и планирования, а иногда и прямо в приложения по планированию бюджета и прогнозированию финансовых показателей. С помощью подобных приложений можно моделировать и сравнивать различные версии построения финансовых структур, оценивая их эффективность по различным наборам критериев. После выбора оптимального решения соответствующие структурные изменения можно переносить из среды моделирования во все рабочие системы посредством централизованного распространения обновленной версии структуры данных.

1.3.3.7 ЮРИДИЧЕСКИЕ ОСНОВНЫЕ ДАННЫЕ

К *юридическим основным данным* относятся сведения о действующих договорах, контрактах, регламентах и иных документах, имеющих юридическую силу в отношении организации или регулирующих ее деятельность. Юридические основные данные позволяют, например, анализировать контракты с различными поставщиками однотипных продуктов или услуг на предмет выбора оптимального варианта, усиления переговорных позиций или объединения разрозненных контрактов в генеральные соглашения.

1.3.3.8 ОСНОВНЫЕ ДАННЫЕ О ПРОДУКТАХ

Основные данные о продуктах могут описывать продукты и услуги, предлагаемые самой организацией, либо полный спектр продуктов и услуг в отрасли (включая предлагаемые конкурентами). Соответственно, различные типы решений по управлению основными данными о продуктах/услугах могут поддерживать различные наборы функций, требующихся различным бизнес-подразделениям.

- **Управление жизненным циклом продуктов**: **системы PLM**[1] фокусируются на управлении жизненным циклом продуктов/услуг на всех этапах, начиная с разработки концепции, включая стадии разработки/проектирования, производства, продаж/поставок и вплоть до снятия с производства и утилизации неликвидных запасов. Многие организации внедряют PLM-системы с целью ускорения вывода продукции на рынок. В отраслях с затяжными

[1] *сокр.* от *англ.* Product Lifecycle Management. — *Примеч. пер.*

жизненными циклами продуктов (например, в фармацевтической промышленности, где одна только фаза разработки нового лекарства длится в среднем 8–12 лет) системы PLM позволяют организациям отслеживать многофакторные статьи расходов и планировать заключение всех требуемых соглашений по мере развития потенциальных продуктов со стадии концептуальных замыслов до выигрышных торговых наименований, включая их диверсификацию в зависимости от действующих лицензионных соглашений.

- **Управление данными о продуктах**: **системы PDM**[1] поддерживают инженерно-технологические функции, позволяя фиксировать и безопасно распространять спецификации и проектную документацию (например, CAD-чертежи), рецепты, рекомендации и инструкции (например, по сборке, приготовлению и т. п.), типовые регламенты, спецификации материалов, перечни компонентов и т. д. Функционал PDM может быть реализован как через специализированные системы, так и с помощью настроек модулей ERP-приложений.
- **Данные о продуктах в системах планирования ресурсов предприятия (ERP)**: здесь речь идет о складском учете и инвентаризации запасов с детализацией на уровне единиц складского хранения (SKU), что позволяет обрабатывать заказы с точностью до штуки заказываемого изделия соответствующего артикула в соответствии с заданными параметрами; возможны различные технические реализации таких решений.
- **Данные о продуктах в системах управления производством**: **системы MES**[2] предназначены для учета запасов сырья, полуфабрикатов и готовой продукции, после чего данные о готовой продукции могут передаваться под управление системы ERP в качестве складских запасов (SKU). Кроме того, данные о фактическом положении дел на производстве служат важными вводными для планирования всей цепи поставок и сбыта.
- **Данные о продуктах в системах управления отношениями с клиентами (CRM)** должны тесно увязываться со стратегией и тактикой маркетинга, продаж и продвижения новых брендов, управлением торговыми/агентскими сетями, территориальным делением, планированием маркетинговых кампаний и т. п.

Многие системы управления основными данными о продуктах тесно связаны с системами управления справочными данными.

1.3.3.9 ОСНОВНЫЕ ДАННЫЕ О МЕСТОНАХОЖДЕНИИ

Основные данные о местонахождении позволяют отслеживать и передавать географическую информацию и выстраивать иерархические связи по принципу административно-территориального подчинения. Граница между справочными и основными данными в этом случае весьма расплывчата. В целом, рекомендуется придерживаться следующего принципа их разграничения.

[1] *сокр.* от *англ.* Product Data Management. — *Примеч. пер.*

[2] *сокр.* от *англ.* Manufacturing Execution Systems. — *Примеч. пер.*

- **Справочные данные о местонахождении** обычно определяются согласно политическим картам и картам административно-территориального деления и включают данные о стране, регионе, городе или районе, населенном пункте, почтовом индексе и адресе, а также могут быть дополнены геолокационными данными (широта, долгота, высота над уровнем моря). Эти данные меняются крайне редко (в случае переезда/передислокации или переименования), и учет таких изменений обычно ведется сторонними организациями. *Справочные данные о местонахождении* могут также классифицироваться по регионам и территориям обслуживания, продаж и т. п., определяемых самой организацией для внутренних нужд.
- **Основные данные о местонахождении** включают точный юридический адрес и точный фактический адрес головного офиса, а также, при необходимости, фактические адреса производственных подразделений, филиалов и т. п. в случае юридических лиц или адреса регистрации, фактического проживания и/или адрес места работы в случае физических лиц. По мере роста, сокращения или изменения направлений деятельности организации данные об адресах меняются, в целом, значительно чаще, чем справочные данные о местонахождении.

В некоторых отраслях действуют специфические требования по учету специфических геофизических данных (сейсмоопасных зон, затопляемых пойм, почв, среднегодовой нормы осадков, риска экстремальных метеоусловий и т. п.), демографических и социологических данных (численность, плотность и этнический состав населения, уровень доходов, террористические риски и т. п.), которые обычно поступают из внешних источников.

1.3.3.10 ОТРАСЛЕВЫЕ ОСНОВНЫЕ ДАННЫЕ: СПРАВОЧНЫЕ КАТАЛОГИ

Справочные каталоги содержат официальные перечни, по сути, готовых основных данных о юридических и физических лицах, продуктах, услугах и т. п., которые распространяются на коммерческой основе и могут браться организацией за основу при учете транзакций. Исходные справочники создаются различными сторонними организациями, а потому данные, позаимствованные из них, подлежат обязательному согласованию и между собой, и с данными, имеющимися в собственных системах организации. Лишь после этого согласованные версии могут включаться во внутренние информационные системы.

Примерами лицензионных справочных каталогов служат всемирная база данных субъектов бизнеса, которую ведет в облачном хранилище корпорация Dun & Bradstreet, и основные файлы сертифицированных врачей Американской медицинской ассоциации (AMA Physician Masterfile).

Справочные каталоги существенно упрощают управление использованием основных данных. Вот лишь два примера их возможного использования.

- **В качестве отправной точки при сравнении и связывании новых записей**: например, в среде с пятью источниками данных данные из каждого источника можно сверять только с данными в справочном каталоге (5 точек сравнения), а не попарно друг с другом (10 точек сравнения).

- **Для добавления элементов данных, отсутствовавших на момент создания записи**: например, о получении врачом лицензии на право предоставления дополнительного вида услуг или о присвоении компании шестиразрядного отраслевого кода-классификатора NAICS[1].

По мере сверки и приведения записей в соответствие со справочными каталогами должны отслеживаться также и связи проверенных записей с записями в других источниках, и их соответствующие атрибуты также должны приводиться в соответствие с обновленными данными согласно установленным правилам преобразования.

1.3.4 Архитектура совместного использования данных

Существует несколько базовых архитектурных подходов к интеграции справочных и основных данных. Чаще всего основные данные ведутся раздельно по предметным областям, и в каждой предметной области имеется собственная система управления записями. Например, база данных сотрудников обычно ведется системой управления человеческими ресурсами, база данных клиентов — системой CRM, а данные о финансовых ресурсах и продукции учитываются в системе ERP.

Звездообразная архитектурная схема доступа к основным данным (см. рис. 77) предусматривает наличие центральной узловой системы (хаба) управления основными данными, которая и координирует все взаимодействия между источниками данных, бизнес-приложениями и хранилищами данных, что обеспечивает интеграцию данных при минимально возможном числе точек интеграции и/или каналов интерфейса обмена данными. Возможно также использование дополнительного локального узла — концентратора данных для частичной разгрузки центрального узла и масштабирования основных данных (см. главу 8).

Каждый из трех наиболее распространенных базовых походов к реализации операционной среды центра (узла) управления основными данными имеет свои плюсы и минусы.

- **Реестр** записей содержит индексы, указывающие клиентским приложениям и внешним системам путь к запрашиваемым записям, которые ведутся в различных системах управления записями основных данных. Эти системы учета управляют основными данными на локальном уровне приложений, входящих в их состав. Доступ к основным данным осуществляется исключительно через главный индексный указатель реестра. Реализовать такую схему относительно просто, поскольку создание реестра практически не требует внесения изменений в системы учета записей. Но зачастую для того, чтобы собрать основные данные из множества разрозненных систем, требуются очень сложные по семантической структуре запросы. К тому же и бизнес-правила приходится разрабатывать отдельно для каждой системы по той же причине их семантической рассогласованности.

[1] NAICS (*сокр.* от *англ.* North American Industry Classification System) — североамериканский аналог российского ОКВЭД, действующий на территории Канады, Мексики и США. — *Примеч. пер.*

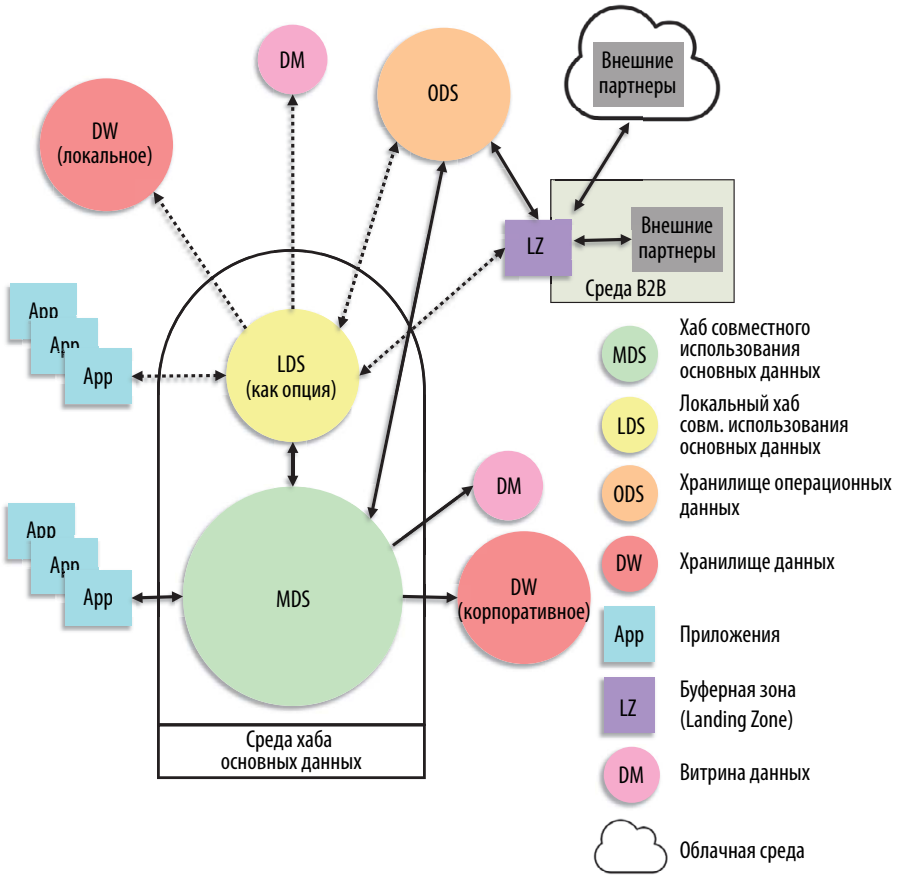


Рисунок 77. Пример архитектуры совместного использования основных данных

- **Транзакционный хаб**: приложения получают доступ к основным данным для считывания и обновления через интерфейсные подключения к центральному узлу. Все основные данные находятся под полным контролем транзакционного центра, и никакие другие приложения не могут сохранять изменения в них самостоятельно. Иными словами, транзакционный хаб исполняет функцию единственной системы управления записями основных данных. Такая система обеспечивает наилучшее высокоуровневое распоряжение основными данными и их полную согласованность. Конечно, лишать существующие системы регистрации транзакционных данных функциональной возможности самостоятельного обновления записей основных данных — очень дорогостоящее удовольствие. Зато все бизнес-правила реализуются в одной-единственной системе — транзакционном центре.
- **Консолидированный подход** правильнее было бы назвать «гибридным», поскольку он предусматривает и реестр, и транзакционный центр. Локальные системы регистрируют основные данные и управляют записями на местах, в локальных хранилищах под управлением входящих в состав этих систем приложений. Все основные данные из локальных хранилищ затем обрабатываются центральным узлом и консолидируются в едином хранилище данных, после

чего совместный доступ к ним обеспечивается опять же через центральный узел, который играет роль справочной системы (см. рис. 77). Это устраняет всякую необходимость доступа к записям, хранящимся непосредственно в системах учета. Консолидированный подход позволяет получать картину данных в масштабах организации без дополнительной нагрузки на системы ведения записей. Все вышеперечисленные плюсы, однако, могут быть перечеркнуты двумя серьезными минусами — дублированием (репликацией) всех данных в центральном хранилище и слишком высоким временем задержки обмена операционными данными между системами на серверах центрального узла.

2. ПРОВОДИМЫЕ РАБОТЫ

Как уже подчеркивалось в разделе 1.3.1, основные данные и справочные данные имеют достаточно много общего по некоторым характеристикам. В частности, и те и другие являются ресурсами совместного доступа, описывающими контекст и смысловое значение других данных, и управление ими должно вестись на уровне организации. Но имеются и важные различия между ними. Наборы справочных данных значительно компактнее и стабильнее массивов основных данных и не требуют сравнения, слияния или связывания и т. п. Поэтому в настоящем разделе будут отдельно описаны сначала направления работ по управлению основными данными (MDM), а затем — по управлению справочными данными.

2.1 Работы по управлению основными данными

2.1.1 Определение драйверов и требований к MDM

В каждой организации существует собственный уникальный набор стимулов для развития систем MDM и препятствий на пути к этому. Факторы влияния включают число, тип и поколение используемых ИТ-систем, поддерживаемые ими бизнес-процессы, назначение и порядок использования самих основных данных в транзакционных и аналитических процессах. Драйверы развития систем MDM часто включают поиск возможностей для совершенствования обслуживания клиентов и повышения эффективности/производительности, а также снижения рисков в области информационной безопасности, защиты конфиденциальных и персональных данных, соблюдения прочих внешних требований. Препятствиями могут являться различия в трактовке и структуре данных в зависимости от системы. Устранению препятствий часто мешают культурные барьеры: например, бизнес-подразделения частенько не желают нести дополнительные затраты на согласование привычных каждому из них систем и процессов между собой, невзирая на очевидную итоговую выгоду от такого согласования для всей организации в целом.

В рамках одного приложения основные данные обычно определить несложно. Куда сложнее выработать согласованный набор стандартных определений для всех имеющихся приложений. В большинстве организаций к решению этой задачи подходят поэтапно, определяя *основные*

данные поочередно для каждой предметной области, а в особо сложных случаях и поочередно для каждого сущностного объекта. Выбор приоритетных направлений проработки *основных данных* рекомендуется подкреплять анализом полезности и окупаемости затрат на предлагаемые усовершенствования с учетом относительной сложности предметных областей, для которых требуется выработать единые наборы *основных данных*. Начинать лучше с простейших категорий и переходить к более сложным, уже наработав определенный опыт.

2.1.2 Анализ и оценка источников данных

За основу при проработке структуры основных данных в системе MDM так или иначе берутся существующие данные приложений. Важно понять структуру и содержание данных каждого приложения и процессы их сбора или создания. Одним из полезных побочных продуктов работ по MDM становится совершенствование структуры метаданных в процессе экспертизы качества существующих данных. Первичная же цель такой экспертизы — понять, каким именно образом можно соотнести всю совокупность имеющихся данных с атрибутами основных данных. В процессе такого соотнесения как раз и вырабатываются четкие определения и устанавливается надлежащий уровень детализации этих атрибутов. На каких-то этапах определения и описания атрибутов основных данных неизбежно встают вопросы семантического характера. Ответственным за данные в предметных областях нужно обязательно согласовывать названия и определения, которые будут использоваться на уровне организации, со всеми бизнес-подразделениями, чтобы итоговый набор основных данных был понятен всем, кто будет иметь к нему доступ (см. главы 3 и 13).

Второе направление экспертной оценки источников состоит в изучении вопросов качества данных. Некачественные данные создадут проблемы с реализацией проекта MDM, так что в случае необходимости следует докапываться до первопричин низкого качества данных в первоисточниках. Главное — никогда не принимать качество данных на веру, ибо за такую доверчивость можно дорого поплатиться. Безопаснее и надежнее исходить из того, что любые данные априори некачественны, то есть ненадежны, неточны, недостоверны, неполны или неадекватны. Всегда проводите экспертизу качества данных и приемлемости источника, прежде чем включать их в среду основных данных.

Самые большие трудности, как уже отмечалось, создают рассогласованные источники данных. Особенно сложно разбираться с ситуациями, когда друг другу противоречат формально высококачественные данные из двух альтернативных источников. Причины такой ситуации могут крыться в трудноуловимых структурных различиях или расхождениях в правилах вычисления или представления значений. Но именно инициативы по согласованию основных данных как раз и дают хорошую возможность для выявления и устранения столь неприятных ситуаций посредством определения и внедрения стандартов сбора или создания информации, которые станут едиными для всех приложений и исключат их рассогласованность по данным.

Для некоторых объектов основных данных, таких как клиент, покупатель или продавец, существует возможность приобретать стандартизированные коммерческие данные (например, справочные каталоги) и уже на их основе вести дальнейшую проработку MDM. Некоторые

поставщики предлагают рафинированные по чистоте каталоги, например, предприятий по отраслям или специалистов по профилям (аккредитованных медучреждений или сертифицированных врачей), с которыми можно сверять внутриорганизационные данные в части названий, имен, контактов и адресов (см. раздел 1.3.3.10). Помимо экспертизы качества существующих данных необходимо также удостовериться и в правильном понимании технологий сбора вводных для выработки основных данных, ведь специфика используемых технологий также оказывает влияние на выбор архитектуры среды MDM.

2.1.3 Определение архитектурного подхода

Выбор архитектуры среды MDM зависит от стратегии бизнеса, платформ существующих источников данных, характера и структуры самих данных, в частности от их генеалогии и волатильности, а также от допусков по запаздыванию синхронизации данных. Архитектура должна согласовываться с моделями потребления данных и совместного доступа к данным. Инструментальное оснащение также будет зависеть не только от потребностей бизнеса, но и от выбранных вариантов архитектурных решений. Оно помогает доопределять основные данные независимо от их ведения в режимах эксплуатации и обслуживания или на этапе ввода модифицируемых систем в эксплуатацию.

Нужно принимать во внимание и число систем-источников, интегрируемых в решение по MDM, и специфику платформ, на которых реализованы эти системы, а также размер организации и ее географический охват. В небольших организациях можно вполне эффективно использовать централизованную схему, построенную вокруг единого транзакционного ЦОД, тогда как в организациях, осуществляющих свою деятельность в глобальных масштабах, неизбежно имеется множество локальных систем, интегрировать которые, вероятно, лучше через реестр. В организации, где бизнес-подразделения практически не сообщаются между собой, работают параллельно и имеют собственные системы, вероятно, целесообразно применять консолидированный (гибридный) подход. Участие в согласовании подхода должны принимать и эксперты по предметным областям, и проектировщики архитектуры данных и корпоративной архитектуры.

Архитектура с использованием центрального хаба совместного доступа к основным данным особенно полезна при отсутствии четко проработанной системы учета/ведения записей основных данных. В этом случае данные поступают в центр управления основными данными из множества систем. При поступлении новых или обновленных данных из одной из периферийных систем именно такая звездообразная архитектура обеспечивает надлежащий контроль согласования данных в других системах с новыми данными. Центр управления совместным доступом становится единым источником контента основных данных для всех хранилищ или витрин данных, упрощает и ускоряет формирование выборок и обработку данных с целью их преобразования, исправления и согласования. В хранилищах данных должны вестись журналы изменений данных, передаваемых на центральный хаб, чтобы их всегда можно было отследить и при необходимости отменить, а вот в центральной узловой системе управления совместным доступом при такой архитектуре вполне можно обойтись одним лишь текущим представлением *основных данных*.

2.1.4 Моделирование основных данных

Управление основными данными — это процесс интеграции данных из множественных источников. Для обеспечения согласованности получаемых результатов и возможности оперативного подключения новых источников по мере расширения организации необходимо иметь проработанные модели данных во всех предметных областях. Логическая или каноническая модель может определяться для всех предметных областей в центральном узле совместного доступа, что позволит устанавливать согласованные между собой на уровне организации определения объектов и атрибутов физических моделей данных, используемых в системах различных подразделений (см. главы 5 и 8).

2.1.5 Внедрение процессов распоряжения и ведения основных данных

Технические решения могут отлично справляться с задачами сравнения, слияния и управления идентификаторами записей основных данных. Однако без должного распоряжения этими данными, осуществляемого специалистами, отвечающими за состояние систем MDM, обойтись не получится, поскольку нужно своевременно выявлять и исправлять случаи выпадения отдельных записей из автоматизированного процесса, а также отыскивать и устранять первопричины таких выпадений на уровне самих процессов. На проекты по совершенствованию MDM должны выделяться достаточные для обеспечения качества основных данных ресурсы, и вестись такие работы должны на постоянной основе. Необходимо вести текущий анализ записей, вносить корректировки в системы-источники и вырабатывать рекомендации, служащие вводными для технических специалистов, отвечающих за тонкую настройку и доработку алгоритмов MDM.

2.1.6 Определение руководящих политик в области использования основных данных и обеспечение их соблюдения

Запуск единой корпоративной системы MDM требует массы усилий, которые по-настоящему начинают окупаться (за счет повышения эффективности и качества работы и обслуживания клиентов) лишь после того, как люди и системы приступят к практическому использованию основных данных. Следовательно, в рамках внедрения MDM должна быть создана дорожная карта перевода всех информационных систем на использование значений и идентификаторов основных данных при реализации операционных процессов. Для полной гарантии согласованности значений основных данных на уровне организации установите однонаправленные замкнутые циклы связи между системами.

2.2 Работы по управлению справочными данными

2.2.1 Определение драйверов и требований

Основным драйвером сбора/создания и ведения справочных данных служит необходимость их применения для обеспечения эффективности и качества обработки транзакционных данных. Централизованное управление справочными данными обходится существенно дешевле ведения

собственных наборов справочных данных каждым бизнес-подразделением и устраняет риск рассогласованности систем. К слову, различные наборы справочных данных существенно разнятся и по важности, и по сложности структуры, и по трудоемкости их ведения. Поэтому нужно выделить приоритетные наборы справочных данных и ориентироваться на них при определении требований к системе управления справочными данными. После создания и запуска системы дополнительные новые наборы справочных данных можно будет определять и настраивать в рамках спецпроектов. Существующие же наборы справочных данных должны проходить регулярные циклы регламентного обслуживания согласно опубликованному графику.

2.2.2 Оценка источников данных

В большинстве отраслей имеются стандартные источники справочных данных и уполномоченные организации, которые занимаются их созданием и обновлением. Одни организации распространяют наборы справочных данных бесплатно, другие на коммерческой основе. На отраслевых рынках информационных услуг часто также имеются посредники, красиво переупаковывающие и доукомплектовывающие справочные данные функциональными дополнениями и с выгодой для себя перепродающие их конечным потребителям. В зависимости от числа наборов и типа справочных данных, которые нужны организации, приобретение у коммерческих поставщиков готовых наборов может оказаться более чем удобным решением, особенно если поставщик гарантирует качество и регулярное обновление данных до последних версий.

Большинство организаций, включая те, что используют справочные данные из внешних источников, ведет также и собственные наборы справочных данных, которые создаются внутри организации. Определение внутренних или местных источников справочных данных часто бывает сложнее, нежели в случае стандартных отраслевых справочников. Как и в случае с основными данными, внутренние источники справочных данных должны четко идентифицироваться, а данные из них выверяться в рамках работ по обеспечению качества. Руководство бизнес-подразделений, в чьем непосредственном распоряжении находятся такие данные, должно понимать все плюсы централизованного управления и с готовностью соглашаться на их передачу в общую базу справочных данных организации, а также оказывать всяческое содействие в части их сопровождения.

2.2.3 Определение архитектурного подхода

Перед покупкой готового или началом проектирования собственного программного средства управления справочными данными критически важно правильно оценить и учесть все требования и возможные трудности управления справочными данными, которые необходимы организации. Особого внимания требует оценка данных на предмет их стабильности (большинство справочных данных достаточно статичны, но встречаются и весьма переменчивые), необходимой частоты обновлений и моделей потребления. Определитесь, нужно ли вести и хранить историю изменений определений и/или значений справочных данных. Если планируете пользоваться коммерческими данными от стороннего поставщика, уточните графики их поставки и метод интеграции.

Архитектурный подход должен вырабатываться с учетом того, что какую-то часть справочных данных в любом случае придется обновлять в ручном режиме. Обеспечьте наличие достаточно простого интерфейса прямого обновления, но при этом конфигурируемого, чтобы можно было вносить изменения в базовые правила ввода данных, с целью учета таких особенностей, как, например, иерархические отношения в структуре справочных данных. Средство обновления данных в составе системы RDM должно позволять операторам вносить в справочные данные изменения по мере надобности без обращения за технической поддержкой и включать автоматизированные рабочие процессы — в частности, получения необходимых согласований и отправки уведомлений. Операторы системы RDM должны планировать график обновлений в соответствии с периодичностью публикации новых кодов. Потребители данных должны получать уведомления обо всех изменениях в системе и обновлениях версий справочных данных. В тех случаях, когда от справочных данных зависят какие-либо логические параметры программного обеспечения, потенциальное влияние изменений на работу программ следует изучать и учитывать до внесения изменений в справочные данные.

2.2.4 Моделирование наборов справочных данных

Многие считают, что к справочным данным относятся лишь простые табличные списки кодов с расшифровками или описаниями. Однако многие справочные данные имеют значительно более сложную структуру. Например, публикуемые в США наборы данных о почтовых индексах (ZIP Codes) обычно включают информацию о штате и округе, а также другие геополитические атрибуты. В целях обеспечения возможности использования справочных данных в долгосрочной перспективе и в привязке к метаданным, да и просто для обеспечения точности и согласованности самих справочных данных нелишним и весьма полезным является создание четких логических и физических моделей наборов справочных данных. Модели помогают потребителям данных разбираться в связях внутри набора справочных данных и могут использоваться для определения правил и критериев контроля их качества.

2.2.5 Внедрение процессов распоряжения и ведения справочных данных

Справочные данные требуют ответственного подхода к обеспечению их полноты и актуальности в части значений, а также четкости и понятности в части определений. Для этого назначаются ответственные за ведение различных наборов справочных данных, которым в отдельных случаях может быть поручено и сопровождение, и техническое обслуживание соответствующих систем, а оно в любом случае должно проводиться при их посильном участии. Например, если несколько бизнес-подразделений используют в своей работе один и тот же набор справочных данных в рамках общей для всех концептуальной модели, куратор этого набора может выступать в роли ведущего на рабочих совещаниях, посвященных согласованию понятий и определению общих для всех ценностей.

В процесс администрирования справочных данных полезно включить ведение основных метаданных о каждом наборе справочных данных, например: фамилия ответственного;

организация-источник; частота плановых обновлений; дата следующего обновления; процессы, использующие данные из набора; сохраняются ли резервные копии предыдущих версий; и т. п. (см. раздел 1.3.2.6). Документирование процессов, использующих справочные данные, позволяет оперативно определять адресатов уведомлений об их изменении.

Во многих инструментариях управления справочными данными реализованы рабочие процессы управления рецензированием и утверждением изменений в структуре и значениях справочных данных. Однако для грамотного использования таких программных средств нужно для начала определить ответственных и за все ролевые функции, и за контент каждого набора справочных данных внутри организации.

2.2.6 Определение руководящих политик в области использования справочных данных

Организация получает осязаемую выгоду от централизованно управляемого хранилища справочных данных лишь в случае востребованности этих данных реальными людьми на реальных рабочих местах. Поэтому организационная политика вкупе с правилами на местах должны определять порядок обеспечения качества справочных данных и либо в явном виде предписывать использование справочных данных именно из собственного хранилища, либо на уровне организации обеспечивать автоматическую выдачу данных именно из этого центрального хранилища по любому справочному запросу и/или автоматическое заполнение по перекрестным ссылкам всех применимых полей именно этими данными.

3. ИНСТРУМЕНТЫ И МЕТОДЫ

Управление основными данными требует инструментария, разработанного специально для управления идентификаторами записей. Для реализации MDM могут использоваться средства интеграции данных, программы корректировки данных, операционные хранилища данных, хабы совместного использования данных или специализированные MDM-приложения. Некоторые разработчики программного обеспечения предлагают настраиваемые решения, которые позволяют управлять основными данными, причем одновременно во многих предметных областях. Другие предлагают интеграционные программные продукты в комплексе с услугами по их внедрению у клиентов с целью создания комплексных решений в области MDM в ИТ-среде организаций-клиентов.

Пакетные решения по управлению продуктами, счетами, учетными записями и базами данных, а также пакеты служб проверки качества данных тоже могут послужить отправной точкой для начала внедрения крупномасштабных программ MDM. Включение подобных служб в ИТ-среду позволяет организациям переходить к использованию лучших в своих категориях решений, безболезненно интегрируя их в привычную архитектуру бизнес-процессов оптимальным для их конкретных нужд образом.

4. РЕКОМЕНДАЦИИ ПО ВНЕДРЕНИЮ

Управление и основными, и справочными данными по формальному признаку может быть отнесено к работам по интеграции данных. Следовательно, внедряемые системы MDM и RDM должны обеспечивать соблюдение принципов интеграции и совместимости данных, описанных в главе 8.

Функциональности систем MDM и RDM единым мощным рывком не внедряются и сами собой не складываются. Решения в этой области требуют детального знания как специфики бизнеса, так и информационных технологий. Организации должны рассчитывать на поэтапное внедрение и постепенное наращивание функциональных возможностей решений по MDM и RDM посредством последовательной реализации проектов, предусмотренных дорожной картой внедрения, приоритетность и очередность которых зависит как от нужд бизнеса, так и от общей архитектуры их информационных систем.

Особо отметим, что реализация программ MDM обречена на провал при отсутствии должного высокоуровневого управления. Специалисты по распоряжению данными организации должны в полной мере понимать сложности, присущие MDM и RDM, и трезво оценивать зрелость организации и ее способность справляться с этими трудностями (см. главу 15).

4.1 Строгое следование архитектуре основных данных

Четкое определение и соблюдение эталонной архитектуры — важнейшее требование управления основными данными и совместным доступом к ним на уровне организации. Для этого подходить к интеграции данных следует с учетом организационной структуры бизнеса, числа раздельных систем учета, реализованной модели распоряжения данными, важности обеспечения доступности и допустимых задержек получения данных, числа клиентских систем и приложений, требующих доступа к данным.

4.2 Мониторинг движения данных

Процессы интеграции справочных и основных данных должны быть спроектированы таким образом, чтобы обеспечивать своевременное извлечение и распространение данных в пределах организации. Потоки данных в среде совместного доступа к справочным и основным данным должны отслеживаться и учитываться. Такой мониторинг движения этих данных необходим для того, чтобы:

- наглядно представлять, как именно осуществляется совместный доступ и для чего используются справочные и основные данные на всех участках работы организации;
- выявлять происхождение данных, которыми обмениваются системы и приложения;
- способствовать анализу корневых причин выявляемых проблем;
- демонстрировать эффективность используемых технических приемов интеграции процессов сбора, учета и потребления данных;

- регистрировать время задержки прохождения значений данных от точки поступления в системы-источники до точки потребления;
- определять значимость бизнес-правил и преобразований данных, реализуемых в интеграционных компонентах.

4.3 Управление изменениями справочных данных

Поскольку справочные данные являются информационным ресурсом, находящимся в совместном доступе, изменять их произвольным образом недопустимо. Ключом к успешному управлению справочными данными является готовность организации к отказу от всяческого децентрализованного и самовольного контроля данных в совместном доступе на местах. Для того чтобы этот принцип устойчиво соблюдался, нужно подкрепить его удобными каналами сбора запросов на изменения в справочных данных и отклика на такие запросы. Совет по руководству данными должен гарантировать последовательную реализацию политик и правил внесения изменений в среды справочных и основных данных.

Внесение изменений в справочные данные должно носить полностью управляемый характер. С одной стороны, глобальные по своему характеру изменения могут затрагивать считаные строки данных. Например, после распада СССР на независимые государства сам термин *Soviet Union* и соответствующие ему коды (в частности, SU) были упразднены, а вместо них в справочных таблицах появились коды новых независимых государств (AM, AZ, BY, GE, EE, KG, KZ, LT, LV, MD, RU, TJ, TM, UA, UZ). На другом полюсе детализации, например в здравоохранении, ежегодно обновляются, уточняются, удаляются и вводятся тысячи кодов процедур и диагнозов, а новые версии справочников и вовсе имеют иную, нежели предыдущие, структуру. Например, коды диагнозов Десятого пересмотра Международной классификации болезней (МКБ-10) структурированы принципиально по-иному, чем в МКБ-9. МКБ-10 отличается от МКБ-9 и форматом, и кодами, и значениями, и — самое главное — принципами рубрикации. У кодов МКБ-10 повышена глубина детализации, вследствие чего они стали специфичнее, а потому классификация содержит многократно больше информации: в редакции МКБ-10 2015 года определены 68 000 кодов против 13 000 кодов, имевшихся в последней редакции МКБ-9[1].

Ставшая в 2015 году в США обязательной кодификация всех случаев обращений пациентов за медицинской помощью по МКБ-10 потребовала серьезного планирования. Организациям, занимающимся предоставлением гражданам медицинских услуг, пришлось озаботиться серьезными

[1] В 2018 г. ВОЗ выпустила принципиально новую онлайновую платформенную версию МКБ-11 (ICD-11), построенную по онтологическому принципу, дополненную таблицами перекодирования из МКБ-10 и рядом полезных инструментов, включая формы обратной связи с целью подачи описываемых ниже запросов на изменение *справочных данных*, утверждение которой включено в повестку 72-й ежегодной Всемирной ассамблеи здравоохранения (май 2019 г.). Число кодов в МКБ-11 сократилось примерно до 55 000 за счет устранения присутствовавших в последних редакциях МКБ-10 дублирующих друг друга, а также весьма экзотических (чтобы не сказать анекдотических) диагнозов наподобие W61.02XA Struck by parrot, initial encounter («Травма, нанесенная попугаем, первичное обращение») или V91.07XA Burn due to water-skis on fire, initial encounter («Ожог в результате воспламенения водных лыж, первичное обращение»). — *Примеч. пер.*

системными изменениями, а также радикальной корректировкой отчетности с целью приведения ее в соответствие с новым стандартом.

Типичные изменения в ИТ-среде после публикации новых наборов справочных данных включают:

- приведение строк данных в соответствие с внешними наборами справочных данных;
- структурные изменения в соответствии с внешними наборами справочных данных;
- изменение наборов внутренних справочных данных на уровне строк;
- изменение структуры наборов внутренних справочных данных;
- создание новых наборов справочных данных.

Изменения могут вноситься как в плановом порядке согласно графику, так и по мере возникновения необходимости. Плановые изменения (например, ежемесячные обновления кодов отраслевых стандартов) не требуют особого надзора со стороны высшего руководства, в отличие от особых случаев, когда высокоуровневое распоряжение процессом обновления справочных данных необходимо однозначно. Особенно это касается тех случаев, когда запрашиваются новые наборы справочных данных, поскольку только на высоком уровне можно определить возможные дополнительные применения подобным данным, о которых может быть ничего не известно тем, кто ими заинтересовался изначально.

Запросы на изменения должны обрабатываться согласно строго определенной процедуре (см. рис. 78). По получении запроса все заинтересованные и/или затрагиваемые стороны должны уведомляться о возможных изменениях, чтобы позволить им заранее просчитать и оценить их потенциальные последствия. Если требуется надлежащее согласование и утверждение изменений, все необходимые обсуждения должны быть проведены, а процедуры соблюдены. Все принятые решения об изменениях должны быть доведены до сведения всех затрагиваемых сторон, после чего запрошенные и согласованные изменения вносятся в опубликованные для совместного доступа справочные данные.

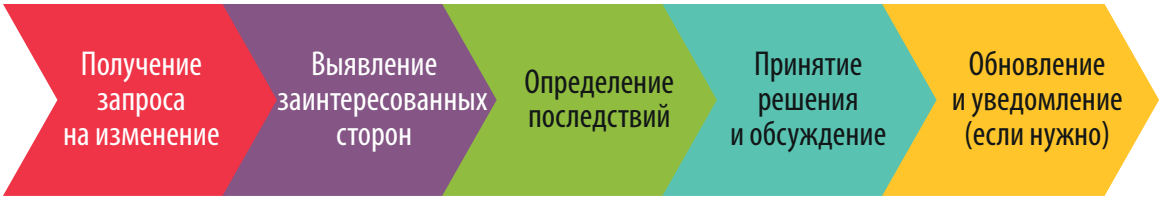


Рисунок 78. Процедура обработки запроса на изменение справочных данных

4.4 Соглашения о совместном использовании данных

Совместный доступ к справочным и основным данным организации с целью их обновления и/или использования требует многостороннего сотрудничества, в том числе, как нередко бывает, и с участием внешних по отношению к организации участников. Для упорядочения доступа

третьих сторон и обеспечения надлежащего использования ими данных, имеющихся в распоряжении вашей организации, желательно иметь продуманные соглашения, четко прописывающие, какие данные могут открываться для совместного доступа и на каких условиях. Наличие подобных соглашений весьма поможет с разрешением возможных проблемных ситуаций, касающихся доступности или качества данных как во внутренней среде, так и распространяемых в порядке совместного доступа. Подобного рода усилия должны обязательно согласовываться в рамках корпоративной программы руководства данными. По мере возможности к выработке решений должны привлекаться все стороны, имеющие отношение к работе с данными: проектировщики архитектуры систем и данных, разработчики моделей, поставщики данных, распорядители данных, разработчики приложений, бизнес-аналитики, контролеры качества, службы информационной безопасности, — иными словами, все, кто имеет малейшее касательство к данным.

Лица, отвечающие за создание среды распространения данных, отвечают и за их качество перед теми, кто поставлен ниже по потоку в силу сложившейся или установленной структуры распределения данных. Но и сами они зависят от качества и своевременности получаемых ими данных из систем, расположенных выше по потоку обработки. Для объективного разрешения проблемных ситуаций в случаях сбоев как раз и предусмотрены соглашения об уровнях обслуживания (SLA) и соответствующие им метрики доступности и качества данных, открытых для совместного доступа. При этом должны быть предусмотрены еще и процедуры выявления корневых проблем с качеством и/или доступностью данных, и стандартный подход к уведомлению затронутых сторон о возникновении проблем и текущем статусе их решения (см. главу 8).

5. ОРГАНИЗАЦИОННЫЕ И КУЛЬТУРНЫЕ ИЗМЕНЕНИЯ

Управление справочными и основными данными приживается в культурно-организационной среде не всегда гладко, поскольку люди испытывают психологический дискомфорт при самой мысли о необходимости отказа от власти над данными и процессами ради совместного создания и использования информационных ресурсов общего доступа. Это порой дается непросто. Специалисты по управлению данными понимают, что хранить чувствительные данные в локальных системах рискованно, но попробуйте объяснить это менеджерам производственных или операционных подразделений, которым данные позарез требуются именно «здесь и сейчас», а усилия по MDM или RDM представляются лишней головной болью и бессмысленной тратой времени, сил и ресурсов, приводящей к тому же к усложнению привычных рабочих процессов.

К счастью, до большинства людей удается донести фундаментальную необходимость этих усилий. Весьма действенны аргументы следующего рода: лучше иметь полное, точное и всестороннее представление о каждом значимом для вас клиенте, чем множество обрывочных, субъективных и однобоких взглядов на него.

Обеспечение доступности и повышение качества справочных и основных данных неизбежно требуют изменения привычных правил и процедур. Состав и масштабы реализуемых решений должны определяться исходя из текущего состояния готовности организации, но обязательно с учетом будущих потребностей, определяемых ее миссией и ви́дением.

Самое же, вероятно, трудно дающееся культурное изменение — переосмысление функций высокоуровневого руководства распоряжением данными и перераспределение ролей. Очень сложно бывает разобраться, кто именно и за что именно отвечает, кто какие решения принимает и кто перед кем отчитывается среди распорядителей бизнес-данных, проектировщиков архитектуры систем и моделей данных, менеджеров и исполнительного руководства. Кроме того, непросто выявить и уровни распределения ответственности между техническими работниками, обслуживающими системы управления данными, координационными комитетами различных программ и Советом по распоряжению данными, и круг вопросов, требующих совместного рассмотрения и коллегиальных решений на основе консенсуса.

6. РУКОВОДСТВО СПРАВОЧНЫМИ И ОСНОВНЫМИ ДАННЫМИ

Будучи ресурсами совместного доступа, справочные и основные данные требуют осуществления деятельности по руководству и распоряжению. Не все несоответствия данных поддаются автоматизированному исправлению. Иногда приходится устранять противоречия в данных путем переговоров между живыми людьми. Без руководства любые технические решения в сфере управления справочными и основными данными так и останутся всего лишь утилитами, используемыми для интеграции данных и не приносящими ощутимой и полноценной отдачи.

В рамках процессов руководства должны приниматься решения по следующему кругу вопросов:

- источники данных для интеграции;
- обязательные правила контроля качества данных;
- условия и правила использования данных;
- виды работ, требующие административного надзора, и периодичность проверок;
- приоритетность и уровни принятия мер в сфере административного надзора;
- представление информации, нужной заинтересованным сторонам;
- стандартные схемы и сроки согласования и внедрения систем RDM и MDM.

Процессы руководства также служат удобным случаем для совместного обсуждения внутренними и внешними надзорными инстанциями, ключевыми интересантами и потребителями информации вопросов минимизации организационных рисков путем определения и утверждения политик и правил информационной безопасности, защиты конфиденциальных и личных данных и соблюдения правил и сроков хранения регламентированной информации.

Будучи процессом непрерывным, руководство данными должно быть выстроено таким образом, чтобы имелась возможность для своевременного рассмотрения и учета новых требований и изменений в существующих правилах с последующим доведением новых принципов, правил и инструкций до сведения пользователей справочных и основных данных.

6.1 Метрики

Существует ряд объективных измеримых показателей для оценки качества справочных и основных данных, а также процессов управления этими данными.

- **Качество и соответствие данных.** Панели мониторинга качества данных могут и должны включать показатели качества справочных и основных данных. В частности, должны отображаться оценки (%) достоверности и соответствия нуждам организации данных по предметным областям, объектам или атрибутам.
- **Контроль изменений.** Аудит происхождения данных, относимых к категории достоверных, — обязательное требование контроля качества данных в среде совместного доступа. Метрики должны обязательно включать частоту изменения значений данных, поскольку она позволяет судить о внутренних характеристиках систем — источников данных и подстраивать под них алгоритмы процессов MDM в среде совместного доступа.
- **Показатели трафика данных.** Данные поступают в системы MDM/RDM из систем-источников, а потребляются системами и процессами, расположенными ниже по информационному потоку. Показатели входящего и исходящего трафика по каналам обмена данными позволяют отслеживать системы, вносящие наибольший вклад в сбор данных, и бизнес-процессы, наиболее нуждающиеся в данных, получаемых по подписке из среды совместного доступа.
- **Соглашения об уровнях обслуживания (SLA)** должны заключаться со всеми сторонами, вносящими вклад в данные совместного доступа и/или подписывающимися на их получение. SLA позволяют обеспечивать гарантированный уровень потребления и принятия данных из среды совместного доступа. Показатели соблюдения условий SLA позволяют проводить углубленный анализ и вспомогательных процессов, и технических проблем, и проблем со структурой данных, замедляющих доступ приложений к системе MDM.
- **Полнота и своевременность обновления данных ответственными лицами.** У каждой категории данных должен иметься распорядитель (должностное лицо или отдел), отвечающий за полноту содержания и своевременное обновление данных. Эти показатели и должны отслеживаться наряду с частотой оценки полноты данных. Иногда эти показатели помогают выявлять и далеко не самые очевидные недостатки в организации работы служб технического обеспечения.
- **Себестоимость данных.** Затраты на получение, хранение и сопровождение данных могут складываться из самых различных статей расходов и зависеть от множества факторов, совокупность которых и должна отражаться подобными показателями. Если смотреть с точки зрения ИТ-решения, затраты могут включать расходы на инфраструктуру аппаратной среды,

приобретение лицензионного ПО, зарплату персонала, гонорары консультантов, расходы на подготовку кадров и т. п. Эффективность и полезность показателей затратности сильно зависят от их последовательного и единообразного учета в масштабах всей организации.

- **Объемы и востребованность совместного доступа.** Потребление данных из внешних источников системами MDM/RDM и спрос на справочные и основные данные отслеживаются по входящему/исходящему трафику и позволяют судить об эффективности работы среды совместного доступа. Эти метрики позволяют регистрировать объем и скорость определения, интеграции и потребления данных подписчиками.

7. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Abbas, June. *Structures for Organizing Knowledge: Exploring Taxonomies, Ontologies, and Other Schema*. Neal-Schuman Publishers, 2010. Print.

Abernethy, Kenneth and J. Thomas Allen. *Exploring the Digital Domain: An Introduction to Computers and Information Fluency*. 2nd ed., 2004. Print.

Allen Mark and Dalton Cervo. *Multi-Domain Master Data Management: Advanced MDM and Data Governance in Practice*. Morgan Kaufmann, 2015. Print.

Bean, James. *XML for Data Architects: Designing for Reuse and Integration*. Morgan Kaufmann, 2003. Print. The Morgan Kaufmann Series in Data Management Systems.

Berson, Alex and Larry Dubov. *Master Data Management and Customer Data Integration for a Global Enterprise*. McGraw-Hill, 2007. Print.

Brackett, Michael. *Data Sharing Using a Common Data Architecture*. Wiley, 1994. Print. Wiley Professional Computing.

Cassell, Kay Ann and Uma Hiremath. *Reference and Information Services: An Introduction*. 3d ed. ALA Neal-Schuman, 2012. Print.

Cervo, Dalton and Mark Allen. *Master Data Management in Practice: Achieving True Customer MDM*. Wiley, 2011. Print.

Chisholm, Malcolm. «What is Master Data?» BeyeNetwork, February 6, 2008, http://bit.ly/2spTYOA Web.

Chisholm, Malcolm. *Managing Reference Data in Enterprise Databases: Binding Corporate Data to the Wider World*. Morgan Kaufmann, 2000. Print. The Morgan Kaufmann Series in Data Management Systems.

Dreibelbis, Allen, et al. *Enterprise Master Data Management: An SOA Approach to Managing Core Information*. IBM Press, 2008. Print.

Dyche, Jill and Evan Levy. *Customer Data Integration: Reaching a Single Version of the Truth*. John Wiley and Sons, 2006. Print.

Effingham, Nikk. *An Introduction to Ontology*. Polity, 2013. Print.

Finkelstein, Clive. *Enterprise Architecture for Integration: Rapid Delivery Methods and Techniques*. Artech House Print on Demand, 2006. Print. Artech House Mobile Communications Library.

Forte, Eric J., et al. *Fundamentals of Government Information: Mining, Finding, Evaluating, and Using Government Resources*. Neal-Schuman Publishers, 2011. Print.

Hadzic, Fedja, Henry Tan, Tharam S. Dillon. *Mining of Data with Complex Structures*. Springer, 2013. Print. Studies in Computational Intelligence.

Lambe, Patrick. *Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness*. Chandos Publishing, 2007. Print. Chandos Knowledge Management.

Loshin, David. *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann, 2001. Print. The Morgan Kaufmann Series in Data Management Systems.

Loshin, David. *Master Data Management*. Morgan Kaufmann, 2008. Print. The MK/OMG Press.

Menzies, Tim, et al. *Sharing Data and Models in Software Engineering*. Morgan Kaufmann, 2014. Print.

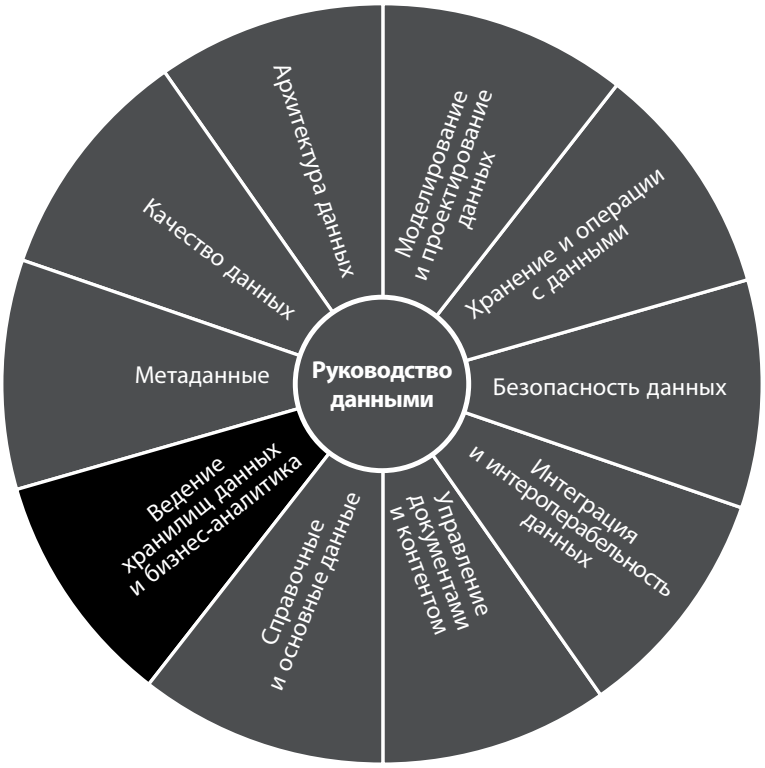
Millett, Scott and Nick Tune. *Patterns, Principles, and Practices of Domain-Driven Design*. Wrox, 2015. Print.

Stewart, Darin L. *Building Enterprise Taxonomies*. Mokita Press, 2011. Print.

Talburt, John and Yinle Zhou. *Entity Information Management Lifecycle for Big Data*. Morgan Kauffman, 2015. Print.

Talburt, John. *Entity Resolution and Information Quality*. Morgan Kaufmann, 2011. Print.

Ведение хранилищ данных и бизнес-аналитика



DAMA-DMBOK2 Data Management Framework

1. ВВЕДЕНИЕ

Понятие *хранилище данных* (*Data Warehouse, DW*) появилось в 1980-х годах для обозначения технологии, позволяющей организациям интегрировать данные из множества разнородных источников в рамках единой модели. Интеграция данных считалась самым многообещающим средством поддержки углубленного изучения операционных процессов и раскрытия новых возможностей для использования данных для обоснования решений и поиска резервов роста эффективности и доходности на уровне организации. Не менее важной ролью хранилищ данных считалась

Рисунок 79. Контекстная диаграмма: ведение хранилищ данных и бизнес-аналитика

и структурная оптимизация за счет упразднения избыточных систем поддержки принятия решений, в большинстве своем так или иначе полагавшихся на данные о работе организации из одного и того же централизованного источника. Концепция же единого хранилища данных сулила возможность устранения избыточности и обеспечения согласованности данных, что делало их более пригодными для принятия оптимальных руководящих решений на уровне организации.

Всерьез строительство хранилищ данных развернулось в 1990-е годы. С тех пор, особенно в связи с одновременным развитием бизнес-аналитики (Business Intelligence, BI) как основного драйвера принятия бизнес-решений, корпоративные хранилища данных успели стать обыденной вещью. В большинстве организаций есть хранилища данных, и их ведение рассматривается как ключевой компонент корпоративного управления данными[1]. Хотя концепция хранилища данных и считается устоявшейся, ее развитие не прекращается. По мере появления всё новых форм данных и концепций, таких как озеро данных (data lake), неизбежно продолжится и эволюция представлений о хранилище данных (см. главы 8 и 15).

1.1 Бизнес-драйверы

Главным драйвером развития хранилищ данных является необходимость поддержки операционных функций, выполнения требований нормативно-правового соответствия и обеспечения деятельности в области бизнес-аналитики (хотя и не вся деятельность в части BI связана с данными из хранилищ). От организаций всё чаще требуют доказательных подтверждений соблюдения нормативно-правовых требований, подкрепленных историческими данными. Следовательно, системы управления хранилищами должны уметь обрабатывать и подобные запросы. Но всё-таки именно поддержка BI остается главной причиной ведения хранилищ данных. BI нужна для всестороннего и глубокого понимания устройства и работы организации, ее клиентов и продуктов. Организация, деятельность которой основана на знаниях, полученных посредством грамотного бизнес-анализа, способна к неуклонному повышению эффективности и, как следствие, получению конкурентных преимуществ. По мере нарастания темпов поступления всевозрастающих объемов данных BI всё более переходит от ретроспективной оценки к предиктивной аналитике.

1.2 Цели и принципы

Организации внедряют хранилища данных в следующих целях:

- поддержка деятельности в области BI;
- повышение эффективности бизнес-анализа и принятия решений;
- изыскание инновационных возможностей по результатам углубленного анализа данных.

При планировании и внедрении хранилища данных следует руководствоваться следующими принципами.

[1] http://bit.ly/2sVPIYr

- **Фокусируйтесь на целях бизнеса.** Хранилище данных должно соответствовать приоритетам организации и способствовать решению бизнес-задач.
- **Изначально помните о желаемых конечных результатах.** Приоритеты и интересы бизнеса плюс потребности в данных BI-приложений должны от начала и до конца диктовать выбор содержания и структуры информационного наполнения хранилища данных.
- **Мыслите глобальными категориями при планировании архитектуры, но руководствуйтесь локальными соображениями при построении.** Ви́дение полной картины конечного результата воплощения архитектурного замысла — требование обязательное, но реализация этого замысла ведется итерационно-поступательными движениями — целевыми проектами или «спринтерскими рывками», обеспечивающими быструю окупаемость вложений.
- **Обобщение и оптимизация производятся на завершающих, а не начальных этапах реализации.** Выстраивайте системную архитектуру на основе максимально детализированных данных. Обобщение, сведение, интеграцию и обобщение с целью приведения структуры данных к стандартным требованиям и повышения производительности систем отложите напоследок, поскольку для восстановления утерянных деталей, если они вдруг понадобятся, всю работу придется откатывать до точки дезинтеграции.
- **Ориентируйтесь на прозрачность и самообслуживание.** Чем больше контекста (то есть всевозможных метаданных), тем проще потребителям разобраться в смысле данных и найти им полезное и выгодное применение. Старайтесь информировать заинтересованные стороны о происхождении данных и процессах их интеграции.
- **Выстраивайте метаданные параллельно с хранилищем.** Критическим фактором успеха хранилища данных является способность объяснять смысл и происхождение данных. В частности, нужно всегда иметь готовые ответы на базовые вопросы вроде: «Почему в поле суммы указано X? Как было рассчитано это значение? Откуда взяты исходные цифры?» Структура метаданных должна моделироваться на стадии проработки модели данных, а учет и управление — входить в состав рабочих процессов и текущих операций.
- **Сотрудничайте** со всеми другими направлениями и проектами в области управления данными, прежде всего с ответственными за руководство данными, обеспечение качества данных и ведение метаданных.
- **Не подходите ко всем с единой меркой.** Различным группам потребителей данных требуются различные инструменты и продукты.

1.3 Основные понятия и концепции

1.3.1 Бизнес-аналитика

Понятие *бизнес-аналитика* (*BI*) имеет два смысловых значения. Во-первых, это вид анализа данных, который нацелен на изучение деятельности организации и выявление открытых и скрытых возможностей для развития бизнеса. Результаты такого анализа используются для совершенствования работы организации и достижения успехов в бизнесе. Говоря о том, что данные — ключ

к успеху и залог превосходства над конкурентами, люди, по сути, расписываются в том, что без BI трудно рассчитывать на радужные перспективы: без правильной постановки вопросов в области сбора и анализа данных организация не получит адекватных ответов о характеристиках своих продуктов, услуг и клиентов, а без всестороннего анализа и учета потребительского спроса руководству не удастся находить оптимальные пути к реализации стратегических целей и планов. Во-вторых, под *бизнес-аналитикой* понимается еще и комплекс технологий, используемых для такого анализа данных. Являясь логическим развитием инструментов поддержки принятия решений, инструменты BI предоставляют возможности по формированию и обработке запросов (querying), извлечению информации (data mining), проведению статистического анализа (statistical analysis), формированию отчетности (reporting), сценарному моделированию (scenario modeling), визуализации данных (data visualization), а также созданию и применению информационных панелей (dashboarding). Средства BI сегодня находят применение во всех областях — от бюджетного планирования до расширенной аналитики (advanced analytics).

1.3.2 Хранилище данных

Хранилище данных (*DW*) включает два ключевых компонента — интегрированную базу данных, необходимых для принятия решений, и увязанное с ней программное обеспечение, используемое для сбора, очистки, преобразования и хранения данных из разнообразных внутренних и внешних источников. Кроме того, для поддержки функций ведения учета исторических данных, операционного и бизнес-анализа DW может включать вторичные витрины данных, то есть выборочные копии данных из основного хранилища. В самом широком контексте под DW может пониматься весь комплекс хранилищ, баз и витрин данных, используемых в организации в целях BI.

Корпоративным хранилищем данных (*Enterprise Data Warehouse, EDW*) называют централизованное DW, предназначенное для информационного обеспечения BI-потребностей всей организации. EDW поддерживает корпоративную модель данных, что обеспечивает согласованность данных, используемых для принятия решений в масштабах организации.

1.3.3 Ведение хранилища данных

Ведение хранилища данных (*Data Warehousing*) включает осуществление текущих операций по извлечению, очистке, преобразованию, контролю и загрузке, обеспечивающих поддержку данных в хранилище в надлежащем состоянии. В процессе ведения DW первоочередное внимание уделяется обеспечению целостности и преемственности данных в историческом и бизнес-контекстах за счет применения к операционным данным адекватных бизнес-правил и реляционных связей. Кроме того, к сфере ведения DW относится также и поддержка процессов взаимодействия и согласования DW с репозиториями метаданных.

В традиционном понимании *ведение DW* относится только к структурированным данным: элементы данных хранятся в полях определенного формата, объединенных в файлы или таблицы, структура которых задокументирована в модели данных. Однако с появлением новейших прогрессивных технологий к области BI/DW стали относить и управление полуструктурированными

и неструктурированными данными. Полуструктурированные данные определяются как электронные элементы, организованные в семантические сущности, атрибуты которых не упорядочены и не связаны (форматы XML и EDI к таковым относятся, HTML — нет). К неструктурированным данным относятся данные, не описываемые какой-либо предопределенной моделью данных. Поскольку существует огромное количество форматов неструктурированных данных (e-mail, текстовые файлы, видео, фото, веб-страницы и т. д.), определение подходящей архитектуры и структуры хранилища, которое обеспечивало бы возможность комплексного анализа всего этого разнообразия в рамках единой системы руководства ведением хранилищ данных, — задача до сих пор не решенная.

1.3.4 Подходы к организации хранилища данных

Основные концептуальные дискуссии относительно того, что именно следует понимать под хранилищем данных, развернулись вследствие того, что два признанных лидера в области разработки концептуальных моделей хранилищ данных — Билл Инмон и Ральф Кимбалл[1] — придерживаются принципиально различных подходов. Инмон определяет *хранилище данных* как «предметно-ориентированный, интегрированный, поддерживающий привязку ко времени, неизменяющийся набор данных, предназначенный для поддержки принятия решений». Такое хранилище строится на основе нормализованной реляционной модели данных. Кимбалл же определяет *хранилище данных* как «копию совокупности транзакционных данных, специфическим образом структурированных для обработки запросов и анализа». Подход Кимбалла подразумевает использование многомерной модели данных (см. главу 5).

Отметим, что при разных подходах к выбору модели данных и Инмон, и Кимбалл отстаивают ряд общих основополагающих идей, лежащих в основе концепции хранилища данных.

- Данные поступают в DW из внешних систем.
- При сохранении данные упорядочиваются и систематизируются с целью повышения их ценности.
- DW делают данные доступными для использования и пригодными для анализа.
- Организации строят хранилища, поскольку нуждаются в контролируемом доступе авторизованных пользователей к надежным, интегрированным данным.
- Данные из хранилищ предназначены для многоцелевого использования — от поддержки рабочих процессов и операционного управления до аналитики и прогнозирования.

1.3.5 Корпоративная информационная фабрика (архитектура Инмона)

Предложенная Биллом Инмоном архитектура DW известна под названием «корпоративная информационная фабрика» (Corporate Information Factory, CIF). DW, согласно определению Инмона, представляет собой «предметно-ориентированный, интегрированный, поддерживающий привязку

[1] Уильям «Билл» Инмон (*англ.* William «Bill» Inmon, p. 1945), Ральф Кимбалл (*англ.* Ralph Kimball, p. 1944) — американские специалисты по информатике. — *Примеч. пер.*

ко времени, неизменяющийся набор сводных и детализированных исторических данных». Исходя из этого определения, назовем и опишем основные концептуальные компоненты CIF и их отличия от оперативных систем (систем поддержки операционной деятельности организации).

- **Предметная ориентированность.** Данные в хранилище организованы по признаку соотнесения их с крупными сущностными объектами бизнеса, а не функциями или приложениями.
- **Интегрированность.** Данные в хранилище унифицированы и связаны. Используются единообразные для всех компонентов хранилища структуры ключей, кодов шифрования, определений данных, условных наименований. Поскольку данные в хранилище интегрированы, они не являются простой копией операционных данных. Вместо этого DW является, по сути, системой записи (system of record) данных.
- **Неизменяемость.** Записи в DW обычно не обновляются, и этим хранилища принципиально отличаются от оперативных систем. Вместо обновления записи с новыми данными добавляются к уже имеющимися. А вот набор записей может отражать хронологию изменений состояния данных в процессе обработки одной и той же транзакции.
- **Привязка ко времени.** Данные в записях DW сохраняются «как они есть» по состоянию на каждый заданный момент регистрации. По сути, записи в DW являются «моментальными снимками» состояния данных об описываемых объектах. Каждый снимок имеет метку времени. Как следствие, сколько бы вы ни запрашивали данные за один и тот же период времени, результаты выдачи будут неизменными вне зависимости от даты и времени обработки запроса.
- **Агрегированные и детализированные данные.** В DW сохраняются как записи о транзакциях на уровне мельчайших деталей, так и обобщенные данные. В операционных системах сводные данные обычно не учитываются. На заре создания DW необходимость обобщения данных диктовалась соображениями экономии вычислительных ресурсов и пространства памяти. В современных средах DW сводные данные могут иметься как на постоянном хранении (в табличной форме), так и формироваться по запросу (в режиме представления). Обычно решающим фактором при принятии решения о необходимости сохранения сводных таблиц является требуемая оперативность доступа к сводным данным.
- **Исторические данные.** Оперативные системы обрабатывают текущие данные, а в DW содержатся записи об истории операций, причем нередко в колоссальных объемах.

Концептуальная модель управления DW в контексте корпоративной информационной фабрики (CIF) описана в книге Инмона, Клаудии Имхофф (Claudia Imhoff) и Райана Соузы (Ryan Sousa) (см. рис. 80).

Архитектура CIF включает следующие компоненты.

- **Приложения**: поддерживают операционные процессы и поставляют детализированные данные в основное хранилище данных и хранилище операционных данных (Operational Data Stores, ODS), где они могут анализироваться и обобщаться.

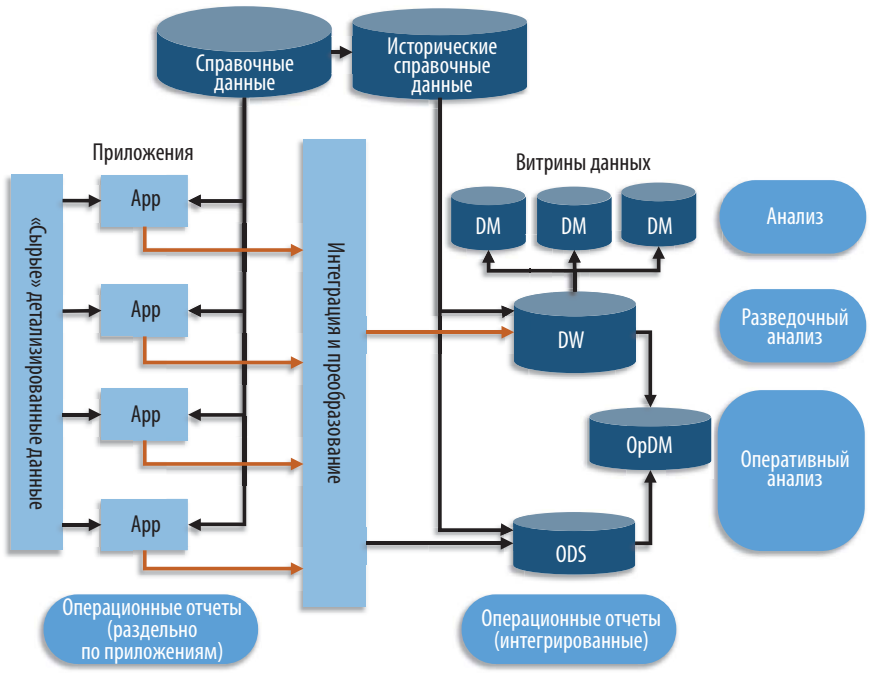


Рисунок 80. Корпоративная информационная фабрика (CIF)

- **Область временного хранения**: база данных, выполняющая роль посредника между базами данных приложений как первоисточниками данных и целевыми базами данных. Именно в области временного хранения (staging area) реализуются операции извлечения, преобразования и загрузки данных в целевые хранилища. Конечные пользователи доступа к данным в этой области не имеют. В основном там присутствуют данные временного хранения, хотя относительно небольшая часть данных из этой области обычно отправляется напрямую на постоянное хранение.
- **Интеграция и преобразование**: в интеграционном слое разрозненные данные из различных источников приводятся к виду, позволяющему включать их в стандартном представлении корпоративной модели данных в DW и ODS.
- **Хранилище операционных данных (ODS)** представляет собой интегрированную базу операционных данных, поступающих от приложений и/или из других баз данных. В ODS обычно содержатся только текущие данные или данные за относительно небольшой отчетный период (30–90 суток), в то время как в главном DW накапливаются еще и исторические данные (часто за многие годы). Главное же отличие ODS от DW заключается в том, что операционные данные динамически изменяются по мере поступления новых данных из среды интеграции в отличие от статичных данных в главном хранилище. ODS используются далеко не во всех организациях, а только в тех, где требуется минимизировать время запаздывания. При его наличии ODS может также служить первоисточником данных для DW или, как альтернатива, контрольным источником для проверки корректности данных в DW.

- **Витрины данных**: в витринах данных (Data Marts, DM) отображаются подготовленные к анализу данные. Обычно это подмножества данных из DW, структурированные специально для нужд узкоспециализированных аналитиков или иных категорий потребителей. Например, в витрины могут выводиться агрегированные данные для быстрого анализа. Многомерные модели часто позволяют использовать приемы формирования быстрых (ненормализованных) выборок данных для пользовательских витрин различного профиля.
- **Витрина операционных данных**: витрина операционных данных (Operational Data Mart, OpDM) стоит особняком, поскольку в ней отображается выборка данных, требующихся для оперативного принятия тактических решений. Данные в OpDM поступают напрямую из ODS, вследствие чего OpDM наследует многие характеристики ODS. В частности, в ней присутствуют только текущие или новейшие данные, которые быстро меняются.
- **Хранилище данных**: главное DW служит единой точкой доступа к полностью интегрированным корпоративным данным, требующимся для принятия управленческих решений, стратегического анализа и планирования. DW получает данные из среды интеграции данных приложений и ODS и отправляет данные в витрины, причем обычно без обратной связи с последними. Данные, не устраивающие пользователей витрин или нуждающиеся в исправлении, просто отклоняются или отбраковываются пользователями, после чего подлежат корректировке в исходном приложении, желательно с последующим повторным пропусканием через всю систему.
- **Операционные отчеты**: технические отчеты генерируются на уровне приложений, а сводные и аналитические являются продуктами витрин данных.
- **Справочные, основные и внешние данные**: помимо транзакционных данных, поступающих из приложений, CIF использует данные, необходимые для понимания транзакций, — например, справочные и основные данные. Доступ к общим для всех компонентов определениям таких данных существенно упрощает интеграцию DW. Хотя приложения используют исключительно текущие значения справочных и основных данных, самому DW требуются и их исторические наборы, и данные о сроках их действия для правильной интерпретации всего массива накопленных данных (см. главу 10).

Рисунок 80 также отражает потоки данных внутри CIF, начиная со сбора или создания данных приложениями (слева) вплоть до выдачи и анализа информации системами витрин данных (справа). При продвижении слева направо данные претерпевают ряд изменений. Например:

- Назначение данных смещается из области оперативного управления в область аналитики.
- В роли конечных пользователей рядовых работников сменяют ответственные руководители.
- Потребление системных ресурсов стандартизованными рабочими процессами сменяется обработкой разовых пользовательских запросов.
- Требовательность ко времени ответа снижается (так как стратегические решения спешки не требуют).
- Для обработки любой операции, запроса или процесса требуется значительно больше данных.

Данные в DW и витринах отличаются от данных приложений следующими особенностями:

- организация по предметным областям, а не по функциональному назначению;
- интеграция вместо узкой специализации;
- рассматриваются переменные значения по времени, а не текущие значения данных;
- допустима значительно большая задержка выдачи данных, чем в приложениях;
- в DW доступно значительно больше исторических данных, чем в приложениях.

1.3.6 Многомерное хранилище данных (архитектура Кимбалла)

Альтернативная ветвь развития архитектурных концепций хранилищ данных основана на различных реализациях многомерной структурной модели данных Кимбалла, который определяет хранилище как «копию транзакционных данных, особым образом структурированную для обработки запросов и анализа» (Kimball, 2002). Тут важно оговориться, что «копия» в данном контексте не означает точной копии оригинала. При переносе в хранилище данные подвергаются реструктуризации сообразно схеме многомерной модели, которая специально проектируется таким образом, чтобы сделать данные предельно понятными и полезными для потребителей, но при этом сохранить и достаточный для обработки запросов уровень формализации[1]. Важнейшим отличием многомерных схем хранения данных от традиционных реляционных является отказ от нормализации.

Многомерные модели, часто называемые также *звездообразными схемами* (*Star Schema*), представляют собой подборки *фактов* (*facts*), под которыми понимаются числовые данные или характеристики бизнес-процессов (например, объем продаж) в проекции на *измерения* (*dimensions*), которые используются для описания атрибутов, соответствующих фактам и позволяющих пользователям правильно интерпретировать фактические данные (например, с объемом продаж сопоставляются артикул продукта X и отчетный квартал). Таблица фактов связана со множественными таблицами измерений, и в графическом представлении такая схема организации данных имеет форму звезды, откуда и обиходное название (см. главу 5). При наличии в модели множественных таблиц фактов они проецируются на общие для различных таблиц так называемые «конформные» (conformed) измерения через «шину» (bus), подобную компьютерной шине[2]. Множественные витрины данных на корпоративном уровне могут интегрироваться посредством подключения их к общей шине конформных измерений.

Матрица шины DW отражает доступные фактические данные на пересечениях строк бизнес-процессов (фактов) и столбцов предметных областей (измерений). Возможности для интеграции через конформные измерения появляются там, где множественные процессы используют одни и те же данные. В нижеприведенном простейшем примере (табл. 27) к бизнес-процессам

[1] http://bit.ly/1udtNC8

[2] Термин «шина» Кимбалл позаимствовал из своей первоначальной области специализации — электротехнической инженерии, где к общей «шине питания» подключаются различные энергопотребляющие компоненты системы.

отнесены Продажи, Запасы и Заказы, и данные обо всех трех бизнес-процессах могут интегрироваться через общие для них конформные измерения Дата и Продукт. Данные о Продажах и Запасах могут интегрироваться через измерение Магазин, а данные о Запасах и Заказах — через измерение Поставщик. Таким образом, лишь четыре измерения из пяти — Дата, Продукт, Магазин и Поставщик — являются кандидатами на роль конформных. А вот измерение Склад общим для каких-либо бизнес-процессов не является и для интеграции данных непригодно, поскольку ему соответствует единственный бизнес-процесс — учет Запасов.

Таблица 27. Пример матрицы шины DW предприятия

	Предметные области				
Бизнес-процессы	**Дата**	**Продукт**	**Магазин**	**Поставщик**	**Склад**
Продажи	X	X	X		
Запасы	X	X	X	X	X
Заказы	X	X		X	
Потенциальная конформность измерения	Да	Да	Да	Да	Нет

Матрица шины корпоративного DW предприятия может использоваться для представления данных в соответствии с долгосрочными информационными потребностями систем DW/BI вне зависимости от технологий их реализации. Столь универсальный инструмент позволяет организации развертывать широкомасштабные работы по управляемому развитию информационных систем. Внедрение каждой новой системы приводит к появлению еще одной пристройки к общей архитектуре предприятия. Рано или поздно количество реализованных многомерных схем перерастает в качество в том смысле, что становится целесообразным их объединение в интегрированную среду корпоративного DW.

Схематическое представление архитектуры DW/BI по Кимболлу (см. рис. 81) часто называют «шахматным» (Kimball's Data Warehouse Chess Pieces — «шахматные фигуры хранилища данных Кимбалла»). Обратите внимание, что в модели Кимбалла функциональная роль самого хранилища данных значительно шире, чем в модели Инмона, и включает все компоненты подготовки/загрузки и представления/выдачи данных.

- **Оперативные системы — источники данных**: оперативные/транзакционные приложения организации. Именно ими создаются исходные данные, интегрируемые в ODS и DW. В принципе, этот компонент эквивалентен области «Приложения» на схеме архитектуры CIF (см. рис. 80).
- **Область временного хранения** по Кимболлу включает комплекс процессов, необходимых для интеграции и преобразования данных в форму, удобную для представления. Эту область можно в грубом приближении уподобить компонентам «Интеграция и преобразование»

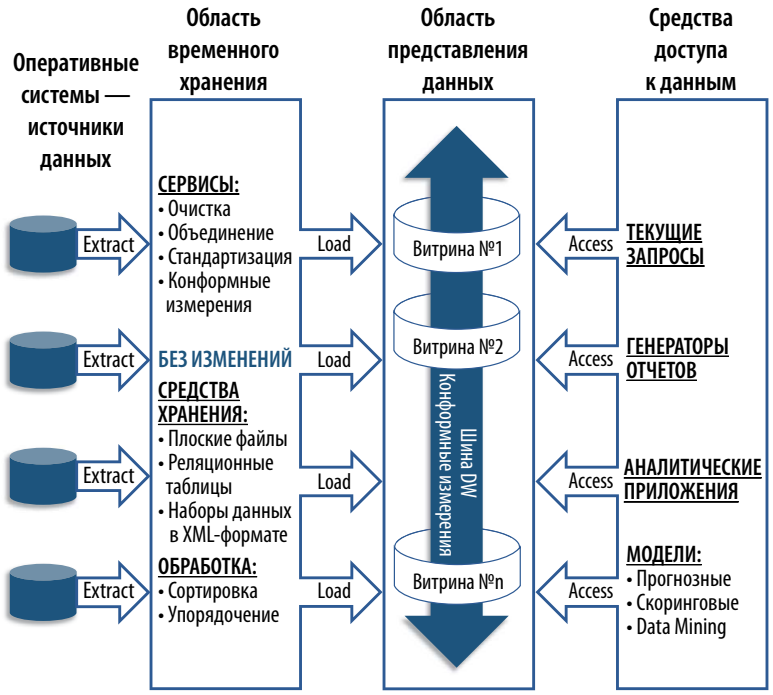


Рисунок 81. Пример «шахматного» представления архитектуры DW по Кимбаллу[1]

и ODS/DW модели Инмона. Однако у Кимбалла основной акцент делается на эффективности подготовки данных для аналитических приложений и конечных пользователей, то есть охватывается более узкий круг задач, чем в архитектурной модели Инмона, ориентированной на полную интеграцию корпоративного управления данными. По сути, вся архитектура DW по Кимбаллу целиком соответствует архитектуре временного хранения по Инмону.

- **Область представления данных**: здесь витрины данных принципиально ничем не отличаются от витрин данных CIF. Ключевым же архитектурным отличием является парадигма интеграции через шину DW, то есть через общие для различных витрин так называемые «конформные» измерения.
- **Средства доступа к данным**: подход Кимбалла изначально ориентирован на потребности конечных пользователей, и средства доступа к данным также определяются строго сообразно их потребностям.

1.3.7 Архитектурные компоненты DW

Среда хранилища данных включает ряд архитектурных компонентов, которые нужно выстроить таким образом, чтобы она наилучшим образом соответствовала потребностям предприятия. Рисунок 82 отражает компоненты архитектуры DW/BI в связке со средой обработки больших данных, обсуждению которой посвящен настоящий раздел. Дело в том, что с развитием концепции

[1] Переработанная версия иллюстрации из: Kimball and Ross (2002). Используется с разрешения правообладателя.

больших данных изменилось и представление об архитектуре DW/BI, поскольку среда обработки больших данных стала дополнительным магистральным каналом притока данных в информационную среду организации.

Рисунок 82 отражает также и фазы жизненного цикла данных. Из систем-источников данные поступают в область временного хранения, где подвергаются очистке и обогащению, интегрируются и отправляются на хранение в DW и/или ODS. Доступ к данным из DW осуществляется через витрины или кубы; также они используются для генерирования разнообразных отчетов. Параллельно аналогичной обработке подвергаются и входящие потоки больших данных, но с одним принципиальным отличием: в большинстве хранилищ данные интегрируются перед загрузкой в таблицы; обработка же больших данных требует их предварительной загрузки до интеграции. BI больших данных может включать функции прогностического анализа и статистического анализа на предмет выявления скрытых закономерностей, а также более традиционные формы аналитической отчетности (см. главу 14).

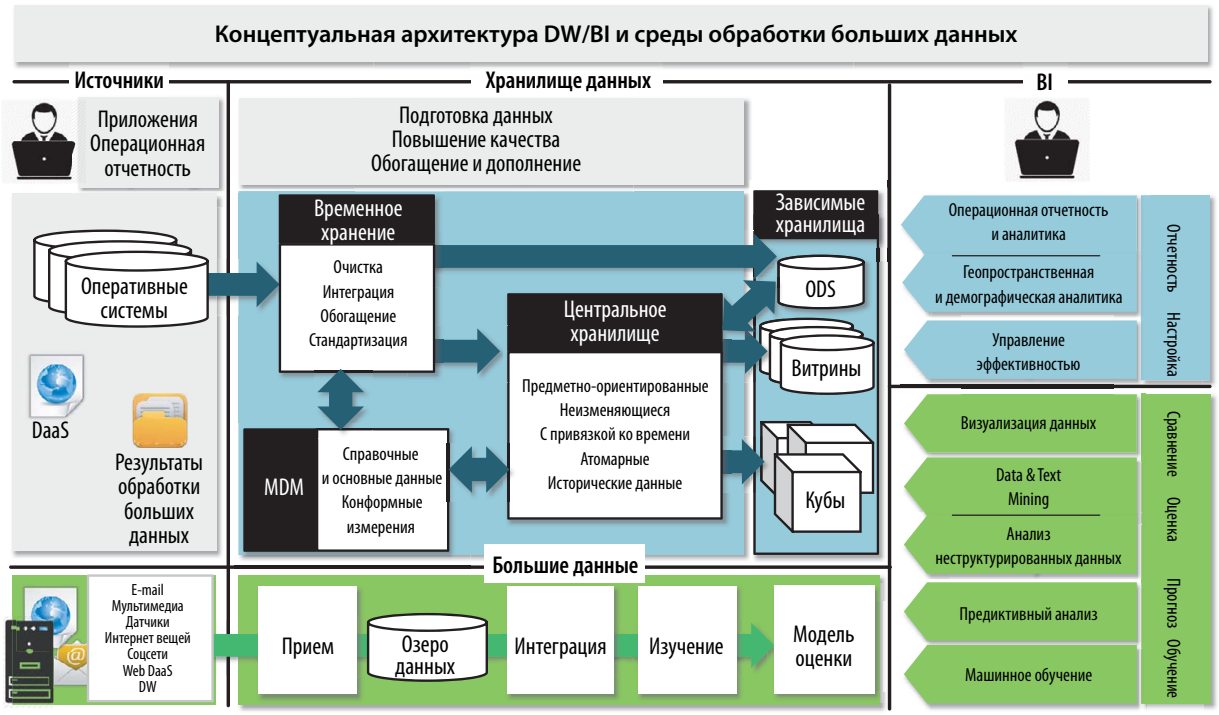


Рисунок 82. Архитектурная концепция DW и BI в условиях доступности больших данных

1.3.7.1 СИСТЕМЫ-ИСТОЧНИКИ

Системы-источники в левой части схемы (рис. 82) включают как внутренние оперативные системы, так и внешние источники данных, привносимых в среду DW/BI. Под оперативными системами в данном контексте понимаются прикладные системы управления взаимоотношениями с клиентами (CRM), бухгалтерским учетом и кадровыми ресурсами, а также прочие системы

оперативного управления, состав которых зависит от отрасли. Внешние данные могут поступать от поставщиков коммерческих баз данных и информационных сервисов (DaaS), включать веб-контент и результаты обсчета различных массивов больших данных.

1.3.7.2 ИНТЕГРАЦИЯ ДАННЫХ

Интеграция данных включает алгоритмы извлечения, преобразования и загрузки (ETL), виртуализацию данных и другие технические средства унификации, формализации и доставки данных по месту назначения. В случае сервис-ориентированной архитектуры (SOA) уровни сервисов данных также относятся к этому компоненту. На иллюстрации (рис. 82) все стрелки соответствуют различным интеграционным процессам (см. главу 8).

1.3.7.3 ОБЛАСТИ ХРАНЕНИЯ ДАННЫХ

Данные в хранилище распределены по различным областям.

- **Область временного хранения**. Промежуточная область хранения данных на пути из первоисточника в централизованное хранилище, где данные доводятся до кондиции. Данные преобразуются, интегрируются и догружаются в хранилище.
- **Конформные измерения справочных и основных данных.** Справочные данные обычно хранятся в отдельном репозитории (на рисунке не показан). По мере поступления подготовленных к интеграции новых основных данных они распределяются по конформным измерениям, которые определяются справочными данными.
- **Центральное хранилище.** По завершении преобразований и подготовки данные обычно отправляются на долгосрочное хранение в центральное хранилище в атомарной форме. В этом же слое содержатся и все предыдущие (исторические) экземпляры массива атомарных данных, и последний экземпляр результатов пакетной обработки этого массива. Структура данных в центральном хранилище разрабатывается в соответствии с потребностями процессов и пользователей, выступающих в роли потребителей данных, однако можно выделить следующие универсальные конструктивные элементы и функционалы, поддерживаемые ими:
  - определение связей между бизнес-ключами и суррогатными ключами для повышения производительности;
  - индексация и определение внешних ключей с целью поддержки множественных измерений;
  - реализация алгоритмов регистрации изменений данных (Change Data Capture, CDC), что позволяет выявлять, учитывать и сохранять историю изменений.
- **Хранилище операционных данных (ODS)** — последняя актуальная версия данных из центрального хранилища для быстрого оперативного доступа за счет минимизации времени отклика. Поскольку в ODS исторические записи отсутствуют, то и время обновления, и время запаздывания выдачи запрашиваемых пользователями или приложениями данных тут существенно ниже. Иногда в ODS в режиме реального времени передаются серии снимков данных,

фиксируемые через заданные интервалы времени, которые и используются для формирования комплексной отчетности и анализа. В последнее время в свете повышения частоты обновлений данных в ODS, ставшей следствием роста требовательности бизнеса, производительности компьютерного оборудования и совершенствования технологий обработки данных в режиме реального времени, всё чаще встречаются архитектурные решения, в которых ODS интегрировано с центральным DW или витринами данных.

- **Витрины данных** — разновидность систем хранения данных, используемых для выдачи или представления особым образом отфильтрованных, отсортированных и упорядоченных данных из DW пользователям или приложениям. Могут также использоваться для представления подмножеств или выборок данных, необходимых различным структурным или функциональным подразделениям в форме интегрированных отчетов; поддерживать обработку запросов и функции анализа исторических данных. Каждая витрина данных ориентирована на определенную предметную область, подразделение или бизнес-процесс. Кроме того, витрины данных могут служить базовыми источниками данных для создания надстройки в виде виртуального хранилища данных, в котором данные из витрин будут переконфигурированы в новые, обобщенные сущностные объекты. По мере интеграции в постоянное хранилище новых данных обновляются и дополняются также и все данные, отображаемые в различных витринах.
- **Кубы** — представления данных, поддерживающие онлайновую аналитическую обработку (OLAP). Могут конструироваться кубы трех типов, названия которых соответствуют типу лежащей в их основе базы данных, — реляционные, многомерные и гибридные.

1.3.8 Типы технологий загрузки

В управлении хранилищами данных реализуются два основных технологических процесса интеграции данных: загрузка исторических данных и текущее обновление. Исторические данные обычно загружаются единожды или максимум несколько раз в случае выявления неких проблем, подлежащих устранению. Текущие же обновления производятся регулярно согласно установленному графику и призваны обеспечивать актуальность данных в хранилище.

1.3.8.1 ИСТОРИЧЕСКИЕ ДАННЫЕ

Одно из преимуществ концепции DW заключается в том, что в центральном хранилище фиксируется во всех деталях ретроспективная история данных. Для регистрации детализированных исторических данных используются различные методы. Организация, желающая вести архив исторических данных, должна выбрать наиболее подходящую методологию и технологию, исходя из предъявляемых к историческим данным требований. Если необходимо обеспечить возможность пошагового воспроизведения всех изменений, требуется совсем иной подход, нежели в тех случаях, когда достаточно иметь хронологические снимки версий текущих состояний.

Архитектура Инмона подразумевает, что в хранилище имеется единственный слой данных, которые предварительно очищены и стандартизированы, а управляются на атомарном уровне.

Реализация интеграции и преобразования в одном и том же слое упрощает многократное использование найденных архитектурных решений. Для успешного внедрения при этом требуется проработанная модель данных предприятия. После валидации данные из единого хранилища становятся доступными различным категориям потребителей через звездообразную структуру витрин.

Архитектура Кимбалла подразумевает, что данные в хранилище структурно распределены по витринам данных подразделений, через которые и обеспечиваются их очистка, стандартизация и управление. В этом случае именно в витринах хранится история данных на максимально детализированном (атомарном) уровне, а на уровень информационных систем предприятия поставляются уже согласованные по конформным измерениям факты.

Наконец, гибридная архитектура хранилища (модель Data Vault) также предусматривает очистку и стандартизацию на стадии подготовки данных к интеграции. История данных хранится в нормализованной форме на атомарном уровне в главном хранилище, а поверх нее определяются суррогатные измерения, первичные и альтернативные ключи. Согласованность и целостность связей между бизнес-ключами и суррогатными ключами также обеспечиваются на уровне хранилища, которое выполняет в данном случае еще и роль архива истории фактов, отражавшихся в витринах данных, сохраняемой на атомарном уровне. В таком виде хранилище предоставляет доступ к данным через различные витрины различным категориям потребителей. Благодаря ведению истории в самом хранилище обеспечивается возможность перезагрузки изменившихся фактов по мере накопления приращений изменений до уровня выше заданного порога гранулированности (детализации). Имеется возможность виртуализации слоя представления данных, что упрощает оперативное отображение накапливаемых изменений и совместную с бизнес-сообществом наработку данных. А вот окончательная материализация данных для нужд производства и конечных пользователей в этом случае реализуется по традиционной звездообразной схеме подключения витрин данных.

1.3.8.2 ПАКЕТНАЯ РЕГИСТРАЦИЯ ИЗМЕНЕНИЙ ДАННЫХ

Хранилища данных часто предусматривают загрузку в них накопившихся изменений раз в сутки (как правило, в ночное технологическое окно) посредством пакетной обработки вводных данных из различных операционных систем. В процессе загрузки могут выявляться и учитываться различные изменения, а используемые для этого технические процедуры регистрации изменений зависят от характера систем-источников и поступающих из них данных.

Ведение транзакционных журналов операционных баз данных с последующим внесением соответствующих изменений в DW — хороший вариант, но лишь для приложений собственной разработки, поскольку коммерческие приложения обычно не предусматривают ни изменений в настройках триггеров, ни надстроек, которые позволяли ли бы такой подход реализовать. Поэтому чаще всего используется загрузка журналов записей об изменениях с метками времени или таблиц регистрации изменений. Наконец, имея дело с устаревшими системами, не имеющими встроенных функционалов обработки меток времени (о да, сохранились еще и такие, равно

как и приложения, вовсе не ведущие собственных баз данных), приходится загружать в хранилище новые версии данных из прикладных систем целиком, перезаписывая прежние версии. Последняя процедура технически идентична процедуре пакетного восстановления данных приложения из резервной копии.

Таблица 28 в обобщенном виде описывает основные различия между известными методами регистрации изменений данных (CDC) по параметрам, определяющим их относительную сложность, оперативность и ресурсоемкость. В столбце «Повторы» указано, возможны ли ситуации дублирования в исходных системах изменений, которые ранее уже были перенесены в целевую среду. Если там стоит «+», значит, не исключены ситуации с многократной перезаписью одних и тех же фактов. В столбце «Удаления» «+», напротив, свидетельствует о пригодности метода CDC для отслеживания удаленных записей в системе-источнике, что бывает весьма полезно для выявления устаревших измерений, фактические данные по которым более не регистрируются, а потому подлежащих изъятию из многомерной модели. В тех случаях, когда удаления в системе-источнике не отслеживаются, приходится предпринимать дополнительные усилия для того, чтобы их фиксировать (см. главу 8).

Таблица 28. Сравнительные характеристики различных методов регистрации изменений данных (CDC)

Метод CDC	Требования к системе-источнику	Уровень сложности	Загрузка фактов	Загрузка измерений	Повторы	Удаления
Загрузка «дельт» данных с метками времени	Система-источник поддерживает ведение записей с метками системных даты/времени	Низкий	Быстрая	Быстрая	+	–
Загрузка таблиц журналов «дельт» данных	Изменения в системе-источнике фиксируются в табличной форме в регистрационных журналах	Средний	Средняя	Средняя	+	+
Журнал транзакций СУБД	Система регистрирует все изменения в БД на уровне журнала транзакций	Высокий	Средняя	Средняя	–	+
Публикация изменений	Система-источник публикует сообщения об изменениях в режиме реального времени	Сложнейший	Медленная	Медленная	–	+
Полная загрузка	Отметки об изменениях отсутствуют. Таблицы переписываются целиком	Простейший	Медленная	Средняя	+	+

1.3.8.3 РЕЖИМ РЕАЛЬНОГО ВРЕМЕНИ И РЕЖИМ, БЛИЗКИЙ К РЕАЛЬНОМУ ВРЕМЕНИ

С появлением операционной бизнес-аналитики (OBI), где приложения настоятельно требуют минимального запаздывания обновления данных и их интеграции в хранилище в режиме, максимально близком к режиму реального времени, возникли и новые архитектурные

подходы, позволяющие отправлять новые данные в хранилища прямиком из оперативной памяти систем-источников. Именно такой подход используется в повсеместно распространенных OBI-приложениях для АТМ (банкоматов), интернет-банков и т. п. По результатам любых банковских транзакций данные о балансе счета клиента банка обновляются в режиме реального времени и выдаются клиенту банка незамедлительно. Две ключевые архитектурно-проектировочные концепции, которые должны быть реализованы для обеспечения обмена данными в режиме, близком к реальному времени, — вычленение и регистрация в центральном хранилище данных о каждом дискретном изменении (Δ — «дельта») и отказ от пакетной обработки данных в пользу альтернативных решений, обеспечивающих незамедлительную запись транзакционных данных из оперативной памяти в хранилище.

Изменения вследствие поступления новых транзакционных данных должны обрабатываться изолированно от основного массива исторических данных постоянного хранения, находящихся в DW. Типичный архитектурный подход к изолированной обработке данных состоит в их сегментировании на подобласти с последующей раздельной обработкой и обратным слиянием. Альтернативные пакетной обработке алгоритмы позволяют минимизировать время запаздывания обновления данных в хранилище, обеспечивая максимально близкий к реальному времени режим. Используются три типа ввода транзакционных данных в хранилища: мелкими партиями, сообщениями и потоковый, которые различаются прежде всего локализацией буферной зоны (см. главу 8).

- **Мелкие партии (буферизация в системе-источнике).** Сам алгоритм регистрации новых данных в данном случае не отличается от алгоритма пакетной загрузки данных раз в сутки, но данные передаются в DW мелкими партиями и значительно чаще (например, каждые 15 минут) или по достижении порогового объема данных в буфере (например, 300 транзакций или 1Gb данных). Это позволяет переносить процессинг на дневное время, не создавая столь интенсивной нагрузки на системные ресурсы, как при ночной пакетной обработке транзакционных данных за сутки. При таком подходе важно предусмотреть очередь ожидания, в которую будут последовательно отправляться последующие пакеты в тех случаях, если обработка какого-то пакета затянулась, чтобы соблюсти правильную хронологическую очередность загрузки данных в хранилище.
- **Сообщения (буферизация в шине обмена).** Взаимодействие систем в режиме, близком к реальному времени, посредством обмена сообщениями о каждой операции изменения, добавления или удаления данных (записи, событии или транзакции) осуществляется через шину предприятия. Системы-источники публикуют сообщения об изменении данных на шине, а системы-подписчики последовательно их считывают и обрабатывают, внося по мере надобности изменения в хранилище данных. Системы-источники и целевые (абонентские) системы-получатели функционируют полностью независимо друг от друга. Этот метод очень часто используется в архитектурах систем предоставления доступа к данным как услуги (DaaS).

- **Потоковый сбор (буферизация в целевой системе).** Система-источник незамедлительно отправляет данные о каждой транзакции или событии в буферную зону или очередь целевой системы, а та их собирает и обрабатывает в порядке поступления. Результаты или некие суммарные данные затем могут с некоторой задержкой передаваться системой-получателем в хранилище данных через другую очередь сообщений.

2. ПРОВОДИМЫЕ РАБОТЫ

2.1 Выработка понимания требований к DW

Проектирование хранилищ данных принципиально отличается от разработки операционных систем. На операционные системы накладываются точные и специфичные требования. В хранилищах данных собираются данные самого разнообразного назначения. Более того, характер использования данных постоянно эволюционирует по мере их анализа и появления новых областей применения. Поэтому на начальных фазах нужно уделить достаточно времени и внимания уяснению функциональных требований к DW и определению источников данных, в полной мере соответствующих этим требованиям. Время, потраченное на получение ответов на концептуальные вопросы в начале проекта, многократно окупится за счет снижения издержек на переработки проекта из-за несоответствий систем обработки данных требованиям или доступным источникам фактических данных.

При сборе требований к проектам DW/BI начать лучше с определения целей и стратегии бизнеса. Выявите и очертите области деятельности, затем выявите ключевых людей в каждой области и детально обсудите с ними, чем именно они занимаются и почему. Зафиксируйте конкретные вопросы, которыми они задаются сегодня, а также вопросы, ответы на которые рассчитывают получить с помощью новых данных. Задокументируйте критерии, по которым они отличают значимую информацию от малозначимой, и классифицируйте важнейшие аспекты значимой информации. По возможности определите и зафиксируйте ключевые рабочие показатели и формулы их расчета. Это поможет в раскрытии и формализации бизнес-правил, используемых для автоматизации контроля качества данных.

Каталогизируйте требования и выберите из каждой группы приоритетные с точки зрения их необходимости на стадии ввода систем DW/BI в эксплуатацию и освоения их сотрудниками; работу над выполнением остальных требований можно отложить на будущее. Выберите самые простые и ценные с точки зрения мгновенного эффекта элементы для реализации в первой версии проекта DW/BI. Описание проекта должно производить должный рекламный эффект на всех заинтересованных лиц, а для этого в нем должны в целостном контексте учитываться требования всех затрагиваемых бизнес-подразделений и/или процессов.

2.2 Определение и сопровождение архитектуры DW/BI

Архитектура DW/BI должна описывать, откуда берутся, куда и когда отправляются данные, зачем и как собираются в хранилище. При этом ответы на вопрос «как» должны быть

детализированными и описывать конкретные аппаратные и программные компоненты систем и организационную модель их интеграции. Технические требования должны включать спецификации производительности, доступности и времени обработки (см. главы 4 и 8).

2.2.1 Определение технической архитектуры DW/BI

Идеальная архитектура DW/BI должна изначально предусматривать механизм обратной связи, обеспечивающий поступление в DW транзакционных и операционных отчетов должного уровня детализации. Этот механизм призван избавить DW от обработки деталей каждой транзакции. Например, можно реализовать механизм просмотра операционных отчетов или форм по транзакционному ключу — например, № счета-фактуры. Клиентов неизменно интересуют все детали, но часть операционных данных, в частности поля текстовых описаний, значимы только в контексте отчета об исходной операции, тогда как никакой аналитической ценности не представляют и в среде DW/BI абсолютно излишни.

Начать следует с выбора архитектурной концепции. Многие работы нужно изначально правильно выстраивать, чтобы обеспечить согласованность нефункциональных требований с нуждами бизнеса. В связи с этим бывает полезным тестирование прототипов с целью оперативного подтверждения или опровержения ключевых гипотез до дорогостоящих вложений в технологии или архитектурные проекты. Кроме того, программы информационно-разъяснительной работы с бизнес-сообществом существенно расширяют возможности команды, реализующей санкционированные бизнесом изменения по содействию переходу на новую систему и обеспечению ее успешной эксплуатации.

Естественным продолжением этого трансформационного процесса является обеспечение полной согласованности архитектуры DW/BI с корпоративной моделью данных (или, как минимум, подтверждение отсутствия явных противоречий между ними). Поскольку основное внимание уделяется изучению структур данных, используемых различными организационными подразделениями на различных участках работы, обязательно проверьте, соответствует ли имеющаяся физическая инфраструктура задокументированной логической модели данных предприятия. При выявлении расхождений или пробелов внесите все необходимые поправки, иначе ошибки на стадии реализации появятся неизбежно.

2.2.2 Определение процессов управления DW/BI

Управление DW/BI в режиме производственной эксплуатации должно осуществляться скоординированным образом и включать полный комплекс необходимых регламентных работ и регулярный выпуск обновлений, а также — по согласованию с бизнес-сообществом — новых версий.

Обязательно должен иметься план-график выпуска стандартных обновлений (см. раздел 2.6). В идеале проектировщикам DW/BI следует выпускать пакетные обновления ПО для развернутых на местах продуктов не только с исправлениями и улучшениями, но и с функциональными дополнениями. Наличие плана-графика выпуска обновлений и/или новых версий

позволяет лучше планировать потребности и ресурсы, а также стандартизировать расписания поставок. Используйте предварительные внутренние выпуски для экспериментов по оптимизации стандартного графика обновлений, распределения ресурсов и оценки полученных результатов.

Установившийся отлаженный процесс выпуска и распространения обновлений также будет способствовать выработке у бизнес-менеджеров понимания, что «патчи» и «релизы» предназначены для совершенствования ИТ-продукта и, как следствие, повышения качества обработки данных, а не для устранения задним числом обнаружившихся проблем. Критически важно работать с прицелом на будущее, в тесном сотрудничестве в рамках кросс-функциональной команды, ибо такой подход способствует неуклонному наращиванию и расширению функциональности продукта, — в отличие от систем техподдержки по заявкам пользователей, снижающих доверие к продукту.

2.3 Проектирование и разработка хранилища и витрин данных

Обычно работы по проектированию DW/BI ведутся параллельно по трем направлениям.

- **Данные**: определяются информационно-аналитические потребности бизнеса и источники данных, позволяющие их удовлетворить. Помимо выявления наилучших источников данных на этом же направлении прорабатываются правила редактирования, преобразования, интеграции и хранения данных, порядок доступа к ним приложений и пользователей, а также выявления и отбраковки некачественных данных.
- **Технологии**: проектирование служебных систем и процессов, обеспечивающих функционирование хранилища и движение потоков данных. Фундаментальным требованием является интеграция технологий DW/BI с существующей архитектурой предприятия, поскольку DW — не вещь в себе. Проектированием включения новых технологий в *архитектуру предприятия* обычно занимаются специалисты по ИТ и проектированию приложений.
- **Бизнес-аналитический инструментарий**: разработка пакета приложений для потребителей данных, позволяющий получать вразумительную картину на основании реализованных программных продуктов по работе с данными.

2.3.1 Мэппинг источников данных в целевые структуры

Мэппинг источников данных в целевые структуры задает правила преобразования для сущностей и элементов данных при передаче от отдельных источников в целевые системы. Помимо правил документируется происхождение каждого элемента данных, начиная с целевой системы вплоть до первоисточника.

Самая сложная часть мэппинга — определение корректных связок или отношений эквивалентности между элементами данных во множественных системах. Задумайтесь, каких усилий требует консолидация в DW входящих из множества автоматизированных систем формирования счетов и проводки платежей или управления заказами. И всегда есть риск неверного

сопоставления данных, особенно если одни и те же данные фигурируют в различных таблицах и полях под разными именами или по-разному структурированы.

Надежная таксономия — непременное условие корректного сопоставления элементов данных в различных системах и обеспечения согласованности и непротиворечивости структуры данных в DW. Обычно таксономия определяется логической моделью данных, а зачастую оба эти понятия и вовсе эквивалентны. В процессе мэппинга допускается присоединение, перестановка или вставка элементов данных в различные структуры при условии соблюдения требования сохранения логической целостности.

2.3.2 Исправление и преобразование данных

Работы по исправлению или очистке данных призваны обеспечить соблюдение стандартов посредством проверки корректности и исправления данных, содержащих недопустимые значения. Особенно важно проверять и исправлять данные при первичных загрузках из источников со значительной предысторией. Чтобы не допускать избыточного усложнения целевых систем, проверку и исправление данных лучше проводить в системах-источниках перед выгрузкой.

Также выработайте стратегию действий в отношении строк данных, некорректность которых обнаружилась уже после загрузки в DW. Политика удаления старых записей сама по себе способна привнести некоторый хаос в связанные таблицы и суррогатные ключи; возможно, лучше предусмотреть как вариант следующую последовательность действий: старая строка помечается как устаревшая, а новые данные записываются посредством добавления новой строки.

При оптимистичном прогнозе можно выбрать стратегию загрузки через создание строк в таблицах измерений, позволяющих размещать импортируемые из системы-источника данные в соответствующих таблицах фактов. При подобной процедуре важно заранее определить порядок учета, обновления и списания таких записей как устаревших.

При пессимистичном прогнозе стратегия должна предусматривать наличие области сбора и переработки отбракованных фактических данных, которые не удается связать с имеющимися ключами измерений. При выявлении таких записей система должна выдавать уведомление или предупреждение и ставить их на контроль, чтобы отбракованные записи можно было впоследствии отследить, по возможности исправить и перезагрузить. При этом алгоритмы обработки задач по загрузке фактов должны предусматривать первоочередную проверку и загрузку ранее отбракованных и исправленных записей и лишь после этого переходить к обработке впервые поступившего нового контента.

Основная задача в части преобразования данных заключается в настройке бизнес-правил в технических системах. Интеграция данных невозможна без их приведения к структуре, определяемой моделью, посредством преобразования. Для корректного определения правил преобразования и интеграции часто требуется непосредственное участие распорядителей данных или экспертов в предметных областях. Все правила должны документироваться, чтобы в дальнейшем ими можно было управлять. Специализированные программные средства интеграции данных позволяют поставить решение всех подобных задач на поток (см. главу 8).

2.4 Заполнение хранилища данных

Самая трудоемкая часть работы по созданию DW/BI — подготовка к приему, обработке и сохранению данных, поступающих из различных источников, в рабочем режиме. Архитектура и модель данных определяют детальное содержание DW и являются ключевым приоритетом при проектировании системной архитектуры DW/BI. Публикация четких правил, расписывающих, какие данные будут доступны только через каналы учета текущих операций (то есть не будут отправляться в DW), — важнейшее условие успеха всего проекта создания DW/BI.

Ключевыми факторами, которые следует учитывать при определении подхода к заполнению DW, выступают требуемое время задержки, доступность систем-источников, «окна» пакетной обработки или интервалы загрузки данных, целевые базы данных, аспекты измерений и согласованность временны́х интервалов обновления данных в хранилище и в витринах. Кроме того, выработанный подход должен предусматривать процедуры контроля качества данных, а также учитывать затраты времени на преобразование данных и задержку обновления измерений.

Еще один важный аспект определения подхода к заполнению DW — обязательность регистрации изменений в ранее накопленных данных, что требует выявления изменений в системах-источниках, интеграции данных об этих изменениях и согласования этих сводных данных по времени. Некоторые базы данных предлагают функциональность ведения журнала всех изменений, что позволяет средству интеграции работать с такой базой данных без обращения к ее таблицам, а просто интегрируя данные об изменениях в записи DW и уведомляя пользователей об изменениях. Если же подобных функций в системе-источнике не предусмотрено, для их реализации приходится писать исполняемые сценарии, макросы или подпрограммы. Наконец, имеется целый ряд хорошо отработанных технических приемов, позволяющих проектировщикам DW обеспечивать интеграцию и минимизировать время запаздывания обработки данных из множества разнородных входящих потоков.

На этом итерационный цикл создания центрального DW можно считать завершенным, и приходит черед наработки дополнительных функциональных возможностей и подключения к проекту новых бизнес-подразделений. Тут никак ни обойтись без освоения новых технологий, процессов и навыков, как и без тщательного планирования с учетом мельчайших деталей и нюансов. Но все приращения ниже по технологическому потоку — суть надстройка над вышеописанным фундаментом, поэтому рекомендуется по максимуму инвестировать средства в обеспечение устойчиво высокого качества данных, технической архитектуры и производственной среды. Постарайтесь предусмотреть и по возможности автоматизировать все процессы, необходимые для полного и своевременного выявления ошибок в данных, поступающих на интеграцию из пользовательских приложений.

2.5 Внедрение портфеля инструментов BI

Внедрение портфеля BI-приложений требует прежде всего грамотного выбора программных средств и инструментов, требующихся различным сообществам пользователей или бизнес-подразделениям. Старайтесь выявлять и по максимуму использовать сходства и аналогии

в бизнес-процессах, аналитических подходах, стилях управления, требованиях и стандартах, с тем чтобы обеспечить как можно более высокую степень единообразия и согласованности BI-инструментария во всех предметных областях.

2.5.1 Распределение пользователей по группам в соответствии с потребностями

Целевые группы пользователей должны определяться на основе изучения спектра их потребностей в BI-средствах. Прежде всего выделите основные группы пользователей и изучите их состав, а затем определите применительно к каждому инструменту BI, какие группы в нем нуждаются, а какие нет. На одном полюсе спектра окажутся разработчики ИТ-решений и специалисты по технической обработке данных, которым требуется доступ к самым продвинутым функционалам. На другом — конечные потребители информационных продуктов, которым важен быстрый доступ к готовым отчетам, сводкам и т. п. Но и потребителям помимо статичных данных могут требоваться некоторые интерактивные аналитические средства: например, настройки глубины детализации, фильтров и сортировки данных в отображаемых представлениях.

Следует предусмотреть возможность перевода пользователей из одной группы или категории в другую: например, при повышении уровня профессионализма, переходе в другое функциональное подразделение и т. п. Уровни доступов к данным различных категорий в каждой группе должны быть дифференцированными. Например, снабженцам достаточно статичных сводных финансовых отчетов, но требуются максимально детализированные интерактивные представления данных из системы учета складских запасов с максимальной аналитической оснасткой. И, напротив, финансовым аналитикам и бухгалтерам производственных подразделений должен быть разрешен полноправный динамический доступ к детализациям приходно-расходных статей, но вполне достаточно статичной сводки наличных материальных ресурсов. Руководителям высшего и среднего звена потребуется сочетание фиксированного набора отчетов в установленной форме, приборных панелей и оценочных карт. Менеджеры и продвинутые пользователи имеют склонность вгрызаться в профильные отчеты и раскладывать данные из них по косточкам, доискиваясь до корневых причин имеющихся проблем. Внешним потребителям могут потребоваться любые из вышеописанных средств BI, поскольку среди них присутствуют лица с самыми разнообразными интересами и опытом.

2.5.2 Обеспечение соответствия инструментария потребностям пользователей

Рынок изобилует всевозможными инструментальными средствами для создания разнообразнейших оперативных, статистических и аналитических отчетов. Крупные поставщики BI-решений сегодня включают в пакеты предлагаемого пользователям ПО функционалы составления классических, выверенных до пикселя отчетов, которые еще недавно были доступны только в узкоспециализированных приложениях по формированию отчетности. Многие разработчики коммерческих приложений встраивают в них аналитические модули, черпающие данные из стандартного справочного набора, содержащегося во включенных в пакет сводных таблицах или кубах данных. Виртуализация размыла границу между внутренними и внешними источниками,

коммерческими и бесплатными данными, что открывает перед пользователями соблазнительную возможность самостоятельно по мере надобности включать в создаваемые отчеты данные, полученные откуда угодно. Иными словами, компаниям, хотя бы из соображений разумной предосторожности, следует использовать единую инфраструктуру и механизмы ведения и распространения внутренней отчетности. Это требование распространяется и на веб-публикации, и на рассылки по электронной почте, и на приложения, включая DW/BI, генерирующие и распространяющие пресс-релизы и отчеты.

Многие поставщики в последнее время занялись интеграцией средств BI в модульные наборы или пакеты инструментов BI. Приобретение готового набора средств BI — удобное решение на уровне корпоративной архитектуры, но лишь при условии его планирования с нуля. В противном случае следует тщательно взвесить все «за» и «против», поскольку в организации наверняка имеются аналогичные программные средства, ранее приобретенные «в розницу» или полученные из открытых источников, и при внедрении купленного «оптом» пакета возникнет масса вопросов с обеспечением совместимости или заменой старого, привычного и проверенного программного обеспечения новым. Важно помнить, что любое BI-приложение, помимо цены, уплаченной за него поставщику согласно прайс-листу, влечет еще и накладные расходы, обусловленные, в частности, потребностями в системных ресурсах, технической поддержке, переобучении пользователей и интеграции в архитектуру предприятия.

2.6 Сопровождение информационных продуктов

Внедренное DW и его ориентированная на потребителей данных часть, включающая клиентские приложения и инструменты BI, превращаются, по сути, в информационный продукт. Последующие усовершенствования платформы DW (дополнения, надстройки и/или модификации) следует внедрять поэтапно, методом приращений.

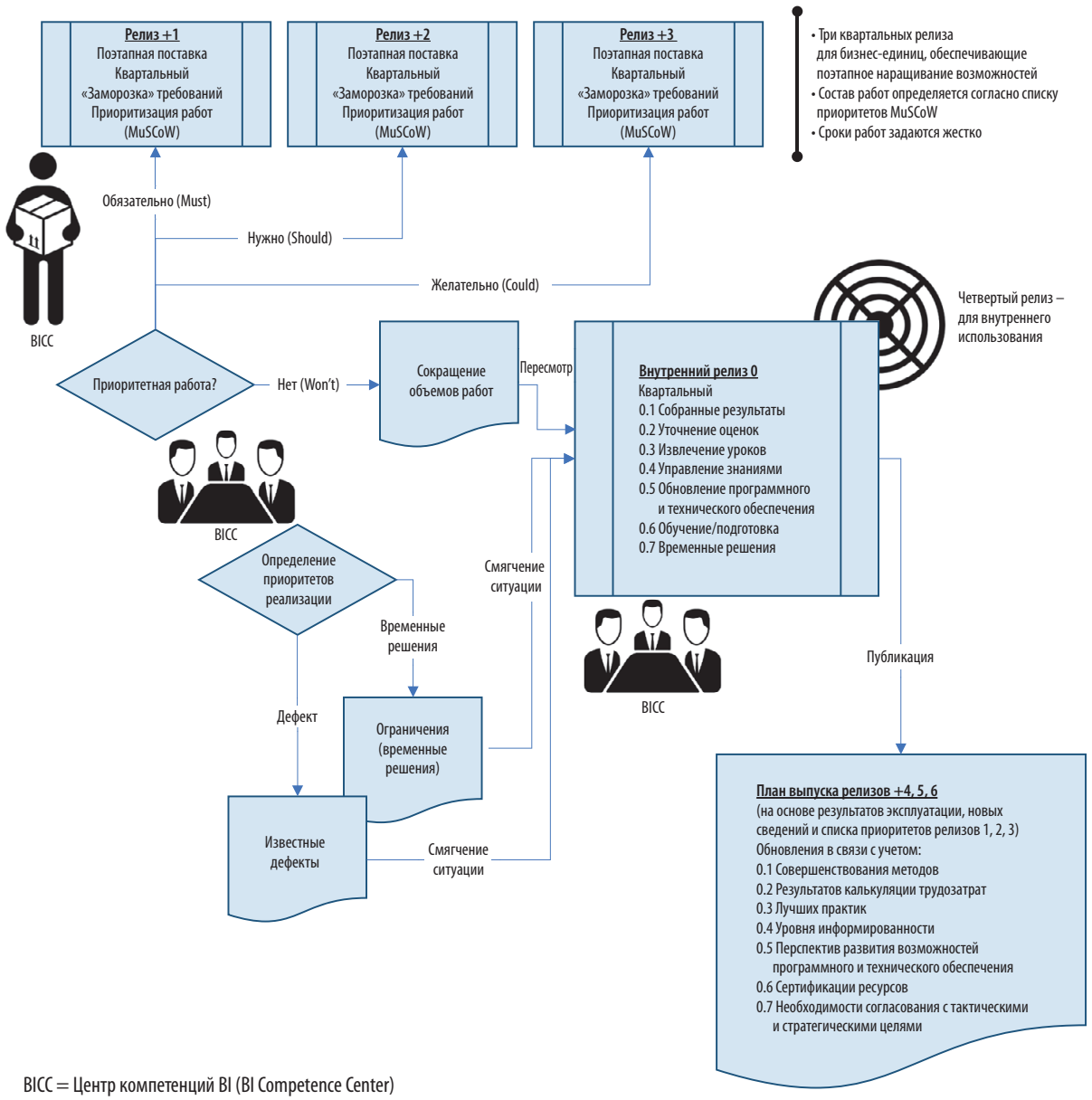
Сопровождение комплекса DW/BI должно включать регулярное наращивание функциональных возможностей по всем ключевым параметрам, вот только добиться этого в динамической рабочей среде бывает не так-то просто. Совместно с бизнес-партнерами очертите круг основных приоритетов — и сфокусируйте основные усилия на самых необходимых доработках и усовершенствованиях.

2.6.1 Управление релизами

Процесс управления релизами — один из важнейших в группе процессов пошаговой разработки любых ИТ-продуктов, необходимый для расширения их возможностей за счет добавления новых функций, повышения производительности и стабильности приложений и эксплуатационной среды, а также просто гарантированного и регулярного сопровождения программного обеспечения всех информационных систем организации. Применительно к DW/BI управление релизами — процесс, необходимый для поддержания хранилища в актуальном рабочем состоянии, а также совершенствования функциональности и повышения производительности приложений. Однако этот процесс не должен передаваться всецело на усмотрение разработчиков ИТ-решений,

а требует участия также и бизнес-подразделений, и проектировщиков модели данных, и бизнес-аналитиков. Это комплекс совместных действий по непрерывному улучшению.

Рисунок 83 содержит иллюстративный пример организации процесса управления выпуском релизов по ежеквартальному плану-графику. За год выпускаются три релиза для бизнес-нужд и еще один — для технологических (служебный релиз для обслуживания хранилища данных). Процесс должен предусматривать поэтапное развитие DW/BI и ведение хронологического архива выполненных технических и бизнес-требований.



BICC = Центр компетенций BI (BI Competence Center)

Рисунок 83. Пример процесса управления релизами

2.6.2 Управление жизненным циклом разработки информационного продукта

Пока потребители данных пользуются текущей версией DW, разработчики ведут подготовку к следующей итерации, отдавая себе отчет в том, что не все их заготовки пригодятся для внедрения в эксплуатационную среду. Согласовывайте состав выпусков релизов со списком приоритетных работ, согласованных с бизнес-подразделениями. Каждая итерация должна способствовать расширению функциональных возможностей существующей реализации за счет усовершенствования имеющихся или добавления новых функций или предусматривать подключение к комплексу DW/BI нового бизнес-подразделения. Новые релизы выпускаются для приведения функциональности DW/BI в соответствие с потребностями бизнес-подразделений, а итерации позволяют менеджеру продукта отлаживать новые BI-функции путем их согласования с конфигурацией DW.

Компоненты, доведенные до состояния, устраивающего бизнес, и достаточного уровня технической проработки, передаются на окончательное согласование, при необходимости корректируются, а затем реализуются в опытно-экспериментальной среде или «песочнице», где бизнес-пользователи изыскивают новые подходы, апробируют новые приемы и разрабатывают новые модели данных или алгоритмы обучения, в том числе и методом проб и ошибок. «Песочница» не требует столь же строгого административного контроля и надзора, какой должен быть непременно предусмотрен в других открытых для бизнес-пользователей областях, однако в той или иной форме расстановка приоритетов необходима и в «песочнице».

По аналогии с традиционной средой контроля качества или тестирования этот опытный участок предназначен прежде всего для проверки пригодности экспериментальных решений для использования в реальных условиях. От результатов пилотных испытаний зависят дальнейшие шаги в отношении апробируемого элемента. Главное — не спешить с бездумным внедрением сработавшего в опытной среде решения без учета его влияния на качество данных и согласования всех административных аспектов внедрения и последующего использования. Выживаемость продукта в эксплуатационной среде — вопрос экзистенциальный: лишь высочайшее качество практической реализации служит основанием для его допуска в эксплуатацию.

Лишь новинки, успешно прошедшие испытания и утвержденные к внедрению руководством как ИТ-, так и бизнес-подразделения, могут внедряться в среде DW/BI в качестве новых информационно-аналитических продуктов. После внедрения итерация считается завершенной.

Решения, тестирование которых в экспериментальной среде дало неудовлетворительные результаты, либо полностью отвергаются, либо возвращаются на доработку. Во втором случае разработчикам ПО могут потребоваться дополнительные консультации с командой DW для повышения шансов на успешное прохождение их продуктом приемочных испытаний на следующей итерации.

2.6.3 Мониторинг и оптимизация нагрузки

Рабочая нагрузка должна контролироваться на уровне всех реализованных в системе процессов, что позволяет отслеживать причины задержек обработки и прохождения данных вверх по

технологической цепочке вплоть до узкого места, которое выступает первопричиной низкой производительности.

По мере надобности используйте средства оптимизации баз данных, включая разбиение на разделы, усовершенствованные стратегии резервного копирования и восстановления. Впрочем, управление архивными копиями в среде DW — отдельная трудная тема для обсуждения.

Пользователи часто считают DW чем-то вроде архива в активном доступе из-за наличия там длительных историй накопления данных, и знать не желают, что DW само по себе нуждается в архивировании и ведении резервных копий, — а потом удивляются пропаже записей, в особенности из интерактивных источников аналитических данных (OLAP) (см. главу 6).

2.6.4 Мониторинг использования и настройка производительности средств BI

Оптимальной считается практика мониторинга и выявления потребностей в настройке BI-систем и приложений посредством определения и отображения текущих показателей удовлетворенности потребителей. Примеры других полезных показателей — среднее время отклика на запрос, число пользователей и/или запросов в сутки, неделю или месяц. Помимо статистических метрик, получаемых непосредственно из систем, полезно проводить и регулярные опросы или анкетирования пользователей комплекса DW/BI.

Регулярный учет статистики и структуры запросов к данным позволяет формировать отчеты о частоте и объемах запросов и использования данных, а данные из этих отчетов использовать для рационального планирования усовершенствований. Тонкая настройка BI-средств, в принципе, аналогична профилированию приложений с целью выявления узких мест и процессов, требующих оптимизации. Индексирование и кластеризацию данных лучше всего производить с учетом распределения и статистики потребления. Колоссальное повышение производительности может достигаться и за счет элементарных организационных решений — например, ежедневной публикации сводного отчета с данными, которые в противном случае будут запрашиваться порознь сотни и тысячи раз на дню.

Прозрачность и наглядность — ключевые принципы мониторинга DW/BI. Чем больше деталей о работе комплекса DW/BI отображается, тем проще потребителям понять, откуда берутся и как анализируются данные, — что повышает доверие к BI и снижает число обращений в службу поддержки за разъяснениями.

Панель отображения основных высокоуровневых показателей доставки данных с поддержкой их дальнейшей углубленной детализации — оптимальное с точки зрения практичности решение, позволяющее удовлетворять информационный спрос как технического персонала, так и пользователей.

Добавление к показателям скорости/времени выдачи показателей качества данных повысит ценность такой информационной панели. Еще одним полезным инструментом мониторинга являются тепловые карты визуализации нагрузок на инфраструктуру, интенсивности потоков данных и соблюдения параметров соглашений об уровнях обслуживания.

3. ИНСТРУМЕНТЫ

Процесс выбора исходного набора инструментов может оказаться долгим и непростым. Ведь нужно постараться сделать так, чтобы выбранный инструментарий обеспечивал удовлетворение и первоочередных насущных потребностей, и нефункциональных спецификаций, и даже еще не сформулированных требований в отношении решений следующего поколения. Ускорить выбор помогают готовые наборы критериев принятия решений и инструментов внедрения процессов, а также привлечение профильных специалистов. Важно учитывать и сравнивать по позициям не только традиционные опции построить самим vs купить готовое, но и предложения категории ПО как услуга (SaaS). Помимо собственно программных средств, концепция SaaS предлагает еще и практический опыт интеграции решений в среду организации, и сбалансированное соотношение затрат/отдачи по сравнению со стоимостью постройки с нуля собственного или издержками интеграции разнообразного покупного ПО. Приплюсуйте сюда еще и затраты на регулярные обновления и (потенциально) замену неподходящих продуктов другими. Выстраивание же отношений с разработчиками на основе соглашений об уровнях обслуживания (SLA) или уровнях операционной поддержки (OLA) служит мостом к предсказуемости затрат и позволяет обеспечить себя всеми необходимыми функциями за привлекательно невысокую абонентскую плату и перекладывать издержки, проистекающие от сбоев в работе систем и/или приложений, на поставщика SaaS либо компенсировать их причитающимися с него пенями или неустойками.

3.1 Репозиторий метаданных

В крупных организациях часто складывается ситуация, когда в различных подразделениях используется множество разрозненных программных средств от разных поставщиков, а нередко еще и в операционных средах различных версий. В таких случаях ключевым компонентом создания единого DW является работа по «сшиванию» данных из лоскутных источников воедино через определение и ведение метаданных. Для автоматизации процессов интеграции метаданных и управления репозиторием метаданных могут использоваться различные методики и решения (см. главу 12).

3.1.1 Словари и/или глоссарии данных

Словарь данных насущно необходим пользователям комплекса DW/BI. В словаре должны содержаться четкие и понятные бизнес-пользователям определения всех нужных им элементов данных (с указанием типа, деталей или структуры данных и применимых правил ИБ и защиты данных). Содержание словарей данных часто берется непосредственно из логической модели данных. Закладывайте фундамент обеспечения качества метаданных на уровне требований, предъявляемых к разработчикам моделей, которые обязаны дисциплинированно подходить к составлению и ведению словарей.

В некоторых организациях бизнес-пользователи также активно участвуют в разработке словарей или глоссариев данных, предлагая термины, определения, а впоследствии исправления и уточнения определений элементов данных в своих предметных областях. Координируйте такие усилия с помощью средств совместной работы, отслеживайте их через Центр компетенций, обеспечивайте гарантированное сохранение созданного контента в логической модели. Обеспечение согласованности между бизнес-контентом и технической терминологией на всех уровнях вплоть до физической модели данных способствует минимизации риска последующих ошибок и переработок из-за неверной интерпретации данных и/или их определений (см. главу 13).

3.1.2 Происхождение данных и модели данных

Многие средства интеграции данных включают инструменты анализа их происхождения, которые принимают во внимание как коды программ для заполнения DW, так и физические модели, и сами базы данных. Иногда предлагаются и веб-интерфейсы для мониторинга и обновления определений и других метаданных. Задокументированное происхождение данных находит множество полезных применений, включая:

- расследование первопричин недостоверности, неточности и иных проблем с качеством данных;
- анализ последствий системных изменений или проблем с данными;
- составление рейтингов источников по показателям надежности и качества данных.

Постарайтесь реализовать интегрированное средство отслеживания происхождения данных, позволяющее разбираться с маршрутами движения и алгоритмами преобразования всех элементов данных в процессе загрузки, а также использования конечными пользователями в отчетах и аналитике. Отчеты с анализом зависимостей и последствий помогут выделять компоненты, на которых скажутся потенциальные изменения, ускорять и оптимизировать выполнение различных задач по оценке состояния и обслуживанию систем обработки данных.

Многие ключевые бизнес-процессы, взаимосвязи и термины выявляются и объясняются на стадии разработки модели данных. Большое количество информации подобного рода, учтенной в логической модели данных, теряется или игнорируется на стадии проектирования физической модели или ее реализации в производственной среде. Поэтому критически важно обеспечивать сохранность архивов проектной документации прежних версий логических и физических моделей данных даже после успешного ввода в эксплуатацию систем, основанных на видоизмененных моделях.

3.2 Средства интеграции данных

Средства интеграции данных используются для заполнения хранилища. Помимо интеграции они обеспечивают возможность планирования расписаний загрузки данных по сложным схемам из множественных источников. При выборе программного средства следует учитывать также, поддерживает ли оно следующие дополнительные функции управления системой:

- аудит, контроль, перезапуск и планирование графиков загрузок, интеграции, оптимизации и иных процессов;
- выборочное извлечение в заданное время конфигурируемой выборки данных из хранилища с передачей ее в систему аудита или контроля качества данных;
- контроль произведенных и неудавшихся операций с последующим перезапуском сбойного или отмененного процесса (см. главу 8).

Множество программных средств интеграции данных выпускается в пакете с портфелем BI-инструментов, поддерживает отправку и прием рабочих сообщений и e-mail или даже управление семантическими слоями. Интеграция рабочего процесса может служить стимулом к дальнейшей проработке процессов контроля качества данных, выявления, разрешения и передачи на надлежащий уровень рассмотрения различных проблем с качеством данных и систем-источников. Отправка сообщений по e-mail или запуск процедур обработки предупреждений по получении e-mail-уведомлений также стали распространенной практикой, особенно в приложениях для мобильных устройств. Кроме того, способность средства интеграции доставлять контент целевым адресатам в виде семантического слоя делает его вполне подходящим и для виртуализации данных в гибких распределенных реализациях.

3.3 Типы инструментов BI

Зрелость рынка и широта спектра BI-приложений, предлагаемых на коммерческой основе, делает разработку компаниями собственных BI-инструментов нецелесообразной[1]. Настоящий раздел посвящен вводному обзору основных типов инструментов, представленных на рынке BI-приложений, и их характеристик, с тем чтобы помочь организациям в выборе наиболее подходящих программных средств с точки зрения пользовательского функционала в зависимости от уровня потребителей *бизнес-аналитики*. Инструменты BI стремительно развиваются и совершенствуются, открывая возможность для перехода от стандартизированной отчетности, диктуемой спецификой используемых информационных технологий, к самостоятельному исследованию данных по направлениям, интересующим бизнес[2].

- **Операционная отчетность** позволяет выявлять и анализировать краткосрочные (помесячные) и среднесрочные (годовые) тенденции и закономерности. Используйте тактический бизнес-анализ (Tactical BI) для выработки и принятия краткосрочных решений в сфере оперативного управления бизнесом.

[1] Материал настоящего раздела по большей части позаимствован из книги: Cindi Howson, «The Business Intelligence Market: Secrets to Making BI a Killer App», McGraw-Hill, 2008, — с разрешения правообладателя и с некоторыми изменениями и дополнениями.

[2] Портал Dataversity описывает эту тенденцию понятием «демократизация технологий работы с данными». Подробнее см.: Ghosh, Paramita. «A Comparative Study of Business Intelligence and Analytics Market Trends». Dataversity. January 17, 2017 (http://bit.ly/2sTgXTJ, ссылка проверена 30.05.19).

- **Управление эффективностью бизнеса** позволяет производить формальную оценку измеримых показателей, соответствующих целям организации. Осуществляется, как правило, на уровне высшего руководства. Используйте стратегический бизнес-анализ (Strategic BI) для формулировки долгосрочных целей и задач.
- **Описательная (descriptive) аналитика, аналитика самообслуживания (self-service)** обеспечивают прикладной анализ текущих вопросов. Сочетание BI-приложений с функциями и процессами операционного управления позволяет задействовать анализ данных в принятии решений в режиме, близком к реальному времени. Требование минимизации времени задержки регистрации и доставки данных (максимальной приближенности к режиму реального времени) диктует выбор архитектуры решений. Для полноценной операционной аналитики сегодня требуются сервис-ориентированная архитектура (SOA) и большие данные (см. главы 8 и 15).

3.3.1 Операционная отчетность

Средства операционной отчетности позволяют генерировать и выводить отчеты непосредственно из транзакционных систем, рабочих приложений или хранилищ данных. Обычно реализуются как функционалы приложений. Очень часть первоначальным применением средств BI является генерирование операционной отчетности, особенно если высокоуровневое распоряжение DW/BI не налажено или в DW содержатся дополнительные по отношению к оперативным/транзакционным данные, учет которых необходим или полезен. Часто операционные отчеты внешне похожи на результаты обработки нестандартных запросов, а на деле представляют собой простые отчеты или вводные для какого-либо рабочего процесса. С точки зрения управления данными ключевым в таких случаях является вопрос о том, достаточно ли приложению собственных данных для генерирования отчета или же ему требуются еще и дополнительные данные из DW или ODS.

Инструменты исследования данных и формирования отчетности иногда еще называют средствами создания произвольных запросов, поскольку они позволяют пользователям создавать «авторские» отчеты или выборки данных, предназначенных для использования в качестве вводных другими пользователями или процессами. Строгого соблюдения каких-либо стандартных требований к структуре/формату документа в данном случае не предъявляется, поскольку речь идет не о счетах-фактурах или чем-то подобном. Зато пользователям часто интуитивно хочется включать в такие отчеты графики и таблицы. Зачастую созданные бизнес-пользователями с помощью произвольных запросов отчеты оказываются настолько удачными, что утверждаются в качестве стандартной формы внутриорганизационной отчетности по затрагиваемому в них кругу вопросов.

Требующиеся бизнесу операционные отчеты часто не совпадают с отчетами, генерируемыми по стандартным запросам, которые обычно (хотя и не всегда) используют в качестве источника DW или как предназначенную для соответствующего бизнес-подразделения витрину данных. Кроме того, стандартные отчеты обычно разрабатываются ИТ-специалистами, а произвольные — продвинутыми бизнес-пользователями с помощью программных средств построения запросов. При необходимости созданные пользователями запросы и отчеты можно утверждать к регулярному использованию в рамках отдела или всего предприятия.

Производственная отчетность часто выходит за рамки DW/BI и включает запросы к транзакционным системам с целью получения вводных для таких оперативных документов, как счета-фактуры или банковские выписки. Запросы и форматы производственных отчетов обычно разрабатываются ИТ-специалистами.

Традиционные инструменты BI включают ряд стандартных средств наглядного представления данных в виде таблиц, секторных и линейных графиков, столбцов, гистограмм и т. д. и т. п. Помимо статичных форматов визуализации, используемых в отчетах для публикации, возможны также динамические и даже интерактивные форматы в онлайновых отчетах, вплоть до поддерживающих масштабирование, навигацию по уровням детализации и/или применение фильтров с целью упрощения анализа данных в визуализированном представлении. Может быть предусмотрено также и пользовательское переключение между различными типами графиков и/или режимами их отображения (см. главу 14).

3.3.2 Управление эффективностью бизнеса

Управление эффективностью бизнеса (Business Performance Management, BPM) — это набор интегрированных процессов организации и приложений, разработанных для оптимизации исполнения бизнес-стратегии. Стандартный набор поддерживает формирование бюджета, планирование, бухгалтерский учет и сводную финансовую отчетность. Наработки в этом сегменте имеются огромные, поскольку производители программного обеспечения как для управления предприятием (ERP), так и для BI видят в данной области огромные резервы роста, к тому же грань между бизнес-аналитикой и управлением эффективностью всё более стирается. Частота приобретения клиентами решений в области BI и управления эффективностью от одного и того же разработчика зависит от возможностей поставляемых им продуктов.

Не вдаваясь в подробности, отметим лишь, что технология BPM позволяет приводить процессы в соответствие с организационными целями. Ключевые элементы BPM — измерения и петля положительной обратной связи. В сфере BI это приняло форму множества приложений для различных стратегических областей деятельности предприятия — бюджетного планирования, прогнозирования, планирования ресурсов и т. п. Другая специализация BI сформировалась строго внутри этой области и включает создание карт балльной оценки в связке с приборными панелями для интерактивного информирования пользователей. Как и в автомобиле, на приборную панель, находящуюся в поле зрения конечного пользователя, выводится сводка текущих значений важнейших показателей (Eckerson, 2005).

3.3.3 Приложения для оперативного анализа

Термин *аналитические приложения* (*analytic applications*) ввел в 1990-х годах Генри Моррис из International Data Corporation (IDC), тем самым подчеркивая их отличия от технологий онлайновой аналитической обработки (OLAP[1]) и бизнес-аналитики (Morris, 1999). Аналитические

[1] *сокр.* от *англ.* **On**line **A**nalytical **P**rocessing. — *Примеч. пер.*

приложения работают по принципу извлечения данных из хорошо известных систем, таких как стандартные ERP или модели данных для представления в витринах, и переработки их в предустановленные показатели и форматы для вывода в отчеты или на информационные панели. По сути, бизнесу предлагаются готовые решения для оптимизации различных функциональных областей (например, управления персоналом) или встраивания в отраслевую вертикаль (например, аналитика розничного рынка). Приложения различных типов могут включать функции анализа клиентов, финансов, цепочек поставок, организации производства, управления персоналом и т. д. и т. п.

3.3.3.1 МНОГОМЕРНЫЙ АНАЛИЗ — OLAP

Термин *онлайновая аналитическая обработка данных* (OLAP) используется для описания подхода, обеспечивающего высокопроизводительную обработку многомерных аналитических запросов. Термин *OLAP* возник отчасти в противовес термину *OLTP*[1], использующемуся для обозначения онлайновой обработки транзакций. Обычно выдача данных в ответ на запросы OLAP происходит в матричном формате. Измерения определяются столбцами и строками матрицы, на пересечении которых выводятся факторы или значения. Концептуально это представление иллюстрируется как куб данных. Многомерный анализ с кубами особенно полезен там, где у аналитиков имеется хорошее представление об общей картине и структуре данных, а разобраться хочется с динамикой и сводной статистикой.

Традиционная область применения OLAP — финансовый анализ, ведь специалисты в этой области привыкли сновать вверх-вниз и туда-сюда по сводным таблицам данных, упорядоченных в рамках хорошо известных иерархий, выискивая и анализируя тенденции и закономерности; а кубы данных позволяют с легкостью переходить на иную шкалу измерений или масштабов даты/времени (годичные, квартальные, месячные, недельные, суточные, почасовые показатели и т. д.), организационной структуры (мир, регион, страна, отрасль, компания, подразделение и т. д.), иерархии продуктов (категория, линия, наименование продукта) и т. п. Многие пакеты программного обеспечения для BI сегодня используют OLAP-кубы в качестве одной из базовых моделей, а некоторые еще и поддерживают автоматизацию и бесшовную интеграцию процессов определения и заполнения кубов данных. А это означает, что любой пользователь может играться в такие «кубики», нарезая данные вдоль и поперек, как душе угодно. Поэтому постарайтесь открывать доступ к подобным функционалам лишь привилегированным категориям продвинутых пользователей из числа специалистов в своих предметных областях заодно с выделенным каналом самообслуживания, чтобы лишь немногие избранные имели возможность и свободу анализировать данные по своему усмотрению, но с пользой для организации.

Обычно прикладные OLAP-системы имеют серверный компонент и клиентское приложение для персонального компьютера или веб-интерфейс. Некоторые компоненты, устанавливаемые

[1] *сокр.* от *англ.* **On**line **T**ransaction **P**rocessing. — *Примеч. пер.*

на ПК, могут быть доступны из электронных таблиц — например, через надстройки или пункты панели инструментов или функционального меню. Функциональность и открывающиеся перед разработчиками возможности зависят от выбранной архитектуры модели OLAP (реляционной, многомерной или гибридной), но в любой из них используется определение данных через кубы, вывод суммарных или усредненных значений, наслоение метаданных и анализ разреженности данных.

Для структурирования куба в соответствии с желаемыми функциональными параметрами может потребоваться дробление крупных измерений на более детализированные с пропорциональным увеличением числа кубов, используемых для размещения, накопления или обсчета данных согласно требованиям аналитической модели. Выбирайте подходящие уровни агрегирования, которые будут обеспечивать обсчет формул и выдачу результатов за приемлемое время. Способность конечного пользователя выбирать слои иерархий в любом случае позволяет находить разумные компромиссы по параметрам обобщения/детализации, скорости расчета и/или плотности/разреженности данных. Кроме того, при явной скудости данных в интересующем кубе может потребоваться добавление или удаление составных структур или повышение уровня разрешения в реализации слоя — источника данных в хранилище.

Для поддержки дифференцированного доступа (например, по уровням ИБ в зависимости от роли) или мультиязычного текста в рамках одного и того же куба данных могут потребоваться дополнительные измерения, функции или расчеты, так что иногда легче попросту разделять такой куб на несколько кубов. Предполагается, что поиск баланса между гибкостью с точки зрения конечного пользователя, производительностью и рабочими нагрузками на сервер в каждом случае будет найден экспериментальным путем и, конечно же, не без переговоров и компромиссов. Согласование подобных вопросов обычно происходит при загрузке данных в хранилища, поскольку достигнутые договоренности могут потребовать изменения иерархий и структуры обобщенных данных или даже создания новых объектов в физической модели данных, реализованной в хранилище. Важно грамотно сбалансировать число кубов, рабочую нагрузку и гибкость выдач, чтобы данные обновлялись достаточно оперативно, кубы при этом содержали достоверную информацию, а запросы обрабатывались без сбоев и не создавали запредельной нагрузки на сервер.

Ценность средств онлайновой аналитической обработки и OLAP-кубов — в минимизации риска неверной интерпретации данных, поскольку они изначально выстраиваются в кубы в соответствии с замыслом аналитика, то есть физическая модель является простой проекцией логической. Аналитик имеет возможность для навигации по базе данных в поисках интересующего его конкретного подмножества данных, фильтрации и отсева лишних данных, пока в кубическом представлении не останутся лишь интересующая его выборка, перегруппировки и сортировки данных, а также определения формул, используемых для расчета аналитических показателей. Продольно-поперечная нарезка данных (на слои и кубики) инициируется пользователем в процессе навигации по интерактивно вызываемым страницам посредством разбиения, вращения и масштабирования. Стандартные операции OLAP включают следующее.

- **Slice — срез**: из многомерного массива (n-мерного куба) данных выделяется подмножество элементов с указанным значением по одному из измерений; в результате получается куб с n – 1 измерениями. Например, из трехмерного куба выделяется двумерный слой.
- **Dice — кубик**: получается путем урезания куба данных по двум и более измерениям.
- **Drill down/up — переходы вниз/вверх по уровням детализации**: позволяет аналитику углубляться в детали данных по любому измерению, начиная с верхнего, наиболее обобщенного уровня и вплоть до самого детализированного (низового) уровня.
- **Roll-up — свертка**: рассчитываются все определенные для одного или нескольких измерений показатели или соотношения и отображаются вместо детализированных данных. Операция свертки возможна лишь в OLAP-модели, допускающей определение отношений или формул расчета неких обобщенных показателей (среднего, итога и т. п.).
- **Pivot — вращение**: позволяет изменять пространственную ориентацию измерений в отчете или на странице/экране.

Классическими считаются три следующие архитектуры систем онлайновой аналитической обработки данных.

- **Реляционная (ROLAP)**: функционалы OLAP реализуются посредством моделирования многомерности через определение связей между атрибутами стандартных двумерных таблиц систем управления реляционными базами данных. Стандартная схема модели данных в среде ROLAP — звездообразная.
- **Многомерная (MOLAP)**: поддержка OLAP в составе или с использованием коммерческих и специализированных многомерных баз данных.
- **Гибридная (HOLAP)**: сочетание ROLAP и MOLAP. Гибридные реализации позволяют хранить часть данных в MOLAP, а часть — в ROLAP. Реализации могут варьироваться в зависимости от имеющихся у проектировщика возможностей по контролю структуры разделов данных.

4. МЕТОДЫ

4.1 Прототипирование с целью уточнения требований

Прежде чем приступать к внедрению чего бы то ни было, нужно оперативно определить приоритетные требования, а для этого создать демонстрационный набор данных и на нем отработать различные прототипы модели DW/BI совместно с будущими бизнес-пользователями. Достигнутый в последние годы прогресс в сфере технологий виртуализации данных существенно облегчает задачи прототипирования.

Профилирование данных — неотъемлемый компонент прототипирования, тем более что оно снижает риск столкновения с непредсказуемыми по структуре или значениям данными на последующих этапах. На этапе загрузки данных из систем-источников в DW чаще всего выявляются

проблемы с качеством как самих вводных данных, так и функций их обработки. Профилирование также помогает ранней диагностике разночтений или расхождений в структуре данных в системах-источниках, которые могут создать серьезные препятствия на пути их интеграции. В каждой системе-источнике данные могут быть вполне высококачественными согласно внутрисистемным критериям, но вот различия в наборах критериев качества, используемых в разных источниках, способны серьезно осложнить процесс интеграции данных.

Предварительная экспертиза состояния исходных данных способствует более точной оценке экономической и технологической целесообразности их интеграции с точки зрения окупаемости затрат и усилий на приведение данных из первоисточника в соответствие с предъявляемыми требованиями. Такая экспертиза важна и для определения реалистичных ожиданий. Планируйте также сотрудничество с командами, отвечающими за обеспечение качества данных и руководство данными, равно как и со специалистами во всех предметных областях по вопросам уточнения определений, устранения рассогласований и минимизации рисков (см. главу 13).

4.2 BI по принципу самообслуживания

Важнейшим принципом организации портфеля BI-приложений является самообслуживание (self-service) в части настроек представлений и выдач данных. Доступные пользователю действия обычно регулируются настройками профиля на портале доступа, где, в зависимости от привилегий, можно выбирать, подключать/отключать и конфигурировать различные функциональности, уведомления, сообщения и предупреждения, периодичность просмотра производственных отчетов, порядок взаимодействия с аналитическими отчетами, разрабатывать собственные отчеты и пользоваться настройками и функциями приборной панели и картами показателей. Отчеты могут выдаваться на портал по стандартному расписанию, чтобы пользователи могли ознакомиться с ними, когда это необходимо. Или же пользователи могут получать отчеты из хранилища с помощью запросов с портала. Наконец, порталы BI позволяют налаживать и обмен контентом между различными организациями.

Открытие доступа к средству совместной работы лицам, не входящим в узкий круг привилегированных пользователей, бывает полезным с точки зрения тонкой дифференциации настроек самообслуживания и публикации сводок о статусе загрузки, общих рабочих показателей, уведомлений о готовящихся обновлениях, а также для ведения различных форумов. Согласуйте темы форумов по техническим каналам, модерируйте их контент, а по мере надобности назначайте и проводите групповые сеансы для заинтересованных пользователей через служебный канал.

Средства визуализации и статистического анализа позволяют быстро разведывать и извлекать нужные данные. Иногда в пакете BI имеются и средства самостоятельного модульного построения пользовательских приборных панелей, ориентированных на бизнес. Их можно оперативно публиковать, распространять, рецензировать, совместно дорабатывать и улучшать. Всё то, что некогда было всецело во власти айтишников и программистов: моделирование, структурирование и формирование выборок данных, расчетные формулы, средства объектного построения и визуализации данных, — в наши дни становится доступным бизнес-сообществу. Это, в свою

очередь, открывает неплохую перспективу для перекладывания части нагрузки по прототипированию интеграции данных на бизнес-пользователей из числа участников форумов с последующей доработкой, оптимизацией и внедрением предложенных ими моделей специалистами по ИТ.

4.3 Открытые для пользователей данные аудита

Для обеспечения преемственности и возможности установления происхождения данных все компоненты и процессы, имеющие функциональную возможность удаления, добавления, модификации или перезаписи данных, должны вести и сохранять детализированные контрольные журналы, которые могут потребоваться для отслеживания изменений и отчетности. Открытие пользователям доступа к таким контрольным данным на уровне просмотра позволяет им самостоятельно убеждаться в том, что используемые ими данные находятся в контролируемом и актуальном состоянии, а это способствует повышению уровня доверия пользователей к данным и системе в целом. Информация из контрольных журналов может также использоваться для поиска первоисточников проблем с данными в случае их возникновения.

5. РЕКОМЕНДАЦИИ ПО ВНЕДРЕНИЮ

Стабильная архитектура с возможностью ее масштабирования по мере роста потребностей — первое непременное условие успешного проекта хранилища данных. Группа эксплуатационного сопровождения и технической поддержки, способная грамотно справляться с ежедневной загрузкой и анализом данных, обеспечением работоспособности всех систем и обработкой заявок пользователей, — второе непременное условие. В дополнение к первым двум условиям для обеспечения устойчиво успешной работы комплекса DW/BI требуется согласованность действий и интересов технических и бизнес-подразделений.

5.1 Оценка готовности / Оценка рисков

Ухватиться за идею нового начинания мало — организации требуется какое-то время для приведения себя в готовность к реализации и устойчивому развитию успеха предприятия. Любой успешный проект начинается с составления перечня необходимых условий. В случае ИТ-проекта в этом перечне, помимо само собой разумеющихся условий поддержки руководства, заинтересованности бизнеса и согласования стратегии, должно фигурировать определение архитектурного подхода, а применительно к проекту DW/BI — также и следующие условия:

- требования ИБ и защиты чувствительных данных и сопутствующие ограничения;
- обоснование выбора инструментов;
- надежные и безопасные источники исходных данных;
- проект процесса оценки и переработки данных из первоисточников.

Выявите элементы данных в хранилище, содержащие чувствительную, конфиденциальную или закрытую информацию, и составьте их опись с присвоением соответствующих грифов. Эти данные необходимо маскировать или скрывать при выдаче информационных продуктов пользователям, не имеющим права доступа к закрытым данным. Также могут потребоваться дополнительные ограничения на доступ к служебным данным, если планируется привлечение сторонних специалистов на стадии внедрения или для технического обслуживания систем.

Все ограничения и требования по обеспечению ИБ и защиты данных должны быть проработаны и учтены до начала выбора программных средств и распределения ресурсов. Также следите за неукоснительным соблюдением всех внутренних правил и процедур распоряжения данными, согласования и утверждения решений. Проекты DW/BI неизбежно сталкиваются с риском перепрофилирования, а иногда и полной их отмены из-за хитросплетений вышеперечисленных факторов.

5.2 Дорожная карта выпуска релизов

Из-за масштабности и трудоемкости подобных проектов хранилища данных строятся поэтапно, методом последовательных приращений. Вне зависимости от выбора конкретной методологии (каскадное, итерационное или гибкое внедрение) проект должен изначально содержать описание желаемого конечного состояния систем DW/BI. Именно поэтому ценнейшим инструментом планирования таких проектов является дорожная карта. В сочетании с плановыми процессами технического обслуживания дорожная карта позволяет гибко и оперативно балансировать не терпящие отлагательства шаги по реализации текущих этапов проекта со стратегическими целями обеспечения многоцелевого использования данных и развития информационно-технологической инфраструктуры.

В рамках поэтапного развертывания систем DW/BI настоятельно рекомендуется по максимуму задействовать матрицу шины DW в коммуникациях и маркетинге. Используйте диктуемые бизнесом приоритеты вместе с оценками рисков для определения степени строгости и запаса прочности, требующихся на каждом витке приращений. К примеру, при добавлении к DW малозначимого потока исходных данных из единственного нового источника вполне можно позволить себе некоторые послабления по части принципиальности соблюдения правил, особенно когда их нарушение, даже если оно будет выявлено, никакими серьезными негативными последствиями для организации не чревато.

Каждое приращение будет приводить к модификации существующих или добавлению новых функционалов; во втором случае речь чаще всего идет о подключении к комплексу DW/BI нового бизнес-подразделения. Для выбора следующего на очереди бизнес-подразделения для «посадки на борт» используйте согласованный и всякий раз один и тот же набор критериев оценки нужд и функциональных возможностей. Ведите учетный перечень недопоставок или недоработок и уделяйте первоочередное внимание их устранению, поскольку именно неисправленные недоделки хуже всего сказываются на функциональных возможностях бизнеса. Определяйте технические зависимости для правильного определения очередности поставок. Определив состав

и очередность исправлений и дополнений, скомпонуйте их в пакет обновлений или включите в новую установочную версию ПО. График выпуска обновлений или новых версий должен быть согласован с руководством бизнеса. В зависимости от специфики организации и систем новые версии могут выпускаться раз в квартал, ежемесячно или еженедельно, а пакеты исправлений (при необходимости) и чаще. Управление версиями осуществляйте совместно с бизнес-партнерами, отмечая их как вехи на дорожной карте, а также в хронологическом перечне выпущенных версий с указанием дат и добавленных или измененных функционалов.

5.3 Управление конфигурациями

Управление конфигурациями комплексов DW/BI осуществляется в привязке к дорожной карте выпуска релизов и обеспечивает предоставление необходимых материалов и скриптов, позволяющих в значительной мере автоматизировать разработку, тестирование и перенос процессов и приложений в среду эксплуатации. Через управление конфигурациями также производится автоматическое отслеживание и контроль номеров версий на уровне баз данных и исходных кодов, с тем чтобы обеспечить их согласованность.

5.4 Организационные и культурные изменения

С начала и до конца жизненного цикла разработки комплекса DW/BI необходимо последовательно фокусироваться на потребностях бизнеса. Внимательно изучайте цепочки создания стоимости своего предприятия, — это более чем полезно для понимания бизнес-контекста. Вычленение конкретных бизнес-процессов в цепочке создания стоимости компании способствует естественному определению понятных для бизнес-пользователей контекстных рамок областей анализа.

Не менее важно обеспечить согласованность между собой всех чисто технических проектов, которые остаются вне поля зрения бизнес-пользователей. Без этого невозможны оценка нужд бизнеса в технической поддержке и последующее их полноценное удовлетворение. При этом обязательно должны приниматься во внимание следующие ключевые факторы успеха.

- **Поддержка руководства.** Кто именно курирует и спонсирует проектирование и внедрение комплекса DW/BI на уровне высшего руководства вашей бизнес-структуры? Имеется ли там координирующий совет или орган, заинтересованный в проекте и гарантирующий его адекватное финансирование? Проекты DW/BI ресурсоемки и без надежного покровительства не реализуемы.
- **Бизнес-цели и задачи** должны быть четко определены и согласованы с направлениями и содержанием работ по проектам DW/BI.
- **Бизнес-ресурсы.** Обещано ли руководством бизнес-подразделений предоставление вам заинтересованных в реализации проекта экспертов в предметных областях? Отсутствие твердых обязательств экспертной поддержки — частая причина неудач и веское основание для того, чтобы отложить проект DW/BI до лучших времен.

- **Готовность бизнеса.** Согласны ли ваши бизнес-партнеры на долгосрочную поэтапную поставку системных и программных решений? Обещают ли создать у себя центры совершенствования для обеспечения устойчивой поддержки будущих новых версий продукта? Насколько расплывчато у целевого сообщества усредненное представление о DW/BI и насколько велики пробелы в знаниях и навыках? Можно ли рассчитывать на возможность преодоления путем разового внедрения продукта, или лучше начать с малых приращений?
- **Согласование ви́дения.** Укладывается ли стратегия развития ИТ в ви́дение бизнеса? Жизненно важно обеспечить концептуальное согласование функциональных требований к комплексу DW/BI с бизнес-функциями, которые поддерживаются или могут им поддерживаться с первых же шагов реализации дорожной карты ИТ. Любые значительные функциональные отклонения или ощутимые расхождения между функционалом ПО и бизнес-функциями чреваты свертыванием и даже ликвидацией программы DW/BI.

5.4.1 Выделенная команда

Во многих крупных организациях создают отдельную команду для осуществления текущей деятельности по поддержке среды эксплуатации (см. главу 6). Поручить команде специалистов оперативное управление продуктами DW/BI бывает полезно хотя бы для оптимизации распределения рабочей нагрузки на персонал, поскольку в свободное от выполнения регламентных работ, предусмотренных графиком, время этих же людей можно задействовать в качестве консультантов службы технической поддержки, которая часто оказывается перегружена обращениями как раз таки по завершении поставок дополнительных функциональных модулей или новых версий.

Администраторы команды сопровождения обеспечивают взаимодействие с командой эксплуатации ИТ-инфраструктуры с целью выстраивания конструктивных рабочих отношений и обеспечения проведения всех необходимых подготовительных работ перед выпуском очередных релизов, а также уведомляет инженеров о выявленных недостатках, которые желательно оперативно устранить. Технологи группы сопровождения, отвечающие за эксплуатацию комплекса DW/BI, в свою очередь отвечают за своевременность и правильность необходимых изменений в конфигурации, при необходимости докладывают по инстанциям о тревожных сигналах или симптомах, а также ведут мониторинг и учет пропускной способности и прочих эксплуатационных параметров систем.

6. РУКОВОДСТВО DW/BI

В сильно зарегулированных отраслях со строгой дисциплинарной ответственностью и отчетностью полезно иметь хранилище данных, в отношении которого осуществляется надежное руководство. Критичным для эксплуатационной поддержки систем и планирования выпуска новых релизов и обновлений является обеспечение выполнения всех формальностей, касающихся руководства данными, на стадии реализации. Всё больше организаций расширяют жизненный

цикл разработки ПО за счет включения в него процедур, связанных с обеспечением руководства данными. В любом случае процессы руководства хранилищем данных должны быть согласованы с процессами управления рисками. При этом они должны быть ориентированы на практические результаты, а также учитывать специфику отрасли, которая и определяет бизнес-потребности (например, в маркетинговых и рекламных компаниях данные используются совершенно иначе, нежели в кредитно-финансовых учреждениях). Процессы руководства призваны минимизировать риск, а не функциональные возможности.

Самыми критичными из функций руководства в отношении хранилищ данных являются те, от которых зависит своевременное предоставление обязательной отчетности в надзорные органы или устранение выявленных нарушений, а также безупречное качество данных и безукоризненный порядок в самом хранилище. Поскольку требования повышения качества данных ставятся во главу угла всех инициатив в области DW/BI, использование квитирования во всех каналах связи и отлаженных процедур приема-передачи данных также являются обязательными для создания экземпляров, обработки, отправки и удаления данных в подобных средах. Соблюдение требований архивирования данных и сроков хранения архивов является ключевым элементом рамочных соглашений, поскольку помогает избежать неконтролируемого разрастания архивов. Правила мониторинга этих сред и планы-графики их очистки от неактуальных данных должны включаться в программу обязательных учебных занятий для пользователей, а не только рабочих совещаний администрации. Загрузка данных в хранилище требует выделения достаточного времени, ресурсов и усилий на обеспечение согласованности, достоверности и качества данных, поступающих в распоряжение сообщества конечных пользователей, но не в ущерб своевременности их обновления.

Разовые, редкие или ограниченные по масштабу инциденты также следует учитывать при планировании жизненного цикла и, как вариант, не выпускать их за пределы опытно-экспериментальной среды или пользовательской «песочницы». Процессы анализа в режиме реального времени можно использовать для получения вводных данных и накопления статистики, если автоматизировать отправку однотипных результатов с метками времени обратно в хранилище данных. Для этого должна быть определена соответствующая политика и процедуры передачи результатов из среды интерактивной обработки в хранилище для общего пользования, а также меры административного контроля соблюдения этих требований.

Применяйте правила выявления данных, свидетельствующих о проблемах или рисках, через сопоставление входящих данных с каталогами или матрицами рисков. Выявленные таким образом элементы классифицируются как указывающие на потенциальную опасность, недостаточную защищенность или затруднительность ранней диагностики, а потому подлежат передаче на рассмотрение администрации на предмет оценки риска и принятия мер. При высокой чувствительности данных, свидетельствующих о потенциальной опасности, бывает целесообразно предусмотреть выделенное рабочее пространство для их изучения уполномоченными лицами в автономном режиме. Тщательный анализ данных совместно с представителями службы безопасности или юридического отдела — последняя страховочная сетка на подобные случаи.

6.1 Обеспечение одобрения со стороны бизнеса

Ключевой фактор успеха — приемлемость данных для бизнеса, что подразумевает их понятность, подтверждаемое и проверяемое качество и происхождение. Письменное подтверждение согласия с составом и структурой данных должно оформляться отдельным пунктом акта приемки комплекса DW/BI по результатам пользовательского приемочного тестирования. Проводите тестирование случайной выборки структурированных данных на предмет их качества и совместимости с BI-средствами при первичной загрузке данных из любой новой системы-источника. Повторное тестирование следует провести после нескольких циклов обновлений, чтобы гарантировать их соответствие согласованным с бизнесом критериям. Соответствие этим критериям — первоочередное условие приемлемости комплекса DW/BI вне зависимости от его технической реализации. Поэтому изначально заложите в проект следующие важнейшие архитектурные субкомпоненты и соответствующие им направления деятельности.

- **Концептуальная модель данных.** Какие данные относятся к информационному ядру организации? Как формулируются и увязываются друг с другом ключевые концепции бизнеса?
- **Система контроля качества данных с контуром обратной связи.** Как выявляются и устраняются проблемы? По каким каналам уведомляются о выявленных дефектах владельцы систем — источников некондиционных данных? Перед кем отчитываются об устранении проблем? Какие процессы используются для выявления и устранения проблем, обусловленных недостатками в интеграционных процессах, реализованных в системе управления самим DW?
- **Исчерпывающие метаданные**. Каким образом в архитектуре DW реализуется поддержка полной интеграции метаданных со всеми процессами и потоками данных? В частности, предусмотрен ли архитектурой доступ к определениям, смысловым и контекстным описаниям, соответствующим различным объектам и элементам данных? Каким именно образом потребители данных получают ответы на базовые вопросы следующего типа: «В чем смысл этого отчета?» или «Что именно характеризует этот показатель?».
- **Исчерпывающая и верифицируемая генеалогия всех данных.** Все ли выдаваемые бизнес-пользователям элементы данных отслеживаются до первоисточника? Ведется ли автоматизированный и последовательный учет происхождения и хронологии изменений всех данных? Все ли данные охвачены системой учета записей?

6.2 Удовлетворенность клиентов/пользователей

Субъективное восприятие качества данных напрямую влияет на степень удовлетворенности потребителей данных, предлагаемых комплексом DW/BI. Однако важно помнить, что на показатели удовлетворенности серьезное влияние оказывают и другие факторы, в частности понимание потребителями смысла данных, оперативность выявления и устранения проблем, отзывчивость и компетентность службы поддержки. Сбор и анализ отзывов клиентов с целью принятия оперативных мер по устранению недостатков может быть реализован как через онлайн-формы обратной связи, так и в формате регулярных встреч и обсуждений с представителями различных

групп пользователей. Подобное взаимодействие также помогает администрации и специалистам DW/BI своевременно информировать потребителей данных о готовящихся новых выпусках, а самим составлять более полное и точное представление, кем, как и для чего используются данные и инструментальная оснастка DW/BI.

6.3 Соглашения об уровне обслуживания

В случае удаленной или распределенной реализации проекта DW/BI бизнес-требования и технические параметры различных сред определяются соглашениями об уровне обслуживания (SLA). Часто такие показатели, как время отклика, сроки хранения данных и уровни доступности данных, в различных отраслях и видах деятельности отличаются на много порядков, и эти специфические особенности оказывают безусловное влияние на выбор систем-компонентов (ODS/DW/витрина), предлагаемых заказчику.

6.4 Стратегия в области отчетности

Обеспечьте наличие стратегии в области отчетности по каждому направлению BI и по всему портфелю BI-решений в целом. Стратегия в области отчетности включает стандарты, процессы, руководства, рекомендации и процедуры, призванные обеспечить наличие у всех пользователей ясной, четкой и актуальной информации, в частности по следующему кругу вопросов:

- средства обеспечения ИБ и защиты данных, гарантирующие только санкционированный доступ авторизованных пользователей к чувствительным элементам данных;
- механизмы доступа пользователей к данным и порядка взаимодействия с ними через системные интерфейсы с целью обработки, формирования отчетов, изучения и просмотра данных;
- тип сообщества пользователей и надлежащие инструменты его интеграции;
- природа сводных и детализированных отчетов, порядок, периодичность или график их подготовки, проверки, распространения или публикации и хранения, включая форматы;
- доступные средства и функции визуализации данных для выдачи графических представлений;
- компромиссы между скоростью/производительностью и разрешением/детализацией.

Стандартные отчеты подлежат периодической переоценке на предмет сохранения ими ценности хотя бы по той причине, что их составление само по себе затратно, так как требует системных ресурсов и места в хранилище. В особом внимании нуждается проработка процессов внедрения, эксплуатации и управления. Согласование инструментов отчетности с потребностями бизнес-сообщества — критический фактор успеха. В зависимости от размера и природы организации могут использоваться самые разнообразные инструменты отчетности, которые отражали бы характеристики не менее разноплановых рабочих процессов. Главное, чтобы средства формирования отчетности обеспечивали ее понятность и полезность с точки зрения целевых аудиторий; чем образованнее и искушеннее пользователи, тем они требовательнее к степени детализации и проработки отчетов. Придерживайтесь матрицы выбора решений на основе анализа актуальных

запросов с целью их своевременного удовлетворения за счет обновления структуры запросов и планирования обновления и усовершенствования будущего инструментария отчетности.

Жизненно необходимым требованием является административный мониторинг и контроль источников данных. Следует обеспечить надежные средства ограничения доступа к данным различных уровней защищенности кругом лиц, имеющих право доступа к ним, а также надлежащие правила управления подписками на рассылки в соответствии с установленными уровнями допусков.

Можно создать Центр компетенций, отвечающий за профессиональную подготовку, разработку пакетов материалов для начинающих, методологических рекомендаций, практических советов и подсказок по работе с различными источниками данных и другие точечные решения или материалы в помощь бизнес-пользователям на пути перехода к модели самообслуживания. Помимо управления знаниями, этот центр может использоваться для своевременного обмена всей необходимой информацией между разработчиками, проектировщиками, аналитиками и сообществами пользователей и/или подписчиков.

6.5 Метрики

6.5.1 Показатели использования

Основными показателями использования и востребованности комплекса DW/BI обычно являются число зарегистрированных пользователей, а также число подключенных или активных пользователей за отчетный период или в среднем за сутки / рабочий день. Подобные метрики позволяют объективно оценивать число/процент сотрудников различных подразделений организации, активно пользующихся DW/BI. Ознакомиться с числом подписчиков на каждую рассылку и/или лицензированных пользователей каждого инструмента BI бывает для начала вполне достаточно, особенно для проверяющих. Однако с точки зрения администрирования и особенно планирования развития мощностей значительно важнее знать число фактических подключений к каждому инструментальному средству и запросов (или эквивалентных им обращений) к данным из различных источников, желательно с профилированием этой статистики по сообществам пользователей. Поэтому предусмотрите раздельный учет и анализ структуры потребления данных пользователями различных категорий: например, аудиторами, аналитиками, использующими функционалы самостоятельного формирования сложных запросов, и рядовыми пользователями — потребителями сводных отчетов.

6.5.2 Доли востребованных данных по предметным областям

Процентные показатели востребованности данных по различным предметным областям позволяют выявить, какие участки хранилища (с точки зрения топологии данных) пользуются спросом у каждого подразделения. Кроме того, они дают возможность понять, какие данные используются совместно разными подразделениями, а какие нет, хотя и могли бы.

Мэппинг источников данных в целевые структуры — естественное дополнение предыдущей группы показателей. К тому же он позволяет проверять соблюдение требований отслеживания

происхождения данных и полноты метаданных, описывающих уже имеющиеся данные, а также проводить углубленный анализ востребованности различных систем-источников различными подразделениями. Последнее полезно и с точки зрения планирования оптимизации аналитических средств, поскольку и без того массово востребованные источники и задействованные инструменты от каких-либо серьезных изменений лучше оградить.

6.5.3 Показатели времени ответа и производительности

Большинство программных средств фиксируют время отправки запроса и получения ответа. По этим данным и рассчитываются усредненные показатели времени отклика или производительности. Данные из журналов приложений также можно использовать для учета числа пользователей различных категорий.

Собирайте показатели времени загрузки данных для каждого информационного продукта в «сыром» формате, непосредственно в ходе процессов заполнения. Эти показатели также должны отображаться и в процентном выражении от ожидаемой нормы: например, если технический регламент предусматривает обновление данных в витрине раз в сутки во время четырехчасового технологического перерыва, то 100-процентным соблюдением считается лишь полное обновление данных к моменту возобновления доступа пользователей по истечении этих четырех часов. Применяйте такую же процедуру оценки к любым процессам генерирования выборок на любых участках технологического потока.

Большинство инструментов сохраняют (в журнале или репозитории) строки запросов, а также значения времени отправки и времени получения. Распределите эти запросы по категориям «запланированные» и «выполненные» — и ведите простейший учет процента успешно обработанных обращений. В случае высокой доли задержек при обращении к популярным объектам соответствующие алгоритмы и процедуры реализации, очевидно, требуют доработки, и приступать к ней нужно незамедлительно, пока проблема не успела сказаться на показателях удовлетворенности потребителей. Исходя из этого, планируйте все необходимые меры по выявлению и анализу дефектов в рамках регламентных работ, а также выделению дополнительных ресурсов, если регулярные отказы затрагивают целую группу объектов. Что касается средств исправления, они могут серьезно варьироваться в зависимости от типа систем и программных средств, но случаются и приятные неожиданности, когда все проблемы устраняются единственным элементарным исправлением — например, добавлением пропущенного или удалением лишнего индекса (см. главу 6).

Также при выявлении подобных проблем нужно проверить и в случае необходимости скорректировать параметры уровней обслуживания. Если оперативно устранить регулярно возникающую проблему к следующему выпуску пакета исправлений или новой версии нереалистично по техническим или финансовым причинам, заявленный уровень поддержки соответствующих функций должен быть снижен до фактического.

7. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Adamson, Christopher. *Mastering Data Warehouse Aggregates: Solutions for Star Schema Performance*. John Wiley and Sons, 2006. Print.

Adelman, Sid and Larissa T. Moss. *Data Warehouse Project Management*. Addison-Wesley Professional, 2000. Print.

Adelman, Sid, Larissa Moss and Majid Abai. *Data Strategy*. Addison-Wesley Professional, 2005. Print.

Adelman, Sid, et al. *Impossible Data Warehouse Situations: Solutions from the Experts*. Addison-Wesley, 2002. Print.

Aggarwal, Charu. *Data Mining: The Textbook*. Springer, 2015. Print.

Biere, Mike. *Business Intelligence for the Enterprise*. IBM Press, 2003. Print.

Biere, Mike. *The New Era of Enterprise Business Intelligence: Using Analytics to Achieve a Global Competitive Advantage*. IBM Press, 2010. Print. IBM Press.

Brown, Meta S. *Data Mining for Dummies*. For Dummies, 2014. Print. For Dummies.

Chorianopoulos, Antonios. *Effective CRM using Predictive Analytics*. Wiley, 2016. Print.

Delmater, Rhonda and Monte Hancock Jr. *Data Mining Explained; A Manager's Guide to Customer-Centric Business Intelligence*. Digital Press, 2001. Print.

Dyche, Jill. E-Data: *Turning Data Into Information With Data Warehousing*. Addison- Wesley, 2000. Print.

Eckerson, Wayne W. *Performance Dashboards: Measuring, Monitoring, and Managing Your Business*. Wiley, 2005. Print.

Han, Jiawei, Micheline Kamber and Jian Pei. *Data Mining: Concepts and Techniques*. 3rd ed. Morgan Kaufmann, 2011. Print. The Morgan Kaufmann Ser in Data Management Systems.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer, 2011. Print. Springer Series in Statistics.

Hill, Thomas, and Paul Lewicki. *Statistics: Methods and Applications*. Statsoft, Inc., 2005. Print.

Howson, Cindi. *Successful Business Intelligence: Unlock the Value of BI and Big Data*. 2nd ed. McGraw-Hill Osborne Media, 2013. Print.

Imhoff, Claudia, Lisa Loftis, and Jonathan G. Geiger. *Building the Customer-Centric Enterprise: Data Warehousing Techniques for Supporting Customer Relationship Management*. John Wiley and Sons, 2001. Print.

Imhoff, Claudia, Nicholas Galemmo, and Jonathan G. Geiger. *Mastering Data Warehouse Design: Relational and Dimensional Techniques*. John Wiley and Sons, 2003. Print.

Inmon, W. H., Claudia Imhoff, and Ryan Sousa. *The Corporate Information Factory*. 2nd ed. John Wiley and Sons, 2000. Print.

Inmon, W. H., and Krish Krishnan. *Building the Unstructured Data Warehouse*. Technics Publications, LLC, 2011. Print.

Josey, Andrew. *TOGAF Version 9.1 Enterprise Edition: An Introduction*. The Open Group, 2011. Kindle. Open Group White Paper.

Kaplan, Robert S and David P. Norton. *The Balanced Scorecard: Translating Strategy into Action*. Harvard Business Review Press, 1996. Kindle.

Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3rd ed. Wiley, 2013. Print.

Kimball, Ralph, et al. *The Data Warehouse Lifecycle Toolkit*. 2nd ed. Wiley, 2008. Print.

Kimball, Ralph. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Amazon Digital Services, Inc., 2007. Kindle.

Linoff, Gordon S. and Michael J. A. Berry. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. 3rd ed. Wiley, 2011. Print.

Linstedt, Dan. *The Official Data Vault Standards Document (Version 1.0) (Data Warehouse Architecture)*. Amazon Digital Services, Inc., 2012. Kindle.

Loukides, Mike. *What Is Data Science?* O'Reilly Media, 2012. Kindle.

Lublinsky, Boris, Kevin T. Smith, and Alexey Yakubovich. *Professional Hadoop Solutions*. Wrox, 2013. Print.

Malik, Shadan. *Enterprise Dashboards: Design and Best Practices for IT*. Wiley, 2005. Print.

Morris, Henry. «Analytic Applications and Business Performance Management». *DM Review Magazine*, March, 1999, http://bit.ly/2rRrP4x

Moss, Larissa T., and Shaku Atre. *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications*. Addison-Wesley Professional, 2003. Print.

Ponniah, Paulraj. *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals*. Wiley-Interscience, 2001. Print.

Provost, Foster and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, 2013. Print.

Reeves, Laura L. *A Manager's Guide to Data Warehousing*. Wiley, 2009. Print.

Russell, Matthew A. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. 2nd ed. O'Reilly Media, 2013. Print.

Silverston, Len, and Paul Agnew. *The Data Model Resource Book Volume 3: Universal Patterns for Data Modeling*. Wiley, 2008. Print.

Simon, Alan. *Modern Enterprise Business Intelligence and Data Management: A Roadmap for IT Directors, Managers, and Architects*. Morgan Kaufmann, 2014. Print.

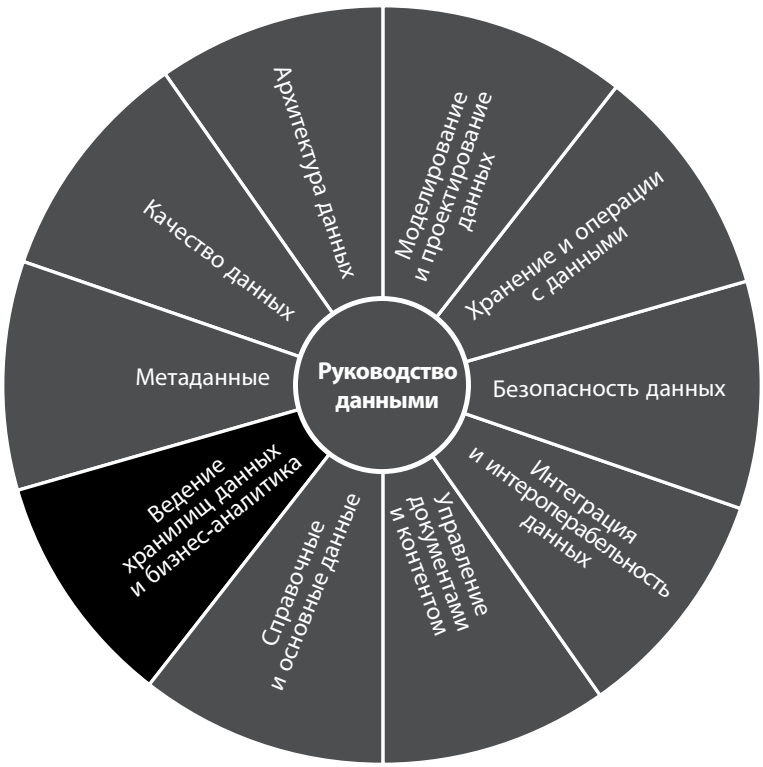
Thomsen, Erik. *OLAP Solutions: Building Multidimensional Information Systems*. 2nd ed. Wiley, 2002. Print.

Vitt, Elizabeth, Michael Luckevich and Stacia Misner. *Business Intelligence*. Microsoft Press, 2008. Print. Developer Reference.

WAGmob. *Big Data and Hadoop*. WAGmob, 2013. Kindle.

Wremble, Robert and Christian Koncilia. *Data Warehouses and Olap: Concepts, Architectures and Solutions*. IGI Global, 2006. Print.

Управление метаданными



DAMA-DMBOK2 Data Management Framework

1. ВВЕДЕНИЕ

Наиболее распространенное определение *метаданных* (*metadata*) — «данные о данных» — вводит в заблуждение своей простотой. В реальности к метаданным можно отнести очень широкий спектр сведений, включая информацию о технологических и бизнес-процессах, правила обработки данных, ограничения, определения логической и физической структуры данных и т. д.

Метаданные могут описывать не только данные как таковые (базы данных, элементы данных, модели данных и т. д.), но и представляемые ими объекты (бизнес-процессы, системы и приложения, элементы ИТ-инфраструктуры и т. п.), а также связи (отношения) между данными и объектами. Метаданные помогают организации правильно понимать смысл имеющихся в ее

распоряжении данных, функционирование систем, структуру и содержание рабочих процессов. Они позволяют проводить оценку качества данных и неразрывно связаны с управлением базами данных и другими приложениями. Обобщая вышесказанное: метаданные необходимы для обеспечения возможности обработки, сопровождения, интеграции, хранения, защиты, проверки и контроля всех прочих данных организации.

Для полноты понимания незаменимости метаданных в сфере управления данными представьте себе огромную библиотеку с миллионами книг и журналов на полках, но без картотеки. Читателю будет весьма проблематично найти не только интересующую его книгу, но и стеллаж с книгами соответствующей тематики. И совсем другое дело — библиотека с каталогизированной картотекой, не просто содержащей всю необходимую информацию о библиотечном фонде (какие книги и периодические издания имеются в наличии, в каких залах и на каких стеллажах они хранятся), но еще и позволяющей отыскивать нужные материалы по различным признакам или исходным данным (предметная область, автор, название и т. п.). Без каталога отыскать в огромной библиотеке конкретную книгу практически нереально. Таким образом, организация без метаданных уподобляется библиотеке без карточного каталога.

Для управления данными метаданные нужны не меньше, чем для их поиска и использования (свидетельством чему служат регулярные упоминания метаданных на протяжении всей этой книги, посвященной всестороннему описанию универсальной концепции управления данными DAMA-DMBOK). Все крупные организации производят и используют данные в огромных объемах. Внутри организации на разных уровнях и в различных подразделениях работает множество самых разных людей, и у каждого из них собственный набор представлений о данных, которыми располагает организация, — но никто не имеет и не может иметь исчерпывающего и достоверного представления о данных организации. Поэтому и требуется скрупулезный учет данных, а без ведения подобной документации организация рискует перестать понимать саму себя. А метаданные служат главным средством регистрации, формализации и упорядочения знаний о данных, имеющихся у организации.

Однако управление метаданными не сводится к одному лишь управлению знаниями о данных; управление метаданными — это еще и средство управления риском. Без метаданных невозможно обеспечить выявление и защиту конфиденциальной и чувствительной информации, управление жизненным циклом данных, а также соблюдение внутренних и внешних требований.

Без надежных метаданных организация не имеет представления ни о том, какими данными она располагает; ни о том, что эти данные отражают, откуда берутся, как перемещаются внутри систем и между системами; ни о том, кто имеет доступ к данным; ни об их качестве и средствах контроля качества. Без метаданных организация не сможет не только управлять данными как ценным ресурсом или активом, но и просто хоть как-то ими управлять.

С развитием технологий колоссально выросли темпы и объемы генерирования всевозможных данных, и технические метаданные сделались незаменимым средством управления передачей и интеграцией данных. Обмен данными в гетерогенных информационных средах на основе

Рисунок 84.
Контекстная диаграмма: метаданные

определений метаданных регламентируется стандартом ISO/IEC 11179 «Регистры метаданных»[1]. В XML и ряде других форматов документов без метаданных невозможно интерпретировать и использовать остальные данные. В других случаях метаданные используются для маркировки данных, предназначенных для обмена, сведениями об их принадлежности, авторстве, конфиденциальности и т. д. (см. главу 8).

Как и любые другие данные, метаданные нуждаются в управлении. С ростом способности организаций собирать и накапливать колоссальные массивы данных роль метаданных в сфере управления данными неуклонно возрастает. Чтобы быть «управляемой на основе данных» (data-driven) организация должна быть «управляемой на основе метаданных» (metadata-driven).

1.1 Бизнес-драйверы

Без метаданных управление остальными данными невозможно. Однако и сами метаданные требуют управления. Надежные и качественно управляемые метаданные обеспечивают:

- повышение доверия к данным за счет предоставления их контекста и поддержки возможности измерения качества данных;
- повышение ценности стратегической информации (в частности, основных данных) за счет ее многоцелевого использования;
- повышение эффективности работы информационных систем через выявление и устранение избыточных данных и процессов;
- своевременное выявление и отбраковку устаревших или неверных данных;
- оптимизацию планирования и проведения статистических исследований;
- лучшее взаимопонимание между потребителями данных и специалистами по ИТ;
- точность вводных данных, используемых для аналитического прогнозирования последствий, что способствует минимизации риска провала проектов;
- ускорение внедрения за счет сокращения времени, уходящего на разработку систем;
- снижение затрат на обучение и негативные последствия текучки кадров за счет исчерпывающей документации данных, включая контекст, источники и историю;
- выполнение требований действующего законодательства и надзорных органов.

Метаданные также способствуют согласованности и непротиворечивости данных и единообразному представлению информации, оптимизации потоков данных и рабочих процессов, надлежащей защите чувствительной информации, что особенно важно для отраслей с повышенными нормативно-правовыми требованиями.

Чем выше качество данных, тем выше их ценность для организации. Качество данных зависит от руководства данными. Метаданные играют критически важную роль в осуществлении руководства данными, поскольку без них невозможно понимание данных в контексте

[1] См.: ГОСТ Р ИСО/МЭК 11179. Информационная технология (ИТ). Регистры метаданных (РМД). — *Примеч. пер.*

функционирования организации. По сути, метаданные являются путеводителем по всем данным, имеющимся в распоряжении организации. Следовательно, управление метаданными должно быть безупречным. Плохо управляемые метаданные приводят к следующим негативным последствиям:

- появление избыточных данных и бессмысленных процессов управления ими;
- дублирующие друг друга избыточные, устаревшие или вовсе не используемые словари, репозитории и иные хранилища метаданных;
- противоречивые определения объектов и элементов данных;
- неверные и противоречивые оценки рисков, соответствующих различным категориям данных, в том числе проистекающих от их нецелевого использования или утечки;
- конфликтующие между собой источники и версии метаданных и, как следствие, подрыв доверия пользователей к любым определениям данных, используемых в организации.

Хорошо организованное управление метаданными обеспечивает полное и согласованное представление об информационных ресурсах организации и способствует эффективному налаживанию взаимодействия между организациями при проведении совместной разработки приложений.

1.2 Цели и принципы

Основные цели управления метаданными:

- управление задокументированными на уровне организации знаниями о данных в привязке к бизнес-терминологии с целью обеспечения единообразной трактовки данных всеми, кто их использует;
- сбор и интеграция метаданных из различных источников с целью обеспечения понимания пользователями сходств и различий между данными, поступающими из различных частей организации;
- обеспечение качества, согласованности, актуальности и защищенности метаданных;
- предоставление стандартных каналов доступа к метаданным всем потребителям данных (пользователям, системам, приложениям и процессам);
- выработка и утверждение собственных или контроль соблюдения предписываемых стандартов технических метаданных с целью обеспечения возможности обмена данными.

Руководящие принципы успешного внедрения решения по управлению метаданными таковы:

- **Приверженность со стороны организации.** Высшее руководство организации должно понимать необходимость адекватного управления метаданными для эффективного распоряжения корпоративными информационными активами, рассматривать его как часть стратегии работы с данными и выделять достаточные объемы финансирования.

- **Стратегия** работы с метаданными должна описывать общеорганизационный план создания, сопровождения, интеграции и использования метаданных. На основе стратегии вырабатываются требования, которые будут предъявляться к создаваемым ИТ-решениям или приобретаемым программным продуктам по управлению метаданными. Стратегия работы с метаданными должна соответствовать бизнес-приоритетам.
- **Корпоративная перспектива.** Осуществление управления метаданными в масштабах организации позволяет гарантировать возможность расширений и дополнений в будущем, но внедряться оно в любом случае должно поэтапно, чтобы обеспечить получение ощутимой финансовой отдачи за счет оптимизации после первых циклов внедрения.
- **Социализация.** Проводите разъяснительную работу, доказывайте и обосновывайте необходимость и назначение метаданных каждого типа; донесение до всеобщего понимания ценности метаданных послужит стимулом к их использованию бизнес-подразделениями и, что не менее важно, позволит привлечь к работе экспертов в различных предметных областях.
- **Доступность.** Не забывайте удостоверяться в том, что все сотрудники знают порядок доступа к метаданным и умеют их использовать.
- **Качество.** Важно понимать, что метаданные часто возникают в качестве побочного продукта различных процессов (моделирования данных, проектирования систем, определения бизнес-процессов и т. п.); необходимо сделать так, чтобы лица, отвечающие за эти процессы, несли персональную ответственность и за качество метаданных, порождаемых этими процессами.
- **Аудит.** Установите стандарты метаданных и строго контролируйте их соблюдение.
- **Совершенствование.** Создайте механизмы обратной связи и обрабатывайте поступающие от потребителей сигналы об ошибочных или устаревших метаданных.

1.3 Основные понятия и концепции

1.3.1 Метаданные как особая категория данных

Как уже было заявлено во введении к настоящей главе, метаданные — это разновидность данных, и, подобно любым другим данным, метаданные нуждаются в управлении. Некоторые организации сталкиваются с затруднениями как раз в вопросах проведения четкой границы между метаданными и всеми прочими данными. Теоретически эта граница соответствует переходу от абстракции к конкретике. Например, в докладе, предавшем огласке факты массового прослушивания телефонов граждан США Агентством национальной безопасности, данные о номерах телефонов и времени звонков как ни в чем не бывало называются «*метаданными*», и это как бы подразумевает, что к «*настоящим*» данным относится лишь содержание телефонных разговоров. Однако здравый смысл подсказывает, что номера телефонов граждан, хронология звонков и продолжительность соединений также должны классифицироваться как просто данные, без приставки «мета-»[1].

[1] Cole, David. «We kill people based on metadata». *New York Review of Books*. 10 May 2014, http://bit.ly/2sV1ulS

Возможно, «золотое правило» состоит в том, чтобы выделять пограничный слой данных, которые могут являться метаданными или информативными данными в различных контекстах, в частности в зависимости от степени осведомленности лица, получающего к ним доступ. Нечто, внешне выглядящее точь-в-точь как метаданные (например, список имен полей или названий столбцов), на поверку может оказаться фактическими данными, — если, например, этот список названий используется для сравнительного анализа терминов, используемых различными организациями, с целью разобраться с фактическим содержанием данных, имеющихся у каждой из них.

Впрочем, для управления собственными метаданными организации ни к чему вдаваться в тонкости. Лучше сосредоточить усилия на своевременном определении дополнительно требующихся метаданных для грядущих нужд в зависимости от того, для чего эти новые метаданные будут использоваться (для создания новых данных, углубленного анализа имеющихся, кроссплатформенной интеграции, управления доступом, публикации/распространения данных и т. п.), и источников данных, удовлетворяющих этим требованиям.

1.3.2 Виды метаданных

Обычно метаданные подразделяют на три основные категории:

- бизнес-метаданные;
- технические метаданные;
- операционные метаданные.

Так людям понятнее, насколько широкий спектр информации и явлений оказывается «под одним зонтом» метаданных, и проще разбираться с функциями определения метаданных. К слову, эта же троичная классификация способна привнести в умы и немало путаницы, особенно если люди залипают на вопросах типа «К какой именно категории относится этот набор метаданных?» или «Кто отвечает за ведение этого набора метаданных и кто имеет право вносить в него изменения?». Лучше решать подобные вопросы, исходя из места происхождения метаданных, а не места или порядка их использования. Кстати, в процессе использования метаданных грани между тремя описанными типами оказываются весьма размытыми и не столь уж и значимыми. Техническому персоналу приходится иметь дело и с «бизнес-метаданными», и с «операционными метаданными»; то же самое касается и двух других категорий пользователей.

Впрочем, за пределами информационных технологий могут использоваться другие классы метаданных. Например, в библиотечном деле стандартными категориями метаданных считаются:

- описательные метаданные (автор, заглавие, предмет и т. д.), которые помогают идентифицировать ресурсы и извлекать их из хранилищ;
- структурные метаданные, которые описывают связи между ресурсами и внутреннюю компоновку и строение ресурсов (число томов, частей, глав, страниц и т. п.).

- административные метаданные (номера версий, датировки архивных копий и т. п.), используемые для управления ресурсами на протяжении их жизненного цикла.

Главное назначение всех подобных категорий — обеспечивать полезной дополнительной информацией тех, кто отвечает за определение и проработку требований к метаданным.

1.3.2.1 БИЗНЕС-МЕТАДАННЫЕ

Бизнес-метаданные описывают, по большому счету, содержание и состояние данных, а также детали, необходимые для реализации функций распоряжения данными. К бизнес-метаданным относятся нетехнические наименования и определения понятий, названия и атрибуты предметных областей; типы данных и иные свойства атрибутов; описания диапазонов данных; расчетные формулы; алгоритмы и бизнес-правила; области допустимых значений данных и их определения. Примерами бизнес-метаданных могут служить:

- определения и описания наборов, таблиц и столбцов данных;
- бизнес-правила, правила преобразований, расчетные и логические формулы;
- модели данных;
- правила и результаты измерения показателей качества данных;
- расписания обновления данных;
- первоисточники и происхождение данных;
- стандарты данных;
- условные обозначения, используемые в системе записи и учета элементов данных;
- ограничения по допустимым значениям;
- контактная информация ответственных (например, владельцев или распорядителей данных);
- классы секретности/конфиденциальности данных;
- известные проблемы с данными;
- примечания по использованию данных.

1.3.2.2 ТЕХНИЧЕСКИЕ МЕТАДАННЫЕ

Технические метаданные детально описывают всевозможные технические характеристики данных, систем их хранения и процессов перемещения данных между системами. Примеры:

- названия таблицы и столбцов таблицы, используемые в физической модели данных;
- свойства столбца;
- свойства объекта БД;
- права доступа;
- правила создания, замены, обновления и удаления записей (create, replace, update and delete; CRUD);
- физические модели данных, включая имена таблиц данных, ключи и индексы;

- задокументированные связи между моделями данных и физическими ресурсами;
- детализация операций по извлечению/передаче/загрузке данных (ETL);
- определения схем данных в файловых форматах;
- карты соотнесения данных между системами-источниками и адресатами;
- документация, описывающая происхождение данных, включая влияние изменений на информацию выше и ниже по потоку обработки;
- названия и описания используемых программ и приложений;
- расписания заданий по загрузке/обновлению контента и зависимостей между ними;
- правила резервного копирования и восстановления данных из резервных копий;
- права доступа, группы и роли пользователей.

1.3.2.3 ОПЕРАЦИОННЫЕ МЕТАДАННЫЕ

Операционные метаданные детально описывают процессы обработки данных и управления доступом к ним. Примеры:

- журналы выполнения заданий пакетной обработки данных;
- история и результаты выгрузки выборок данных;
- сбои в расписаниях;
- результаты аудита, балансировки и контрольных измерений;
- журналы ошибок;
- структура, частота и время/скорость обработки запросов данных и отчетов;
- планы-графики исправлений, обновлений и выпуска новых версий и степень их соблюдения;
- правила резервного копирования, периодичности и сроков хранения резервных копий, порядок активации плана аварийного восстановления и т. п.;
- требования и условия соглашений об уровнях обслуживания;
- схемы регистрации и распределения потоковой нагрузки;
- правила архивирования данных, сроки хранения архивов, правила обеспечения связности архивных данных;
- критерии окончательного удаления (утилизации) архивных данных;
- правила совместного доступа к данным;
- технические роли и обязанности, контактные данные.

1.3.3 Стандарт ISO/IEC 11179

Стандарт ISO/IEC 11179 предоставляет рамочную структуру для организации регистра метаданных (Metadata Registry, MDR). Он разработан для того, чтобы обеспечить обмен данными, управляемый на основе метаданных (metadata-driven) и базирующийся на точных определениях данных, начиная с их отдельных элементов. Стандарт состоит из следующих частей[1]:

[1] Данные приведены по состоянию на 05.06.19 (стандарты находятся в состоянии динамической доработки). Для справки добавлены ссылки на соответствующие национальные стандарты РФ. — *Примеч. пер.*

Часть	Последняя версия ISO/IEC ISO/IEC 11179 IT — MDR	Соответствующий ГОСТ ГОСТ Р ИСО/МЭК 11179 ИТ. РМД
1.	ISO/IEC 11179-1:2015 Framework	ГОСТ Р ИСО/МЭК 11179-1-2010 Основные положения
2.	ISO/IEC TR 11179-2:2019 Classification	ГОСТ Р ИСО/МЭК 11179-2-2012 Классификация
3.	ISO/IEC 11179-3:2013 Registry metamodel and basic attributes	ГОСТ Р ИСО/МЭК 11179-3-2012 Метамодель регистра и основные атрибуты
4.	ISO/IEC 11179-4:2004 Formulation of data definitions	ГОСТ Р ИСО/МЭК 11179-4-2012 Формулировка определений данных
5.	ISO/IEC 11179-5:2015 Naming principles	ГОСТ Р ИСО/МЭК 11179-5-2012 Принципы наименования и идентификация
6.	ISO/IEC 11179-6:2015 Registration	Регистрация
7.	ISO/IEC DIS 11179-7 Metamodel for data set registration * ** на финальной стадии согласования*	Метамодель регистрации набора данных † *† информация о статусе в открытом доступе отсутствует*

1.3.4 Метаданные, используемые для описания неструктурированных данных

По своей природе все данные имеют какую-то структуру, хотя не все из них могут быть отструктурированы в виде привычных строк и столбцов реляционных баз данных. Любые данные, представленные не в виде баз данных или упорядоченных файлов данных, а в другой форме, включая зафиксированные на различных носителях документы, рисунки, фото, аудио- и видеозаписи и т. д., относятся к неструктурированным (см. главы 9 и 14).

Метаданные для управления неструктурированными данными, вероятно, необходимы даже в большей мере, чем для управления структурированными данными. Возвращаясь к умозрительной аналогии с библиотечным делом из вводной части главы: книги и периодика — хороший пример неструктурированных данных. Главное назначение метаданных в этом случае — упорядочение картотеки-каталога этих изданий, чтобы каждый владелец читательского билета мог найти искомые материалы вне зависимости от их формата.

Метаданные, используемые для учета и систематизации неструктурированных данных, включают:

- описательные метаданные (рубрики каталога, ключевые слова и т. п.);
- структурные метаданные (теги, поля, форматы и т. п.);
- административные метаданные (источники, расписания обновлений, права доступа, навигационная информация и т. п.);
- библиографические метаданные (записи в каталоге библиотеки и т. п.);
- учетные метаданные (сроки и правила хранения и т. п.);
- метаданные о порядке долговременного хранения (архив, условиях хранения, правила архивирования и т. п.; подробнее см. в главе 9).

Большинство вышеперечисленных утверждений о метаданных для неструктурированных данных относятся к области традиционного управления контентом. Что касается всплеска интереса к управлению неструктурированными большими данными, накапливаемыми в так называемых озерах данных на платформах типа Hadoop, то тут, как выясняется, не обойтись без автоматизированной каталогизации обрабатываемых данных с целью обеспечения возможности последующего доступа к ним. В большинстве подобных платформенных решений предусмотрены процедуры сбора метаданных в процессе приема данных. Каждому объекту при его размещении в озере присваивается минимальный набор атрибутов метаданных (имя, формат, источник, версия, дата поступления и т. п.). По этим данным и выстраивается каталог содержимого озера данных.

1.3.5 Источники метаданных

Одного взгляда на список типов метаданных достаточно, чтобы понять, что они собираются из множества различных источников. Кроме того, если управление метаданными отлажено и ведется на уровне всех приложений по отдельности, собрать эти метаданные воедино, согласовать и интегрировать не составит труда. Увы, во многих организациях управлением метаданными не занимаются и на уровне отдельных приложений, тогда как приложения имеют свойство генерировать метаданные автоматически в качестве побочного, но никак не конечного продукта (то есть метаданные с указанием автора, названия, даты и т. п. создаются и сохраняются, но никак не заточены под нужды потребителя). В результате, как и в случае с данными других категорий, интеграция метаданных требует больших объемов подготовительной работы.

Операционные метаданные так или иначе генерируются по мере обработки данных. Тут ключевая задача состоит в том, чтобы наладить их сбор в удобной для использования форме, назначить ответственных за их интерпретацию, формализацию и сопровождение — и обеспечить их всеми необходимыми для этого инструментами. Стоит отметить, что для интерпретации данных из таких источников, как, например, журналы регистрации ошибок, требуются метаданные, описывающие записи в журналах. Значительную часть технических метаданных также можно собирать из системных источников — например, объектов баз данных.

Немало бизнес-метаданных можно получить методом декомпиляции существующих систем, а также из сопроводительных глоссариев, моделей и документации процессов (Loshin, 2001; Aiken, 1995). Однако такой подход следует признать весьма рискованным, прежде всего из-за невозможности точно выяснить степень тщательности и добросовестности исходных определений. Если разработчики действующих систем в свое время что-то не учли или недоработали, то и полученные на основе их определений *метаданные* окажутся неточными или неоднозначными, то есть не будут выполнять своей функции, заключающейся в разъяснении потребителям точного смысла используемых ими данных.

Так что лучше целенаправленно выработать определения с нуля самостоятельно, чем полагаться на существующие. Для формулировки определений требуется немало времени

и определенные навыки (как минимум, владение языком технических описаний и умение переводить их на общедоступный язык). Именно поэтому разработка бизнес-метаданных должна осуществляться под опекой ответственных распорядителей данных бизнес-подразделений (см. главу 3).

Значительная часть технических метаданных для СУБД и бизнес-метаданных для пользовательских приложений может оперативно собираться в процессе их проектирования и при необходимости дорабатываться. Например, при моделировании данных так или иначе обсуждается точный смысл каждого элемента данных и связей между ними. Высказанные мнения должны документироваться, а потом их можно обобщить и взять за основу определений, используемых в словарях данных, бизнес-глоссариях и прочих хранилищах метаданных. Сами по себе модели данных также служат немаловажным источником детализированных сведений о физических характеристиках данных. Важно не жалеть времени на обеспечение наличия в сопроводительной документации к каждому проекту тщательно проработанных метаданных как артефакта, полностью соответствующего стандартам предприятия и готового к последующему использованию в интеграционных решениях.

Качественно определенные наборы бизнес-метаданных можно использовать без каких-либо изменений, перенося из проекта в проект и лишь дополняя новыми определениями по мере появления новых понятий и уточняя связанные определения. Такой подход весьма способствует выработке устойчивого и согласованного понимания отображения стандартных для бизнеса понятий в различных представлениях, описываемых различными наборами данных. По мере выработки универсальных метаданных многоцелевого назначения организации можно начать задумываться и о планировании проекта интеграции метаданных. Для начала, например, можно составить опись всех используемых систем с указанием всех метаданных, относящихся к каждой из них, и промаркировать все элементы метаданных тегами систем, в которых они используются.

Но помните, что создание метаданных — не самоцель. В большинстве организаций руководство едва ли согласится финансировать проекты по созданию метаданных ради метаданных, а не ощутимого экономического эффекта от них, — а если и согласится, то на финансирование эксплуатационных расходов на поддержание в рабочем состоянии систем, не приносящих отдачи, в долгосрочной перспективе можно не рассчитывать. В этом отношении метаданные ничем не отличаются от любых других данных — как, впрочем, и во всех иных отношениях. Любые данные должны создаваться как продукт тщательно определенного процесса и с использованием средств обеспечения их в целом высокого качества. Распорядители данных и специалисты по управлению данными должны следить за наличием и надлежащим функционированием механизмов ведения метаданных, относящихся к подконтрольным им процессам. Например, если организация собирает критически важные метаданные из моделей данных, нужно сделать так, чтобы процесс управления изменениями гарантировал перенос изменений в любой модели в метаданные, и наоборот.

Для иллюстрации широты диапазона и глубины проникновения метаданных в жизнь любой организации ниже приводятся примеры систем выступающих в роли источников метаданных[1].

1.3.5.1 РЕПОЗИТОРИИ МЕТАДАННЫХ ПРИЛОЖЕНИЙ

Репозиторием метаданных называют набор физических таблиц, в которых хранятся метаданные. Таблицы метаданных часто встраиваются в средства моделирования данных, BI и иные приложения. По мере созревания организации появляется естественное желание интегрировать метаданные из репозиториев различных приложений в единый комплекс хранения метаданных, чтобы у потребителей была возможность получить целостное представление обо всем спектре имеющейся информации.

1.3.5.2 БИЗНЕС-ГЛОССАРИЙ

Назначение бизнес-глоссария — документирование и хранение используемых в деловой практике организации терминов и определений, а также связей между ними.

Во многих организациях бизнес-глоссарий ведется в формате простой электронной таблицы. Однако в зрелых организациях часто используются имеющиеся в продаже или разрабатываемые собственными силами глоссарии сложной иерархической структуры, содержащие надежно проверенную информацию и поддерживающие управление изменяющимися со временем терминами, определениями и связями. Подобно любым системам, ориентированным на работу с данными, бизнес-глоссарии должны проектироваться таким образом, чтобы они были архитектурно согласованы с аппаратным и программным обеспечением, базами данных, процессами и человеческими ресурсами, распределенными по различным ролевым функциями и сферам ответственности. Приложение, реализующее бизнес-глоссарий, должно строиться таким образом, чтобы функционально оно отвечало потребностям трех основных целевых аудиторий.

- **Бизнес-пользователи**: аналитики данных, бизнес-аналитики, исследователи, руководители и менеджеры всех уровней используют бизнес-глоссарий для правильного понимания и истолкования терминологии и данных.
- **Распорядители данных** используют бизнес-глоссарий для управления жизненным циклом терминов и определений, а также для развития всестороннего понимания знаний, накопленных на уровне предприятия, посредством сопоставления со всеми без исключения элементами данных терминов, определяемых в глоссарии; а со всеми терминами — компонентов, к которым они относятся: например, рабочих метрик, отчетов, аналитических показателей качества данных или технологических процессов. Распорядители данных также выявляют проблемы рассогласованности терминов и определений, используемых на различных участках

[1] В англоязычном оригинале DMBOK2 источники метаданных (и подразделы с их описаниями) представлены просто в алфавитном порядке, а не в соответствии со степенью важности источника для организации (поскольку для отдельных организаций приоритеты могут быть разными). В данном издании сохранен порядок следования подразделов оригинала. — *Примеч. науч. ред.*

организации, и координируют выработки устраивающих все стороны решений по приведению терминологии организации к общему знаменателю.

- **Технические пользователи**: все, кто использует бизнес-глоссарий в процессе решения текущих задач архитектурного и системного проектирования, развития ИТ-инфраструктуры, проведения анализа последствий планируемых изменений и т. д. и т. п.

В бизнес-глоссарии должны фиксироваться свойства каждого термина, включая, например:

- **термин** (слово или словосочетание) и его определение, допустимые сокращения или аббревиатуры и синонимы (если таковые имеются);
- **бизнес-подразделение и/или приложение**, управляющее данными, к которым относится термин;
- **лицо, ответственное за сопровождение термина**, и дата последнего обновления определения;
- **классификационная или таксономическая принадлежность термина** (или соответствующая ему бизнес-функция);
- **определения, противоречащие определению термина**, с описанием характера противоречий, плана и сроков их разрешения/устранения;
- **распространенные неверные трактовки термина** с разъяснением сути ошибок;
- **технические алгоритмы** выработки определения;
- **происхождение данных**;
- **источник данных**, описываемых термином (официальный или авторитетный).

Любой бизнес-глоссарий на практике должен дополняться базовым набором отчетов, предназначенных для использования в административных процессах. Организациям настоятельно рекомендуется реализовывать глоссарии в виде электронных справочных ресурсов, а не в печатной форме, поскольку содержание глоссариев динамично меняется. За составление, ведение, использование, обработку и учет глоссариев обычно отвечают распорядители данных. Функция учета включает отслеживание новых терминов и определений, нуждающихся в рецензировании и утверждении, мониторинг статуса терминов и определений, находящихся в процессе согласования, и ведение отдельного сводного отчета с проблемными терминами, у которых отсутствуют определения или иные обязательные атрибуты (см. раздел 6.4).

Структура и функциональность бизнес-глоссариев могут варьироваться в широких пределах. Чем проще дается пользователям поиск по бизнес-глоссарию, тем выше вероятность обращения к его контенту. Однако ради простоты в обращении нельзя жертвовать самым главным качеством бизнес-глоссария, а именно — четкостью, полнотой и однозначностью трактовок определений.

1.3.5.3 ИНСТРУМЕНТЫ БИЗНЕС-АНАЛИТИКИ (BI)

Инструменты бизнес-аналитики генерируют собственные наборы метаданных, относящихся к бизнес-аналитическим моделям, включая описания обзорной информации, классов, объектов,

производных и рассчитываемых величин, фильтров, отчетов, полей и структуры отчетов, категорий пользователей, которым адресованы отчеты, периодичности публикации и каналов распространения отчетов.

1.3.5.4 СРЕДСТВА УПРАВЛЕНИЯ КОНФИГУРАЦИЯМИ

Средства управления конфигурациями или базы данных управления конфигурациями (Configuration Management Databases, CMDB[1]), обеспечивают функциональную возможность ведения метаданных в привязке к конкретным ИТ-ресурсам, включая управление связями между ними и деталями договоров, регулирующих доступ к каждому ресурсу. В CMDB каждый ИТ-ресурс называется элементом конфигурации (Configuration Item, CI). В стандартной архитектуре CMDB сбор метаданных и управление ими реализованы на уровне типов CI. Многие организации практикуют интеграцию CMDB с процессами управления изменениями, что позволяет выявлять связанные между собой CI (ресурсы и/или приложения) и обеспечивать согласованное изменение настроек остальных связанных CI при изменении какой-либо настройки одного из CI. При этом архитектура хранилищ предусматривает механизмы привязки ресурсов репозитория метаданных к элементам физической реализации CMDB, что обеспечивает возможность получения полной картины распределения данных по платформам.

1.3.5.5 СЛОВАРИ ДАННЫХ

Словари определяют структуру и содержание наборов данных, используемых чаще всего в отдельно взятой базе данных, приложении или хранилище. Словарь можно использовать для управления именами/названиями, описаниями, структурой, характеристиками, сроками и правилами хранения, значениями по умолчанию, связями/отношениями и/или ссылками, свойствами уникальности и иными атрибутами на уровне элементов данных модели. Также словарь должен содержать определения таблиц или файлов данных. Функции управления словарями данных встраиваются также и в программные средства создания, сопровождения и/или обработки различных массивов данных. В любом случае, чтобы сделать эти метаданные доступными потребителям данных, их нужно извлечь из источников — баз данных или средств моделирования. Словари данных могут также описывать на языке бизнес-терминологии, какие элементы данных доступны сообществу пользователей, какие ограничения по их выдаче предусмотрены требованиями информационной безопасности и защиты данных, применимых к различным бизнес-процессам. Для экономии времени, которое будет впоследствии затрачиваться на определение, публикацию и ведение словарных статей, полезно изначально включить в проект еще на уровне логической модели семантический слой с поддержкой анализа и учета контента. Однако, как уже отмечалось, полагаться на существующие определения весьма рискованно, особенно в организации, пребывающей на зачаточном уровне понимания смысла и процессов управления метаданными.

[1] *сокр.* от *англ.* Configuration management database. — *Примеч. пер.*

Многие ключевые бизнес-процессы, связи и термины выявляются и определяются на стадии разработки модели данных. При этом значительная часть этой бесценной информации часто утеривается при переходе от логической модели к физической или при реализации физической модели в производственной среде. Словарь данных помогает обеспечить сохранность этой информации, которая может очень и очень пригодиться организации на последующих этапах приведения физической модели в соответствие с логической и согласования с ними решений, развернутых в производственной среде.

1.3.5.6 СРЕДСТВА ИНТЕГРАЦИИ ДАННЫХ

Многие средства интеграции данных используются также и в качестве исполняемых программ для переноса данных из системы в систему или между различными модулями одной и той же системы. При этом они обычно генерируют промежуточные временные файлы, содержащие копии переносимых данных или производные от них. Подобные средства способны загружать в свою рабочую область данные из самых разнообразных источников и проводить над ними всевозможные операции: группировать, исправлять, переформатировать, объединять, фильтровать и т. п., — а полученные на выходе результаты распространять по целевым адресам. При этом средства интеграции обязательно ведут учет движения данных между системами и попутных преобразований, обеспечивая полное документирование их происхождения. Любое успешное решение по управлению метаданными должно опираться на использование метаданных, описывающих происхождение и полную цепь преобразований данных в процессе интеграции, чтобы всегда можно было проследить полный и единственный путь каждого элемента данных от первоисточника до конечного пункта назначения.

Средства интеграции данных дают возможность внешним хранилищам метаданных собирать данные о происхождении информации и получать доступ к временным файлам метаданных через интерфейсы приложений (API). После завершения сбора всей необходимой информации хранилищем метаданных некоторые средства позволяют генерировать целостные диаграммы происхождения любого элемента данных. Средства интеграции данных также поддерживают ведение метаданных об исполнении различных задач по интеграции данных, включая данные о последнем успешном запуске, продолжительности обработки и статусе исполнения задания. Некоторые хранилища метаданных способны также извлекать статистику последней обработки метаданных и отображать ее рядом с элементами данных (см. главы 6 и 8).

1.3.5.7 КАТАЛОГИ БАЗ ДАННЫХ И СУБД

Каталоги баз данных — один из важнейших источников исходной информации, требующейся для определения и обеспечения актуальности метаданных. В каталогах описывается наполнение баз данных и содержатся сведения о форматах, свойствах и длине полей, версиях и статусе развертывания ПО, времени доступности сетевой инфраструктуры, систем и данных для пользователей, а также множество других атрибутов операционных и технических метаданных. Традиционно самая распространенная модель данных — реляционная. Системы управления реляционными базами данных рассматривают данные как набор связанных таблиц с различным числом столбцов, в которых

содержатся значения различных свойств или атрибутов строчных элементов данных, которые могут индексироваться, фильтроваться, запрашиваться для просмотра или обработки (см. главу 5). Решение по управлению метаданными должно «уметь» подключаться к базам данных различной архитектуры и наборам данных под управлением различных приложений с целью считывания всех без исключения метаданных, которые можно вытянуть из каждой базы данных. Некоторые средства управления хранилищем метаданных поддерживают интеграцию метаданных, извлеченных из различных СУБД и репозиториев приложений, что бывает весьма полезно для получения более целостной картины о фактически имеющихся у организации на физическом уровне ресурсах данных.

1.3.5.8 ИНСТРУМЕНТЫ УПРАВЛЕНИЯ МЭППИНГОМ ДАННЫХ

Инструменты управления мэппингом данных используются на стадии анализа структуры и проектирования архитектуры данных предприятия для перевода проектных требований на язык спецификаций мэппинга, после чего полученные спецификации могут использоваться либо программным средством интеграции, либо программистами для написания утилит, используемых для преобразования данных на стыках между системами с целью их интеграции. Документация, описывающая мэппинг, часто ведется в масштабе организации в формате электронных таблиц, например Excel. В последнее время поставщики платформенных решений стали использовать вариант корпоративной архитектуры данных предприятия с централизованным хранилищем спецификаций мэппинга, поддерживающим контроль и сравнение версий. Модули мэппинга часто включаются в пакеты интеграционных решений, что позволяет автоматизировать программирование решений по интеграции данных, а зачастую и обмен данными с хранилищами других метаданных и справочных данных (см. главу 8).

1.3.5.9 ИНСТРУМЕНТЫ ОЦЕНКИ И КОНТРОЛЯ КАЧЕСТВА ДАННЫХ

Инструменты проверки данных позволяют оценивать качество данных, определяя и фиксируя степень их соответствия установленным критериям точности, достоверности, актуальности и т. п. Большинство подобных средств поддерживают функции обмена балльными оценками и профилями качества данных с репозиториями других метаданных, что позволяет системе управления главным репозиторием метаданных присваивать оценки качества соответствующим ресурсам физических данных.

1.3.5.10 СПРАВОЧНИКИ И КАТАЛОГИ

Словари и глоссарии данных содержат детальную информацию о терминологии, таблицах и полях, а справочники и каталоги — сведения технического характера о системах, использующих данные, их источниках и местах хранения внутри организации. Справочник метаданных особенно полезен разработчикам и суперпользователям (администраторам, распорядителям и аналитикам данных), поскольку позволяет составить полное представление обо всем спектре данных, имеющихся в распоряжении организации, и очертить круг потенциальных источников проблем или входных данных для новых приложений.

1.3.5.11 СРЕДСТВА ОБМЕНА СООБЩЕНИЯМИ О СОБЫТИЯХ

Средства обмена сообщениями используются для передачи данных из системы в систему без необходимости какой-либо интеграции самих систем. Соответственно, для понимания смысла таких сообщений требуется множество метаданных на обеих сторонах канала передачи сообщений. Средства передачи сообщений генерируют метаданные, описывающие структуру и порядок их передачи, а средства обработки входящих сообщений их интерпретируют и регистрируют соответствующие события в своей системе. Подобные инструменты обычно включают графические интерфейсы управления логикой перемещения данных. Детали реализации интерфейсов, алгоритмов и статистического учета движения и обработки сообщений они могу экспортировать в другие репозитории метаданных.

1.3.5.12 ИНСТРУМЕНТЫ И РЕПОЗИТОРИИ ДЛЯ МОДЕЛИРОВАНИЯ ДАННЫХ

Инструменты моделирования данных используются для построения моделей данных различного уровня — концептуальных, логических и физических. В процессе моделирования эти программные средства производят метаданные, описывающие компоненты архитектуры проектируемого приложения или системы, такие как предметные области, логические сущности и атрибуты, связи между сущностями и их атрибутами, родительские типы и подтипы, таблицы, столбцы, индексы, первичные и внешние ключи, ограничения по целостности и прочие свойства, которые предусмотрены выбранными моделями данных. Данные о параметрах всех разрабатываемых моделей сохраняются в репозиториях метаданных соответствующих инструментов моделирования, а затем импортируются и интегрируются в корпоративное хранилище метаданных. Кроме того, средства моделирования часто служат источником терминов и определений для словарей данных.

1.3.5.13 РЕПОЗИТОРИИ СПРАВОЧНЫХ ДАННЫХ

Справочные данные документируют описания и значения различных данных, требующихся для бизнеса. Эти данные классифицируются по типам и контекстам (областям) использования в системе. Средства управления справочными данными позволяют также определять связи между различными кодифицированными величинами как внутри области, так и в более широком контексте. Эти наборы инструментов обычно поддерживают отправку подборок справочных данных в хранилище метаданных, где, в свою очередь, могут быть предусмотрены механизмы включения полученных справочных данных в бизнес-глоссарий и соответствующие столбцы таблиц или поля записей физических моделей, где эти данные используются.

1.3.5.14 РЕЕСТРЫ СЕРВИСОВ

При использовании сервис-ориентированной архитектуры (SOA) вся техническая информация о запущенных сервисах, потоках обработки и конечных адресатах данных хранится в реестре сервисов, с помощью которого осуществляются все конфигурационные настройки систем и приложений. Через реестр с могут задаваться, например, определения, интерфейсы, операции,

параметры ввода/вывода, политики, версии и типовые сценарии использования служб данных. Важнейшими метаданными о любом сервисе являются: номер версии, местонахождение, ЦОД, статус доступности, дата подключения, сервисный порт, IP-адрес, порт статистики, время ожидания подключения, повторного подключения и число попыток подключения до выдачи сообщения об ошибке, и т. п. Имеется возможность адресовать запросы к реестрам сервисов на предмет получения интересующих данных: например, перечня всех активных или доступных сервисов, информации о текущих версиях, списка устаревших сервисов или детальной информации о каком-то конкретном сервисе. Также можно анализировать имеющиеся в реестре данные о сервисах на предмет выбора подходящего кандидата для повторного использования. Информация из подобных источников нередко позволяет выявлять важные факты, касающиеся данных и схем их перемещения между системами и/или приложениями. Метаданные, извлекаемые из реестров сервисов, можно интегрировать с метаданными, полученными из других источников, для получения полной картины обмена данными между всевозможными системами.

1.3.5.15 ПРОЧИЕ ХРАНИЛИЩА МЕТАДАННЫХ

Метаданные могут обнаруживаться в самых разных хранилищах, документах и источниках, включая специализированные перечни, такие как журналы регистрации событий, списки источников или интерфейсов, наборы кодов, всевозможные словари, пространственные и временны́е схемы, картографические привязки, наборы цифровых данных, распределенных по географическому признаку, архивы репозиториев, бизнес-правила и т. д.

1.3.6 ТИПЫ АРХИТЕКТУРЫ МЕТАДАННЫХ

Как и любые другие данные, метаданные имеют жизненный цикл. Поэтому все решения по управлению метаданными включают следующие архитектурные уровни, соответствующие различным фазам жизненного цикла метаданных:

- создание или получение метаданных;
- хранение метаданных в одном или нескольких репозиториях;
- интеграция метаданных;
- доставка метаданных потребителям;
- использование метаданных;
- контроль и управление метаданными.

Для подключения к источникам, а также для сбора, хранения, интеграции и сопровождения метаданных и управления доступом к ним могут использоваться различные архитектурные подходы.

1.3.6.1 ЦЕНТРАЛИЗОВАННАЯ АРХИТЕКТУРА МЕТАДАННЫХ

Централизованная архитектура предусматривает единое хранилище метаданных, копируемых туда из различных источников. Организациям с ограниченными ИТ-ресурсами, как

и стремящимся к максимально возможной автоматизации управления метаданными, такой вариант архитектуры, как правило, противопоказан. А вот организации, стремящиеся к согласованности метаданных, извлекают максимальную пользу от хранения их в централизованном хранилище.

Преимущества централизованного репозитория метаданных:

- Высокая доступность вследствие автономности от систем-источников.
- Высокая скорость обработки запросов на извлечение метаданных, поскольку все запросы обрабатываются без обращений за пределы центрального репозитория.
- Структуры баз данных представлены в хорошем разрешении и не зависят от сторонних или коммерческих систем.
- Извлеченные из хранилища метаданные допускают преобразование, настройку или дополнение метаданными из других источников с целью повышения их качества.

Недостатки централизованного подхода:

- Сложность процессов оперативного воспроизведения изменений в метаданных систем-источников в хранилище метаданных.
- Дороговизна оборудования и эксплуатационного сопровождения централизованного хранилища.
- Возможная потребность в разрабатываемых под заказ модулях сопряжения или межплатформенном ПО для обмена метаданными между хранилищем и системами.
- Проверка корректности работы и техническое сопровождение компонентов ПО, разрабатываемого под заказ, повышает потребность в высококвалифицированных ИТ-специалистах — как в штате организации, так и на стороне поставщика ПО.

Рисунок 85 иллюстрирует реализацию сбора метаданных в отдельном централизованном хранилище, заполняемом посредством импорта метаданных из различных источников (стрелки). Конечным же пользователям предлагается адресовать запросы к центральному репозиторию через портал доступа к метаданным. При такой конфигурации пользователь лишен возможности запрашивать метаданные напрямую у программных средств, которые их генерируют. Зато поддерживается глобальный поиск по всем метаданным, собранным в хранилище.

1.3.6.2 РАСПРЕДЕЛЕННАЯ АРХИТЕКТУРА МЕТАДАННЫХ

Полностью распределенная архитектура предусматривает единую точку доступа к метаданным через портал, обеспечивающий извлечение запрашиваемых данных систем-источников в режиме, близком к реальному времени. Центральное хранилище при такой архитектуре отсутствует; вместо него в среде портала управления метаданными ведутся каталоги данных, содержащихся в системах-источниках, и действуют общие правила оптимизации обработки запросов, а обращение

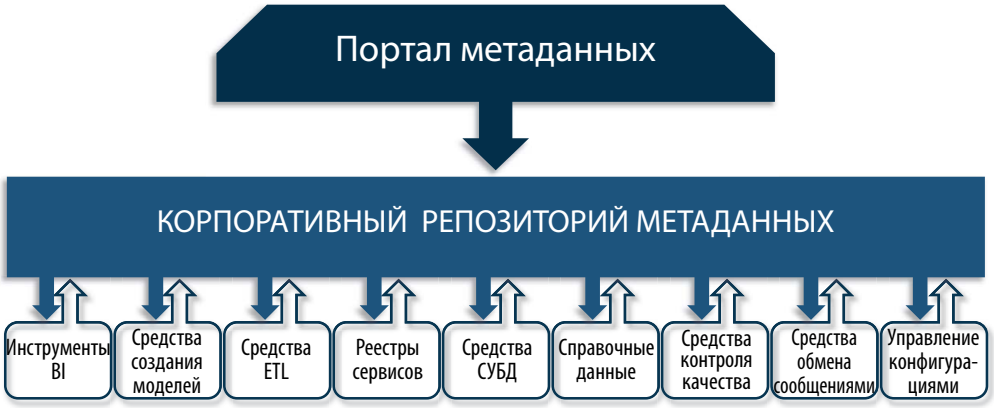


Рисунок 85. Централизованная архитектура метаданных

непосредственно к системам-источникам осуществляется посредством протоколов, используемых промежуточным ПО, — например, с помощью брокера объектных запросов (ORB[1]).

Преимущества распределенной архитектуры метаданных:

- Метаданные не требуют обновления и проверки актуальности, поскольку всякий раз запрашиваются из первоисточника.
- Запросы, обращенные к распределенным источникам, сразу же распределяются по множественным каналам обмена данными и, как следствие, могут обрабатываться оперативнее.
- Метаданные из проприетарных систем так или иначе ничего, кроме отправки запроса и получения ответа, не предусматривают, поскольку структура данных в таких системах защищена патентами; следовательно, ничего свыше этого минимума реализовывать и не требуется.
- Обработка запросов к метаданным упрощается еще и за счет автоматизации.
- Минимизируются потребности в пакетной обработке, поскольку метаданные при такой архитектуре не требуют ни тиражирования, ни синхронизации.

Минусы распределенных архитектур:

- Отсутствие всякой возможности реализовать поддержку пользовательских определений или добавлений записей метаданных за отсутствием какого бы то ни было буферного хранилища между порталом доступа и системами-источниками.
- Стандартизированное представление метаданных из различных систем без учета их специфики.
- Недоступность метаданных, описывающих данные из систем-источников, в случае отсутствия связи с последними.
- Качество метаданных всецело зависит от контроля качества на стороне систем-источников.

[1] *сокр.* от *англ.* Object Request Broker. — *Примеч. пер.*

Рисунок 86 иллюстрирует распределенную архитектуру метаданных. Централизованный репозиторий метаданных в данном случае отсутствует, а портал переадресует пользовательские запросы напрямую программным средствам управления источниками, которые их и исполняют. За неимением централизованного хранилища для сбора метаданных из различных источников каждый запрос приходится переадресовывать соответствующей системе-источнику; как следствие, отсутствует возможность реализации функций глобального поиска метаданных по всем доступным источникам.

Рисунок 86. Распределенная архитектура метаданных

1.3.6.3 ГИБРИДНАЯ АРХИТЕКТУРА МЕТАДАННЫХ

Из самого ее названия явствует, что гибридная архитектура метаданных сочетает в себе элементы, свойства и характеристики как централизованной, так и распределенной архитектур. Метаданные всё так же поступают в центральный репозиторий непосредственно из систем-источников, вот только сохраняются они там выборочно. Обычно система управления таким хранилищем предусматривает сохранение критически важных стандартизованных элементов метаданных из систем-источников и последующее добавление дополнительных элементов по запросу пользователей, в том числе в ручном режиме из сторонних источников.

Преимущества подобной архитектуры включают получение запрашиваемых метаданных из систем-источников в режиме, близком к реальному времени, и возможность эффективной проработки расширенных метаданных, когда это нужно пользователю. Гибридный подход способствует снижению потребности во вмешательстве ИТ-специалистов с целью ручной настройки и программирования межплатформенных интерфейсов доступа к проприетарным системам, когда такой доступ требуется. Актуальность метаданных гарантируется регулярными обновлениями базового набора и запросом пользователями остальных элементов по мере надобности из первоисточников.

А вот в плане доступности метаданных из систем-источников гибридная архитектура наследует все минусы распределенной: поскольку запросы обрабатываются удаленными системами, повлиять на их доступность и производительность со стороны портала метаданных нереально. Кроме того, минусом гибридной архитектуры является и дополнительный расход вычислительных ресурсов на привязку слоя текущих метаданных, полученных по оперативному запросу из

системы-источника, к стационарному слою метаданных централизованного хранения перед выдачей сводного набора запрошенных метаданных конечному пользователю портала.

Для многих организаций, однако, преимущества гибридной архитектуры перевешивают недостатки. В целом, показаниями к использованию гибридной архитектуры служат: частое внесение изменений в операционные метаданные; необходимость полагаться исключительно на актуальные, но при этом согласованные, консолидированные и унифицированные метаданные; ситуации быстрого роста объемов и числа источников метаданных.

Организациям с относительно статичными метаданными, не планирующим в обозримом будущем резкого расширения профилей метаданных, подобный вариант архитектуры не то чтобы не подходит, а может не оправдать ожиданий, ничего не добавив по сравнению со стандартной централизованной архитектурой и практически не раскрыв своего потенциала.

1.3.6.4 ДВУНАПРАВЛЕННАЯ АРХИТЕКТУРА МЕТАДАННЫХ

Другим усовершенствованным вариантом архитектуры метаданных является двунаправленная (bi-directional) архитектурная модель, допускающая изменение метаданных на любом участке (источник, среда интеграции, пользовательский интерфейс) с последующим согласованием изменений системой управления хранилищем (брокером) и передачей результата в систему-источник.

Такой подход сопряжен со множеством всевозможных трудностей. Само проектное решение требует, чтобы в главном хранилище содержались исключительно новейшие версии метаданных из систем-источников, а в случае рассогласованности вынуждает вносить изменения еще и в исходные метаданные на уровне источников. Все изменения подлежат систематической регистрации и последующему согласованию. Как следствие, приходится выстраивать целый промежуточный слой с наборами интерфейсов сопряжения репозитория с источниками с прямой и обратной связью, обеспечивающих двунаправленное согласование данных между ними, которые также требуют настройки и управления.

Рисунок 87 иллюстрирует порядок сбора общих метаданных, предназначенных для обеспечения совместного доступа к данным. Из различных систем-источников они поступают в профильные разделы централизованного хранилища метаданных. Пользователи адресуют запросы к порталу метаданных; портал передает пользовательский запрос в центральный репозиторий; система управления центрального репозитория пытается выполнить запрос, используя метаданные в совместном доступе, ранее собранные в тематические разделы хранилища из различных систем-источников. Если же запрос оказывается слишком специфическим или требующим более детализированных метаданных, нежели те, что имеются на уровне централизованного репозитория, он перенаправляется на самый нижний уровень и адресуется системе-первоисточнику, которая и пытается разыскать в своих недрах запрашиваемые специфические детали. Благодаря наличию набора общих метаданных в централизованном репозитории такая архитектура обеспечивает возможность глобального поиска с использованием всего спектра доступных средств.

Рисунок 87. Гибридная архитектура метаданных

2. ПРОВОДИМЫЕ РАБОТЫ

2.1 Определение стратегии работы с метаданными

Стратегия работы с метаданными описывает намерения организации по управлению метеданными и этапы перехода из текущего состояния к оптимальной в ее представлении практике в обозримом будущем. Стратегия должна служить для команд разработчиков рамочной структурой совершенствования управления метаданными. При этом выработка требований к самим метаданным помогает прояснить и основные факторы влияния на выбор стратегии, и потенциальные препятствия на пути ее претворения в жизнь.

Стратегия должна включать определение корпоративной архитектуры метаданных, а также фазы ее внедрения, требуемые для решения стратегических задач. Комплекс мероприятий в области стратегического планирования включает следующие шаги.

- **Инициирование деятельности по стратегическому планированию в области метаданных**. На этом шаге команде по выработке стратегии работы с метаданными необходимо определиться с ближайшими и долгосрочными целями. В рамках стратегического планирования следует разработать проект общих положений, очерчивающих круг работ и задач, решаемых в области метаданных в контексте общеорганизационных усилий по руководству данными, а также подготовить план коммуникаций с целью обеспечения поддержки проводимых мероприятий. К планированию должны быть привлечены все ключевые заинтересованные стороны.

- **Обсуждение с ключевыми заинтересованными сторонами** из числа руководителей бизнес- и ИТ- подразделений позволяет заложить надежный фундамент из всесторонних знаний для обоснования стратегии управления метаданными.
- **Оценка существующих источников метаданных и информационной архитектуры** позволяет определить степень сложности и возможные пути решения проблем путем обсуждения с ключевыми специалистами блока ИТ и изучения документации с описанием архитектуры систем, моделей данных и т. д.
- **Разработка будущей архитектуры метаданных**. На этой стадии определяется общая концепция и разрабатывается целевая архитектура среды управления метаданными. При этом должны в полной мере учитываться все стратегические аспекты, такие как организационная структура, согласование архитектуры метаданных с основными направлениями деятельности в области руководства и распоряжения данными, механизмы управления архитектурой метаданных и средствами доставки метаданных, техническая архитектура и архитектура систем безопасности.
- **Разработка плана поэтапного внедрения** включает проверку и подтверждение, интеграцию и приоритизацию результатов обсуждений и анализа данных, после чего документируется окончательный вариант стратегии работы с метаданными и определяется подход к поэтапному внедрению изменений, необходимых для перехода от имеющейся к будущей среде управления метаданными.

Со временем стратегия будет развиваться, корректироваться и уточняться, равно как и требования к метаданным, их архитектура, а также понимание жизненного цикла метаданных.

2.2 Выработка понимания требований к метаданным

Определение требований к метаданным начинается с содержательной части. Необходимо выяснить, что именно должны описывать метаданные на каждом уровне архитектурного проекта. В частности, нужно определиться с именами таблиц и столбцов в логической и физической моделях данных. Контент метаданных может варьироваться в весьма широких пределах в зависимости от нужд бизнеса и потребителей технических данных (см. раздел 1.3.2).

Комплексное решение по управлению метаданными обязательно должно удовлетворять ряду функциональных требований, в частности тем, которые затрагивают следующие категории вопросов.

- **Изменения.** Как часто будут пересматриваться и обновляться наборы и атрибуты метаданных?
- **Синхронизация.** Как спланировать график обновлений метаданных в привязке к изменениям в источниках?
- **История.** Сохранять ли в архивах предыдущие версии *метаданных* и на какой срок?
- **Доступ.** Кто будет иметь право доступа к метаданным? Как будет осуществляться доступ? Какие именно функции должен поддерживать пользовательский интерфейс?

- **Структура.** В соответствии с какой моделью будет организовано хранение метаданных?
- **Интеграция.** Степень и правила интеграции метаданных из различных источников.
- **Сопровождение.** Процессы, правила и процедуры обновления метаданных (ведение журналов и порядок согласования и утверждения изменений).
- **Управление.** Распределение ролей и обязанностей в области управления метаданными.
- **Качество.** Требования к качеству метаданных и механизмы контроля их соблюдения.
- **Безопасность.** Часть метаданных может не подлежать раскрытию, поскольку само их существование свидетельствует о наличии у организации данных с высокой степенью конфиденциальности.

2.3 Определение архитектуры метаданных

Система управления метаданными для начала должна уметь извлекать сами метаданные из различных источников. Следовательно, архитектура системы обязана обеспечивать возможность регулярного сканирования разнообразных источников метаданных на предмет появления новых и изменения имеющихся элементов метаданных и обновления соответствующих записей в центральном хранилище. Кроме того, система должна поддерживать ввод и редактирование метаданных в ручном режиме, обработку поисковых и справочных запросов по метаданным, поступающих от различных групп пользователей.

Среда управляемых метаданных призвана надежно изолировать от конечного пользователя множественные источники разрозненных метаданных. Следовательно, ее архитектура должна предусматривать единственную точку доступа пользователей к централизованному хранилищу метаданных. Через эту точку доступа (портал) пользователю открывается связная и прозрачная картина метаданных из всех источников. Пользователи должны иметь доступ к метаданным, оставаясь в неведении об их происхождении из разнородных сред-источников. В аналитических приложениях, включая ориентированные на обработку больших данных, могут предусматриваться определяемые пользователем функции (UDF[1]) обращения к данным из различных наборов, основанные на использовании метаданных. Чем меньше в решении возможностей для применения UDF, тем чаще и глубже будут конечные пользователи докапываться до первоисточников, собирать, просматривать и анализировать напрямую исходные наборы данных и служебные определения метаданных в рабочих системах, что, как правило, чревато всяческими рисками утечек чувствительных данных и прочими угрозами информационной безопасности.

Выбор архитектуры зависит от специфики потребностей организации. С технической точки зрения выделяют три основных подхода к построению репозитория метаданных, которые в целом повторяют архитектурные подходы к построению хранилищ данных, — централизованный, распределенный и гибридный (см. раздел 1.3.6). Выбор подхода производится с учетом технических возможностей в плане реализации хранилища и механизмов обновления метаданных.

[1] *сокр.* от *англ.* user-defined function(s). — *Примеч. пер.*

2.3.1 Создание метамодели

Создание модели данных для репозитория метаданных (или метамодели — metamodel) — один из первых практических шагов по проектированию, следующий после завершения разработки стратегии работы с метаданными и уяснения бизнес-требований. Метамодель может создаваться по мере необходимости на различных уровнях обобщения/конкретизации — от высокоуровневой концептуальной модели, объясняющей отношения и связи на уровне систем, до глубоко детализированной метамодели, исчерпывающим образом прописывающей все атрибуты, элементы и процессы. Являясь прежде всего инструментом планирования и формулировки требований, метамодель еще и сама по себе служит ценным источником метаданных.

В представленном примере (рис. 88) модели репозитория метаданных (метамодели) прямоугольники представляют основные высокоуровневые сущности, содержащие данные.

2.3.2 Применение стандартов метаданных

Решение по управлению метаданными должно соответствовать внутренним и внешним стандартам, а перечень применимых и обязательных для соблюдения требований составляться, согласовываться и утверждаться еще на стадии стратегического планирования. Мониторинг же их соблюдения — одна из важных функций руководства данными. Внутренние стандарты метаданных организации определяют допустимые наименования, форматы и свойства, требования защиты, прозрачности и документирования обработки данных, и т. п. Внешние стандарты метаданных, применимые к организации, включают форматы и протоколы обмена данными, требования к API и прочие технические регламенты.

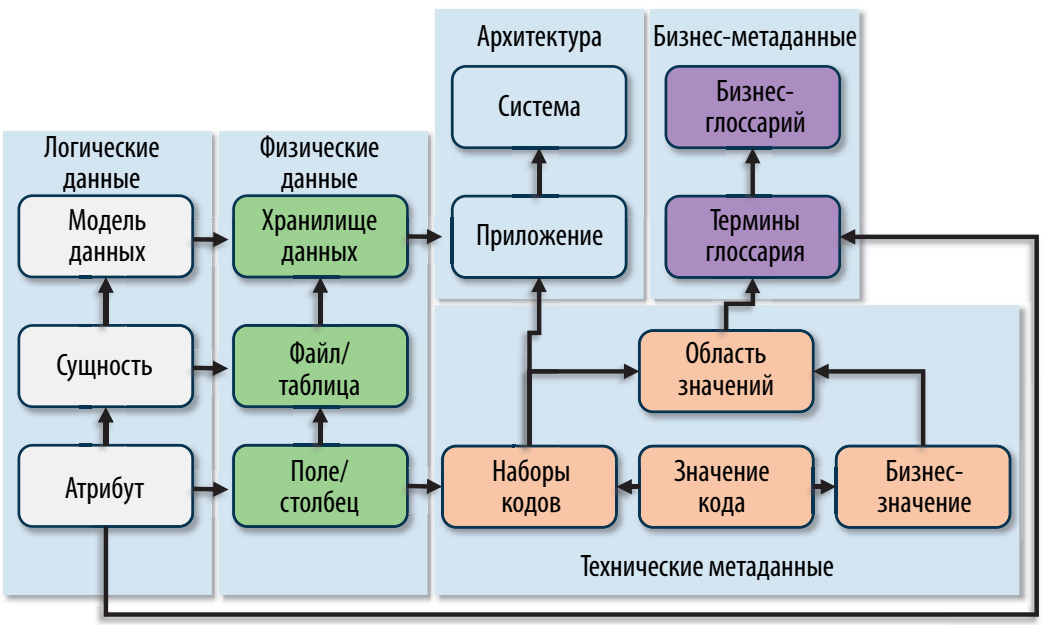


Рисунок 88. Пример метамодели (модели репозитория метаданных)

2.3.3 Управление хранилищами метаданных

Следует проработать и реализовать комплекс механизмов управления средой метаданных. К области управления хранилищами относится весь спектр операций по перемещению и обновлению метаданных во всех системах-источниках и центральном хранилище под контролем специалистов по метаданным. В основном это административные по своей сути функции — мониторинг, обработка отчетов, реагирование на предупреждения, контроль рабочих журналов, разрешение проблем, выявленных в среде хранения метаданных, и т. п. Многие работы в рамках этих контрольных функций являются стандартными для эксплуатации и обслуживания баз данных и интерфейсов в любых информационных средах. Контрольные мероприятия должны осуществляться под общим руководством администратора или органа, отвечающего за распоряжение данными. Ниже обобщены основные направления оперативного управления хранилищами метаданных.

- Контрольные мероприятия и текущие задачи управления, включая:
  - планирование и контроль соблюдения графика регламентных работ;
  - анализ статистики нагрузки, трафика и т. п.;
  - резервное копирование и восстановление, архивирование и полное удаление;
  - текущие изменения конфигурационных настроек;
  - отладку и настройку с целью оптимизации и повышения производительности;
  - анализ статистики запросов;
  - запросы и генерирование отчетов;
  - управление средствами обеспечения ИБ.
- Оперативные мероприятия по управлению качеством, включая:
  - обеспечение/контроль качества;
  - дифференцированную настройку частоты обновления различных наборов данных;
  - выявление и учет недостающих метаданных;
  - выявление и учет устаревших метаданных.
- Работы по управлению метаданными, включая:
  - загрузку, сканирование, импорт и маркировку массивов данных;
  - картирование источников и конфигурирование каналов передачи данных;
  - управление версиями;
  - управление пользовательским интерфейсом;
  - ведение метаданных связующих наборов данных (при схемах NoSQL);
  - привязку внутренних данных к источникам (ссылки и метаданные задач);
  - управление лицензиями на доступ к внешним данным или подписками на них;
  - получение метаданных, необходимых для обогащения данных (например, подключение к GIS).

- Организация переподготовки и обучения, включая:
  - обучение пользователей и повышение квалификации ответственных за данные;
  - измерение и анализ показателей эффективности управления данными;
  - тренинги по управлению данными, формированию запросов и отчетов.

2.4 Создание и ведение метаданных

Как явствует из детального описания в разделе 1.3.5, метаданные создаются самыми разнообразными процессами и сохраняются на различных участках работы организации. Однако для обеспечения высокого качества метаданных управлять ими нужно как единым информационным продуктом. Сами по себе качественные метаданные не образуются. Управление ими требует планирования в масштабах организации (см. главу 13).

Ниже перечислены наиболее общие принципы обеспечения качества метаданных.

- **Ответственность.** Нужно понимать, что метаданные чаще всего создаются в рамках формализованных процессов (моделирование данных, разработка систем (SDLC), определение бизнес-процессов и т. п.), и лица, отвечающие за эти процессы, в равной мере несут ответственность и за качество метаданных.
- **Стандарты.** Устанавливайте стандарты метаданных и соответствующие им правила управления метаданными, контролируйте их соблюдение и проверяйте действующие стандарты на предмет их полезности с точки зрения упрощения интеграции и использования метаданных.
- **Совершенствование.** Создайте механизм обратной связи, чтобы потребители имели возможность оперативно информировать группу управления метаданными о неточных или устаревших данных.

Как и любые другие данные, метаданные можно профилировать и проверять на предмет их качества. Кроме того, деятельность по ведению метаданных должна быть регламентирована или осуществляться в рамках подлежащего аудиту компонента работ по проекту.

2.4.1 Интеграция метаданных

Интеграционные процессы заключаются в сборе и консолидации метаданных в масштабах организации, включая не только генерируемые собственными системами и приложениями, но и поступающие извне. В центральном репозитории метаданных должна осуществляться интеграция извлеченных из систем-источников технических метаданных с метаданными, описывающими технологические процессы, бизнес-процессы и административные процедуры. Технически извлечение метаданных из источников может реализовываться с помощью специальных программных модулей — адаптеров, сканеров или сопрягающих приложений. В интегрированных средах иногда возможен вариант прямого доступа системы управления центральным репозиторием метаданных к базам данных источников. Адаптеры или сканеры данных поставляются в комплекте

многих коммерческих инструментов разработки ПО, включая и специализированные средства интеграции метаданных. В крайнем случае их можно разработать самостоятельно, используя предлагаемые инструментами API .

Главные трудности подстерегают проектировщиков там, где необходимо обеспечить интеграцию метаданных при одновременной поддержке высокоуровневого согласования и управления. Скажем, любая попытка интеграции внутренних наборов данных с разнородными данными из внешних источников (официальной статистикой, неструктурированными данными из публикаций, статей и аналитических отчетов и т. п.) неизбежно осложняется вследствие возникновения массы вопросов по поводу обеспечения качества и семантической согласованности.

Сканирование источников на предмет выявления метаданных, подлежащих интеграции в репозитории, может быть реализовано двумя путями.

- **Прямой интерфейс**. Одноэтапный процесс сканирования и загрузки: сканер просто выявляет в системе-источнике метаданные, которые нужно загрузить, а затем напрямую вызывает компонент загрузки данных соответствующего формата, который и загружает выявленные метаданные в репозиторий. Никаких промежуточных форматированных файлов выгрузки/загрузки метаданных не создается; всё делается за один прием.
- **Опосредованный интерфейс**. Двухэтапный процесс: сканер собирает метаданные из системы-источника и выгружает их в промежуточный файл данных установленного формата, который затем считывается и обрабатывается службой сбора данных хранилища метаданных. Подобные интерфейсы присущи более открытым архитектурам, поскольку допускают различные методы формирования, считывания и обработки файлов переноса данных.

В процессе сканирования создаются и используются служебные файлы нескольких типов:

- **Контрольный файл** с информацией о модели данных системы-источника.
- **Файл для повторного использования** с правилами обработки, которые можно использовать повторно при осуществлении загрузок.
- **Файлы журналов**, создающиеся при выполнении всех процедур на каждом шаге обработки каждого цикла сканирования и/или загрузки.
- **Временные файлы или файлы с резервными копиями**, используемые при выполнении процесса как вспомогательные, а также для отслеживания изменений и/или обеспечения аварийного восстановления.

Для хранения временных и резервных файлов используйте область временного хранения (staging area). Эта область обеспечивает возможность как восстановления, так и отслеживания изменений в источниках, повлекших проблемы с качеством метаданных. В зависимости от архитектуры буферная область может быть реализована в виде отдельного каталога файлов или базы данных.

Для интеграции метаданных обычно вполне подходят средства управления хранилищами данных и BI-приложениями (см. главы 8 и 11).

2.4.2 Распространение и доставка метаданных

Для доставки метаданных потребителям данных (пользователям, приложениям или инструментам), которые нуждаются в их регулярном обновлении, могут использоваться следующие каналы, средства и механизмы:

- веб-сайты метаданных во внутренней сети, поддерживающие функции навигации, поиска, обработки запросов, формирования отчетов и анализа;
- отчеты, глоссарии и другие документы;
- хранилища и витрины данных, инструменты бизнес-аналитики (BI);
- средства моделирования данных и программирования приложений;
- обмен сообщениями и транзакционными данными;
- веб-сервисы и API;
- интерфейсные решения сторонних организаций (например, решения по поддержке цепочек поставок).

Решение по управлению метаданными нередко реализуется в связке с BI-решением, что обеспечивает полноту и своевременность синхронизации метаданных с BI-контентом, а также возможность включения метаданных в результаты бизнес-аналитики, выдаваемые конечным пользователям. Аналогичным образом и некоторым системам управления взаимоотношениями с клиентами (CRM), а также другим компонентам систем планирования ресурсов предприятия (ERP) может требоваться интеграция метаданных на уровне доставки.

Обмен метаданными с внешними организациями осуществляется через файлы (в виде плоских файлов, а также файлов XML или JSON) или веб-сервисы.

2.5 Применение метаданных в аналитике и при формировании запросов и отчетов

Метаданные задают направления использования информационных активов. Необходимо в полной мере использовать метаданные в бизнес-аналитике (включая формирование статистических и аналитических отчетов), принятии решений (оперативных, тактических и стратегических) и изучении бизнес-семантики (без знания бизнес-терминологии невозможно понять смысл бизнеса). Репозиторий метаданных обязательно должен иметь портал или клиентское приложение для пользователей с поддержкой функций поиска и извлечения любых метаданных, которые могут потребоваться для вышеописанных потребностей, изучения накопленных данных и управления ими. Можно предусмотреть дифференцированные наборы интерфейсов для бизнес-пользователей, технических пользователей и разработчиков, поддерживающие только те функции, которые требуются представителям каждой из названных групп. Для планирования

развития и анализа потенциальных последствий изменений нужны одни отчеты, для выявления и устранения рассогласования терминов, используемых в различных областях, — другие, для разрешения проблем с данными в хранилищах данных и проектах BI — третьи (например, отчеты о происхождении данных).

3. ИНСТРУМЕНТЫ

Основным средством управления метаданными является репозиторий метаданных. Он включает слой интеграции, а часто еще и интерфейс ручного обновления метаданных. Программные средства, производящие и использующие метаданные, становятся одновременно источниками и потребителями метаданных, интегрируемых в репозиторий.

3.1 Инструменты управления репозиторием метаданных

Инструментальные средства управления метаданными, поддерживающие все необходимые функции, реализуются в среде централизованного хранилища (репозитория) метаданных. Ввод метаданных может производиться вручную или посредством их извлечения из различных источников через специальные подключения. Репозитории метаданных также поддерживают функции обмена метаданными с другими системами.

Средства управления метаданными и сами репозитории служат также источниками метаданных, особенно при гибридной архитектурной модели метаданных или в средах крупных предприятий. Средства управления метаданными позволяют осуществлять обмен собранными метаданными с другими репозиториями метаданных, что делает возможным сбор и аккумуляцию разнообразных метаданных из множества разнородных источников в централизованном репозитории или, как альтернативный вариант, обогащение и стандартизацию метаданных в процессе обмена ими между узлами распределенной (сетевой) модели.

4. МЕТОДЫ

4.1 Отслеживание происхождения и анализ влияния

Ключевым преимуществом выявления и документирования метаданных, которые описывают все информационные активы организации, является получение исчерпывающих сведений о том, как именно преобразуются данные при перемещении между системами. Многие средства управления метаданными предоставляют информацию о том, что именно происходит с данными в их среде. Это обеспечивает возможность просмотра происхождения (lineage) данных при их продвижении через системы и приложения. Текущую версию последовательности преобразований, получаемую по результатам анализа программного обеспечения, называют «происхождением

«как реализовано»» (As Implemented Lineage), и она может отличаться от «происхождения «как спроектировано»» (As Designed Lineage), определяемого спецификациями мэппинга данных, зафиксированными в проектной документации.

Возможности отслеживания происхождения ограничиваются объемом и составом данных о преобразованиях на стороне приложений, имеющихся в системе управления метаданными. Функционально-ориентированные репозитории часто оснащены средствами визуализации метаданных, которые позволяют получать исчерпывающую информацию о преобразованиях данных, но лишь в среде репозитория, поскольку эти средства изолированы от всего, что происходит с данными в иных средах до поступления в среду репозитория.

Системы управления метаданными импортируют происхождения «как реализовано» из программных средств, а затем дополняют их происхождениями «как спроектировано» для тех источников, из которых данные о фактической реализации преобразований получить невозможно. Процесс составления связной картины из собранных элементов происхождения данных называют *сшиванием* (*stitching*). В результате получается целостное преставление о перемещении данных от мест их первоначального хранения (официальных источников или систем записи) до конечных пунктов назначения.

Рисунок 89 содержит простейший пример описания происхождения элемента данных. Расшифровывается оно следующим образом: элемент бизнес-данных «Итого сумма заказа», физически реализованный как столбец zz_total, зависит от трех других элементов данных, а именно: «Цена за шт.» (столбец yy_unit_cost физической модели), «Налог штата» (yy_tax) и «Заказано (шт.)» (yy_qty).

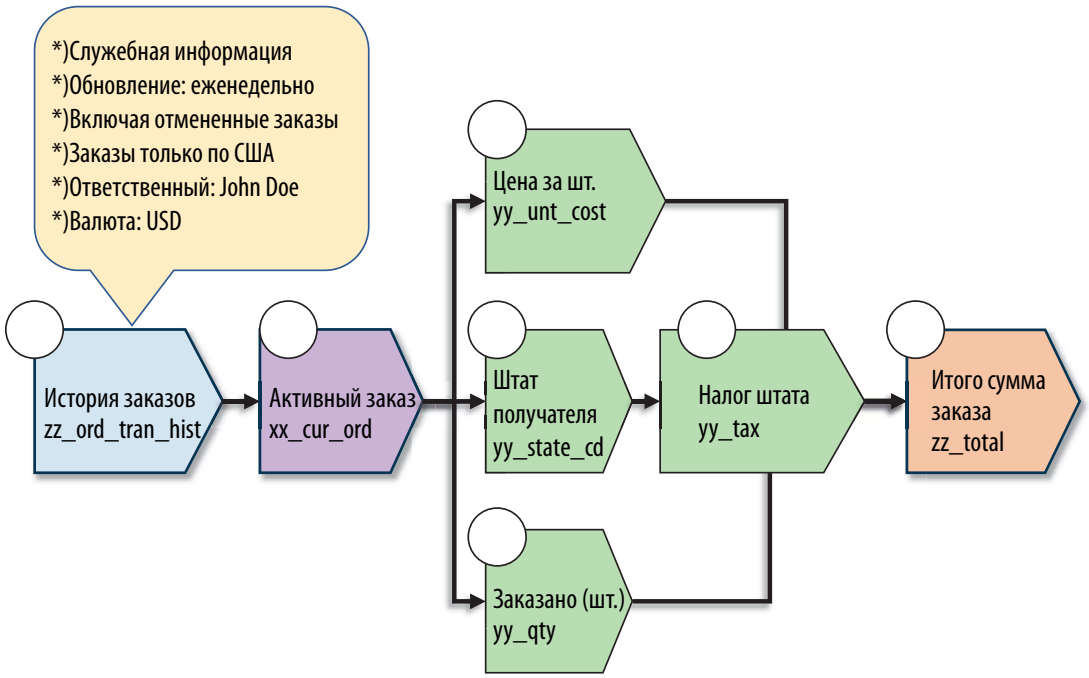


Рисунок 89. Пример схематического представления происхождения элемента данных

При всей кажущейся доходчивости графических схем отображения происхождения элементов данных (наподобие той, что представлена на рис. 89) они бывают понятны далеко не всем бизнес-пользователям, особенно если учесть, что на практике их структура оказывается значительно сложнее, чем в приведенном простом примере. Более высокоуровневые схемы (например, последовательности обработки данных системами — System Lineage) позволяют составлять обобщенное представление о движении данных на уровне систем или приложений. Многие средства визуализации поддерживают функции масштабирования (+/–), позволяющие переходить с уровня просмотра происхождения отдельного элемента на уровень потоков данных между системами, что существенно упрощает понимание происхождения элементов данных в общем контексте архитектуры систем. Рисунок 90 содержит пример наглядного представления последовательности обработки данных системами и/или приложениями.

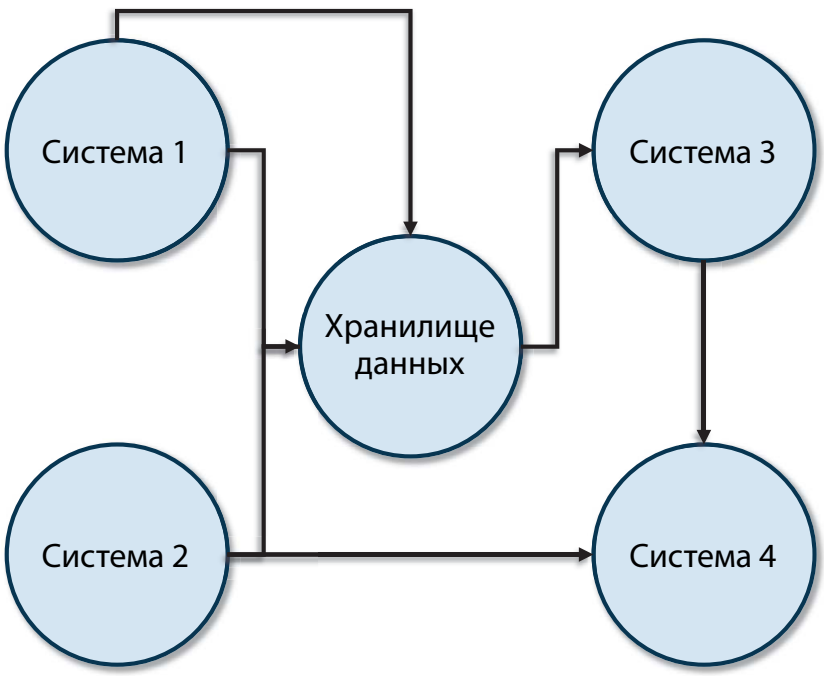


Рисунок 90. Пример схемы потоков данных на уровне систем

С ростом числа элементов данных в системах выявлять их происхождение и управлять потоками данных становится всё сложнее. В целях успешного достижения бизнес-целей необходима тщательно продуманная стратегия и оперативные планы выявления и импорта в репозиторий всех необходимых метаданных. Успешное выявление происхождения данных требует учета как бизнес-потребностей, так и технических особенностей систем.

- **Бизнес-аспекты.** Ограничьте раскрытие происхождения лишь важными с точки зрения бизнеса элементами данных и совместно с ответственными за различные направления

деятельности расставьте их в порядке приоритетности. Затем отследите в обратном направлении маршрут, по которому каждый элемент данных попадает в целевую систему, до системы или приложения-первоисточника. Ограничив сканируемые ресурсы лишь теми, которые реально участвуют в перемещении, передаче или обновлении выбранных элементов данных, вы поможете потребителям бизнес-данных лучше понять, что именно происходит с каждым элементом при прохождении через системы, прежде чем он попадает к ним в том виде, который им привычен. А в сочетании с результатами измерений показателей качества данных задокументированное происхождение помогает отыскивать точки негативного влияния плохо спроектированных процессов на качество данных.

- **Технические аспекты.** Начните с систем-источников и выявите всех первичных потребителей, а затем всех последующих потребителей первого изучаемого набора данных, затем второго, третьего и так далее, пока не выявите все системы, которые их обрабатывают или используют. Пользователи из числа технологов могут почерпнуть много полезного из стратегии раскрытия системного закулисья и получить ответы на различные вопросы об интересующих их данных. Такой подход позволит и техническим, и бизнес-пользователям самостоятельно исследовать происхождение различных элементов данных в масштабах предприятия, получая ответы на вопросы типа «Откуда берется номер карты социального страхования?», и генерировать отчеты о последствиях гипотетических изменений, например: «Системы, требующие перенастройки в случае изменения разрядности данных в столбце N». Такая стратегия, однако, при всей ее практической полезности может оказаться весьма сложной в реализации и управлении.

Многие средства интеграции данных включают инструменты анализа происхождения не только данных, накопленных в хранилище, но и на уровне моделей данных, а также на уровне физической базы данных в целом. Некоторые даже предлагают бизнес-пользователям возможность мониторинга и обновления определений данных через веб-интерфейс, вследствие чего метаданные всё больше уподобляются онлайновым бизнес-глоссариям.

Задокументированное происхождение помогает использовать данные и бизнес-пользователям, и техническим специалистам. Без него масса времени тратилась бы на расследование причин аномальных результатов, моделирование потенциальных последствий изменений или устранение реальных последствий изменений, произведенных без гарантии положительного результата. Поэтому лучше изыскать возможности для разработки собственного или внедрения коммерческого интегрированного решения, поддерживающего анализ последствий изменений в комплексе с учетом происхождения данных и позволяющего разобраться во всех деталях и механизмах движения данных на всех этапах, начиная с загрузки в систему и заканчивая выдачей отчетов и аналитики конечным пользователям. Отчеты с результатами факторного анализа зависимостей позволяют очертить круг компонентов, которые будут затронуты потенциальными изменениями, и оперативно спланировать задачи по их оптимизации, доработке и последующему эксплуатационному сопровождению.

4.2 Метаданные для обработки больших данных

Многие специалисты по управлению данными привычно и профессионально справляются с любыми задачами, возникающими в процессе работы с традиционными хранилищами структурированных данных, где каждый элемент четко идентифицирован и промаркирован. В наши дни, однако, множество данных поступает в менее структурированных форматах. Какие-то источники неструктурированных данных находятся внутри организации, какие-то вне ее. В любом случае необходимость физически собирать сами данные в централизованное хранилище отпала. Новые технологии позволяют программам обращаться к данным дистанционно, не требуя их переноса в операционную среду программ, что способствует радикальному сокращению потоков данных и не менее радикальному повышению скорости их обработки. Тем не менее успешное управление данными в озерах данных зависит от способности эффективно управлять этими метаданными.

Маркировка принимаемого озером данных контента тегами метаданных должна производиться сразу же, чтобы впоследствии данные были доступны. Многие программные средства приема данных (data ingestion) поддерживают их профилирование. В процессе профилирования определяются предметные области, связи и проблемы с качеством. Также возможна маркировка контента тегами. При приеме теги метаданных могут использоваться, к примеру, для маркировки выявленной чувствительной или конфиденциальной информации (в частности, персональных данных). Специалисты в области науки о данных (data scientists) могут добавлять доверительные интервалы, тестовые идентификаторы, коды кластеров с различным поведением и т. п. (см. главу 14).

5. РЕКОМЕНДАЦИИ ПО ВНЕДРЕНИЮ

Среду управляемых метаданных следует внедрять поэтапными приращениями, что позволяет минимизировать риски для организации и упростить восприятие новшеств сотрудниками. Репозитории метаданных лучше реализовывать на базе открытых платформ реляционных баз данных. Такой подход позволяет максимально гибко разрабатывать и внедрять дополнительные механизмы управления и интерфейсы обмена данными по мере выявления необходимости их включения в первоначальный проект репозитория.

Контент репозитория метаданных должен быть предельно обобщенным, а не слепо воспроизводить структуру данных в системах-источниках. Планируйте структуру информационного наполнения репозитория в тесной координации с экспертами в предметных областях и в соответствии с комплексной моделью метаданных предприятия. Планирование должно вестись с учетом необходимости обеспечения интеграции метаданных на уровне, позволяющем потребителям данных просматривать и выбирать различные источники данных и получать к ним доступ. Чем полнее будет реализовано это требование, тем ценнее будет репозиторий с функциональной точки зрения. Помимо текущей версии, в репозитории должны содержаться планируемая к выпуску и предыдущие версии метаданных.

Зачастую первая реализация носит экспериментальный характер и нужна для проверки и доработки спроектированной среды метаданных и средств управления ими. В дальнейшем неизбежна интеграция проектов управления метаданными в общую методологию проектирования систем, и лучше озаботиться этим на ранних стадиях. На практике же порядок внедрения управления метаданными может весьма серьезно варьироваться в зависимости от архитектуры и типов существующих хранилищ данных.

5.1 Оценка готовности / Оценка рисков

Надежная стратегия управления метаданными полезна всем, поскольку помогает принимать эффективные решения. Прежде всего нужно донести до понимания сотрудников, насколько рискованно вовсе не управлять метаданными. Просто оцените все риски, проистекающие от отсутствия качественных метаданных, и их возможные последствия:

- ошибочные суждения вследствие неверных, неполных или устаревших исходных предположений или незнания истинного контекста данных;
- огласка или утечка чувствительных данных, ставящая под удар репутацию, безопасность и благополучие клиентов или сотрудников, подрывающая доверие к бизнесу и подводящая его под штрафные санкции или судебные издержки;
- в случае увольнения экспертов в предметных областях — риск лишиться всех знаний, которыми они располагают.

Принятие организацией надежной стратегии в отношении метаданных начинает приносить плоды в плане снижения описанных выше рисков, начиная с первых же шагов по ее осуществлению. Формальной оценкой степени готовности организации противостоять этим рискам как раз и является уровень зрелости текущей реализации комплекса мер по управлению метаданными. Экспертной оценке подлежит состояние всех важных для бизнеса категорий и элементов данных, имеющихся глоссариев метаданных, процессов составления происхождения и профилей данных, обеспечения и контроля качества данных, управления основными данными и других аспектов распоряжения данными организации. Результаты экспертизы вкупе с бизнес-приоритетами берутся за основу стратегического плана совершенствования практики управления метаданными. Формальная оценка служит также и важным компонентом бизнес-обоснования, позволяющим заручиться поддержкой и финансированием со стороны высшего руководства.

Стратегия управления метаданными должна являться неотъемлемой частью общей стратегии распоряжения данными организации, а при изначальном отсутствии таковой может стать вполне подходящим первым шагом на пути перехода к практике согласованного высокоуровневого управления. Экспертиза имеющихся метаданных и текущего состояния управления ими дополняется выводами из бесед с ключевыми заинтересованными лицами и заключениями с результатами оценки рисков, — и весь этот пакет документов учитывается при выработке стратегии и дорожной карты развития управления метаданными.

5.2 Организационные и культурные изменения

Инициативы по внедрению и совершенствованию общеорганизационной системы управления метаданными, как и все прочие усилия по налаживанию управления данными, часто наталкиваются на неприятие и даже сопротивление в силу их непривычности и несовместимости со сложившейся организационной культурой. Главная проблема обычно заключается в том, что переход из среды с неуправляемыми метаданными в среду с управляемыми метаданными требует труда и дисциплины, а такая работа над собой дается людям непросто, даже если они и понимают всю степень важности и ценности надежных метаданных. Готовность организации — еще один повод для серьезной озабоченности, поскольку управление метаданными требует продуманного и методичного надзора и контроля.

Во многих организациях управление метаданными котируется довольно низко по шкале приоритетов. Проработка исходного набора метаданных требует скоординированных и целенаправленных усилий в масштабах организации. В зависимости от профиля таким набором может являться структура анкетных данных сотрудников, серий и номеров полисов страхования, VIN и регистрационных номеров транспортных средств, спецификаций продуктов, и т. д. и т. п. Тут важно понимание того факта, что любое изменение структуры подобных метаданных автоматически требует масштабной переработки множества информационных систем предприятия. Поэтому постарайтесь выбрать для затравки такой набор метаданных, чтобы его постановка под контроль принесла быстрые и осязаемые плоды в плане повышения эффективности и/или прибыльности работы компании за счет повышения качества данных. В качестве подкрепляющей аргументации можно использовать конкретные позитивные примеры из практики других компаний или учреждений, работающих в вашей отрасли.

Реализация стратегии распоряжения данными предприятия немыслима без поддержки и заинтересованности высшего руководства. Кроме того, требуется тесное сотрудничество поверх функциональных барьеров между бизнес-подразделениями и технологами.

6. РУКОВОДСТВО МЕТАДАННЫМИ

Организациям следует самостоятельно определять специфические требования к управлению метаданными на протяжении их жизненного цикла, а также механизмы руководства, обеспечивающие выполнение этих требований. Рекомендуется определять роли и должностные обязанности лиц, несущих формальную ответственность за целевое использование выделенных ресурсов, особенно на крупных или критически важных участках работы. Процессы руководства метаданными сами по себе требуют надежных метаданных, поэтому вполне можно реализовать подход, при котором команда по руководству будет обкатывать принципы создания и использования метаданных для начала на собственной практике.

6.1 Механизмы контроля процессов

Команда по руководству данными должна нести полную ответственность за определение стандартов метаданных и управление изменением их статуса в организации. Часто для этого используются программные средства управления потоками работ или совместной работой. Этой же команде может быть поручена организация информационно-разъяснительной работы, разработка программ, а возможно, и проведение курсов переподготовки в масштабах организации.

Более зрелое руководство метаданными потребует внедрения практики определения терминов посредством их последовательного многоуровневого согласования с поэтапным повышением в статусе, например: от проекта термина — к согласованию — к утверждению — к публикации — к использованию — к замене или удалению. Кроме того, команде по руководству также можно поручить управление взаимосвязями терминов, а также их классификацией и объединением в группы.

Интеграция стратегии управления метаданными с жизненным циклом разработки систем (SDLC) — мера исключительно важная, поскольку лишь так можно обеспечить своевременный учет системой управления метаданными фактически состоявшихся изменений в составе и структуре, терминах и определениях метаданных, используемых в организации. Такая интеграция гарантирует постоянную актуальность используемого набора метаданных.

6.2 Документация, описывающая метаданные

Следует вести основной каталог метаданных (master catalog of metadata) с указанием источников и получателей данных на уровне элементов. Этот незаменимый как для бизнес-пользователей, так и для ИТ-специалистов справочный ресурс должен публиковаться в открытом для всех пользователей доступе и служить не просто путеводителем по системам, подсказывающим «что где лежит», но и информировать о том, что именно там находится, как и для чего используется описываемый компонент или элемент метаданных, включая следующее:

- текущий статус метаданных;
- источник метаданных;
- получатели метаданных;
- расписание или даты последнего и следующего обновлений;
- сроки хранения и список доступных архивных версий;
- краткое описание контента;
- оценка качества и предупреждения (например, об отсутствующих значениях);
- текущий статус в системе регистрации или источнике, например:
  - доступны данные с DDMMYY по настоящее время;
  - сбор данных прекращен;
  - элемент заменен на [имя элемента];

- программные средства, элементы архитектуры и лица, участвующие в сборе, обработке или использовании;
- чувствительная информация, имеющаяся в источнике, и способ ее защиты при выдаче данных (изъятие, маскировка, подмена и т. п.).

В области управления документами и контентом примерно такую же информацию отражают карты данных. Визуальные представления ландшафта интеграционных решений также относятся к документации, описывающей метаданные (см. главу 9).

6.3 Стандарты и руководства

Стандарты метаданных требуют соблюдения хотя бы ради обеспечения технической возможности обмена данными с деловыми партнерами. Компании вполне отдают себе отчет в ценности предоставления достоверной информации о себе клиентам, поставщикам, партнерам и регулирующим органам. А для правильного понимания и полноценного использования распространяемой информации потребителями она должна укладываться в рамки общепринятых определений, вследствие чего и появились в изобилии отраслевые стандарты метаданных.

Учитывать отраслевые и специализированные стандарты метаданных нужно уже на ранней стадии планирования, поскольку это способствует развитию технологий управления метаданными. Обычно ведущие разработчики программного обеспечения предусматривают поддержку множества разнообразных стандартов выпускаемыми ими продуктами, а некоторые еще и предлагают техническое содействие с конфигурированием поставляемых решений с учетом отраслевой или узкопрофильной специфики.

Коммерческое ПО обычно обеспечивает поддержку обмена данными через интерфейсы XML, JSON и/или REST. При этом все производители используют одну и ту же стратегию компоновки и объединяют предлагаемые программные средства в пакеты решений. Технологии интеграции данных, управления реляционными и многомерными базами данных, управления требованиями, создания отчетов в сфере BI, моделирования данных и управления бизнес-правилами обеспечивают возможности импорта/экспорта данных и метаданных с использованием языка XML. Форматы XML, а также определения типов документов (DTD) или чаще схем XML (XSD) создаются самими разработчиками и надежно защищаются, поскольку относятся к объектам интеллектуальной собственности. Доступ к ним возможен через поставляемые теми же разработчиками фирменные интерфейсы. В связи с этим для интеграции подобных программных средств в существующую среду управления метаданными требуются доработки систем на стороне клиента.

Руководства включают шаблоны, дополняемые примерами их использования. Кроме того, предлагаются курсы обучения, из которых можно получить как информацию о правилах использования, так и полезные советы наподобие «не определяйте один термин через другой». Различные шаблоны ориентированы на определенные типы метаданных, а состав шаблонов может варьироваться в зависимости от выбранного вами решения, — поэтому, прежде чем приобретать

ПО, выясните, имеется ли у разработчика в составе продукта или в качестве дополнительной опции подходящий для вашей организации шаблон. Текущий мониторинг публикуемых инструкций и рекомендаций по обеспечению эффективности и контроль своевременной установки обновлений также входят в круг обязанностей команды по руководству данными.

Стандарты ISO в области метаданных предлагают рекомендации для разработчиков инструментов, поэтому их вряд ли будут подробно изучать организации, реализующие управление метаданными с помощью коммерческого ПО, поскольку все требования стандартов в нем уже учтены. Тем не менее следует хотя бы ознакомиться со стандартами и понять их требования, а также последствия их несоблюдения.

6.4 Метрики

Всю степень значимости метаданных проще всего понять через оценку отрицательного влияния их отсутствия. В части оценки риска измерьте затраты времени потребителей данных на поиск и изучение сопроводительной информации о них — и поймете, насколько с появлением метаданных улучшится ситуация с точки зрения сокращения непродуктивных человеко-часов. Эффект от внедрения централизованной модели метаданных можно оценить также и по показателю полноты самих метаданных, числу процессов управления, в которых применяются метаданные, и частоте обращений к метаданным пользователей и процессов. Ниже перечислены рекомендуемые для любой среды метаданных метрики.

- **Полнота метаданных, представленных в центральном репозитории**, позволяет сравнивать текущую ситуацию с идеальной, когда централизованно управляется 100% метаданных предприятия (все описания объектов и все определения элементов данных). Сведения о требуемых определениях можно почерпнуть из стратегии в области работы с метаданными.
- **Зрелость управления метаданными**: метрики, разработанные для оценки зрелости управления метаданными, основанные на модели оценки зрелости возможностей (CMM-DMM, см. главу 15).
- **Распоряжение метаданными**: приверженность организации управлению метаданными находит отражение в назначении соответствующих распорядителей, деятельность которых должна охватывать всю организацию, а также в наличии документации с описанием ролей и проводимых работ.
- **Использование метаданных**: популярность репозитория метаданных среди пользователей вполне объективно определяется статистикой обращений. Сложнее отслеживать интенсивность использования метаданных потребителями в повседневной практике ведения бизнеса. Не исключено, что тут придется проводить опросы или более подробные обследования.
- **Использование бизнес-глоссария**: число обращений, обновлений и устраненных противоречий; полнота.
- **Соответствие сервисов основных данных требованиям по повторному использованию**: показатель повторного использования имеющихся сервисов основных данных в SOA-решениях.

Метаданные о сервисах данных помогают разработчикам выявлять возможности повторного использования существующих сервисов в разрабатываемых приложениях.

- **Качество документации, описывающей метаданные**, может оцениваться как на основе автоматизации, так и вручную. Автоматизированные методы включают попарную проверку источников на противоречия и совпадения, а также отслеживание тенденций изменения со временем количества выявленных проблем. Также автоматически можно фиксировать и процент атрибутов, имеющих определение, и тенденцию изменения этого показателя. Ручные методы контроля качества документации включают полные или выборочные обследования, в рамках которых структура и содержание рассматриваемых аспектов соответствуют принятым в организации определениям критериев качества. Стандартные измеримые показатели качества позволяют судить о полноте, надежности, актуальности и т. п. метаданных, представленных в репозитории.
- **Доступность репозитория метаданных** (% времени):
  - для пакетной обработки;
  - для обработки пользовательских запросов.

7. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Aiken, Peter. *Data Reverse Engineering: Slaying the Legacy Dragon*. 1995.

Foreman, John W. Data Smart: *Using Data Science to Transform Information into Insight*. Wiley, 2013. Print.

Loshin, David. *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann, 2001.

Marco, David. *Building and Managing the Meta Data Repository: A Full Lifecycle Guide*. Wiley, 2000. Print.

Milton, Nicholas Ross. *Knowledge Acquisition in Practice: A Step-by-step Guide*. Springer, 2007. Print. Decision Engineering.

Park, Jung-ran, ed. *Metadata Best Practices and Guidelines: Current Implementation and Future Trends*. Routledge, 2014. Print.

Pomerantz, Jeffrey. *Metadata*. The MIT Press, 2015. Print. The MIT Press Essential Knowledge ser.

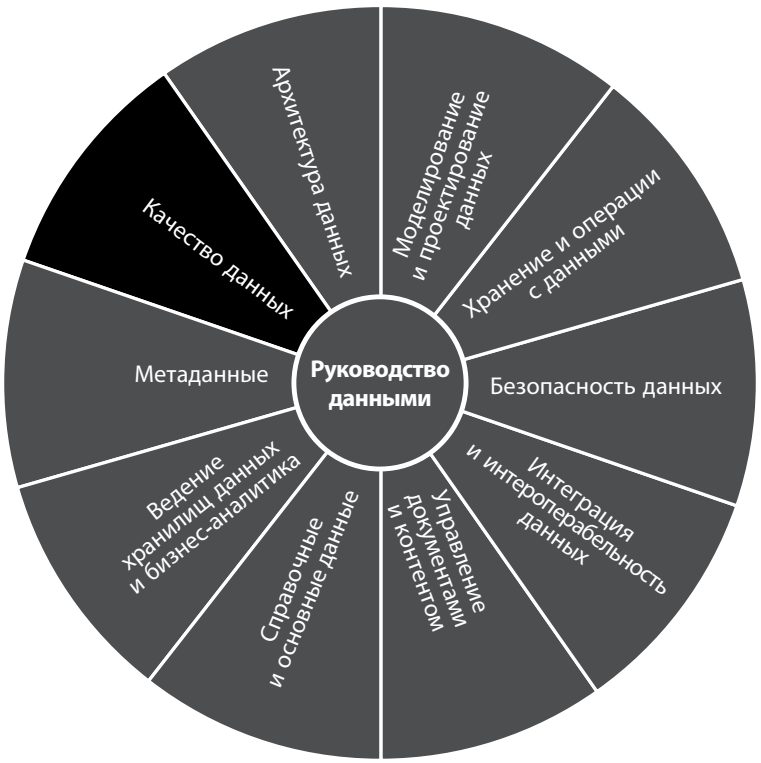
Schneier, Bruce. *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*. W. W. Norton and Company, 2015. Print.

Tannenbaum, Adrienne. *Implementing a Corporate Repository: The Models Meet Reality*. Wiley, 1994. Print. Wiley Professional Computing.

Warden, Pete. *Big Data Glossary*. O'Reilly Media, 2011. Print.

Zeng, Marcia Lei and Jian Qin. *Metadata*. 2nd ed. ALA Neal-Schuman, 2015. Print.

Качество данных



DAMA-DMBOK2 Data Management Framework

1. ВВЕДЕНИЕ

Эффективное управление данными подразумевает наличие в организации комплекса тщательно структурированных и согласованных взаимосвязанных процессов, которые позволяют задействовать данные на благо организации в соответствии со стратегическими целями. Управление данными включает разработку проектных решений по составу и структуре информационных массивов приложений, поддержку надежного хранения и безопасного доступа к данным, осуществление их целевого распространения, извлечение уроков из использования и, конечно же, обеспечение соответствия данных потребностям бизнеса. При этом все заявления о ценности данных относятся только к данным надежным и достоверным, то есть данным высокого качества.

Имеется, однако, множество факторов, способствующих притоку непригодных для использования данных, подрывающих веру в ценность данных как таковых, в том числе: недопонимание губительных для организации последствий работы с данными низкого качества; неумелое планирование; архитектурная обособленность систем; рассогласованность проектов; неполнота документации; недостаточная стандартизация и недостаточный уровень руководства данными. Многие организации не в состоянии даже отличать подходящие данные от негодных.

Все дисциплины (области) управления данными вносят свой вклад в обеспечение качества данных; более того, получение высококачественных данных, которые позволяли бы организации успешно решать поставленные задачи, должно быть целью этих дисциплин. Поскольку к появлению некондиционных данных способны привести невежественные решения или действия любого лица, работающего с информационными ресурсами, обязательным условием получения качественных данных является кросс-функциональное взаимодействие и согласованность предпринимаемых мер по контролю качества. Организациям и отдельным командам следует об этом помнить и планировать мероприятия по обеспечению качества в рамках реализуемых процессов с учетом всех выявленных рисков появления некондиционных данных, включая обусловленные человеческим фактором.

Поскольку ни одна организация не может похвастаться безупречностью технологических и бизнес-процессов, а также практик управления данными, проблемы с качеством данных неизбежны. Однако в организациях, где реализована формальная система управления качеством данных, проблемы возникают реже и решаются проще, чем в организациях, где качество данных — дело случая.

Формальное управление качеством данных выстраивается по аналогии с непрерывным управлением качеством других продуктов. Качество данных контролируется на всех фазах их жизненного цикла посредством определения стандартов и встраивания механизмов обеспечения и контроля их соблюдения в процессы создания, преобразования и хранения данных, включая наборы измеримых показателей соответствия данных стандартам качества на всех этапах. Для внедрения комплексного подхода к обеспечению качества данных обычно требуется команда по реализации программы качества данных (команда программы качества данных — Data Quality program team). Команда программы качества данных отвечает за привлечение к участию и координацию действий профессионалов в области управления данными со стороны бизнеса и технических подразделений при проведении работ, обеспечивающих последовательное применение методов, которые гарантировали бы пригодность любых данных для использования по назначению. Команде программы качества данных, вероятно, потребуется принять участие в серии проектов, прежде чем они смогут внедрить в организации устойчивые процессы с использованием передовых практик непрерывного управления качеством данных. Параллельно должны приниматься экстренные меры по устранению неотложных проблем.

Поскольку управление качеством данных предполагает управление их жизненным циклом, программа качества данных неизбежно накладывает определенные требования и ограничения на использование данных и предусматривает ответственность за обеспечение их соблюдения при

Рисунок 91.
Контекстная диаграмма: качество данных

осуществлении операционной деятельности. В обязанности участников команды программы качества данных может входить, например, составление отчетности об уровнях качества данных; участие в анализе данных и сборе статистики; выявление и приоритизация проблем с данными. Кроме того, команда программы качества отвечает за взаимодействие с потребителями данных при решении вопросов, касающихся обеспечения их потребностей, а с теми, кто задействован в создании, обновлении или удалении данных, — вопросов обеспечения соблюдения правил обращения с данными. Качество данных зависит от всех, кто с ними работает, а не только от профессионалов в области управления данными.

Так же как руководство и управление данными в целом, управление качеством данных осуществляется именно как систематическая программа, а не разовый проект. При этом программа качества данных включает и работы, которые проводятся на проектной основе, и плановую деятельность по сопровождению информационных систем и ресурсов, а также обеспечение эффективных коммуникаций и обучение. Самое главное — помнить о том, что долгосрочный успех программы качества данных, то есть обеспечение устойчиво высокого качества данных, зависит от того, удастся ли побудить организацию к изменению культуры и привить людям образ мышления, ориентированный на качество. Как сказано в *«Лидерском манифесте о данных»*, «фундаментальные и устойчивые изменения требуют лидерства и приверженности руководства, помноженных на вовлечение всех без исключения сотрудников на всех уровнях организации. Людей, использующих данные для выполнения своей прямой работы, — а таких в большинстве организаций очень высокий процент, — нужно сделать драйверами изменений. Первоочередное внимание при планировании и проведении изменений должно уделяться тому, как организация управляет своими данными и повышает их качество»[1].

1.1 Бизнес-драйверы

Бизнес-драйверы, обусловливающие необходимость наличия формализованной программы качества данных, включают:

- повышение ценности данных и информационных ресурсов организации и реальной отдачи от их использования;
- снижение рисков и издержек, обусловленных низким качеством данных;
- повышение эффективности и производительности в масштабах организации;
- защиту и укрепление репутации организации.

Организации, стремящиеся получать полноценную отдачу от имеющихся данных, безусловно сознают, что высококачественные данные ценнее данных низкого качества. К тому же некачественные данные обременены серьезными рисками и чреваты ущербом репутации, штрафами, упущенной прибылью, оттоком клиентов и негативными отзывами в СМИ (см. главу 1). Обеспечивать

[1] Полный текст «Лидерского манифеста о данных» (*The Leader's Data Manifesto*) см.: http://bit.ly/2sQhcy7. — *Примеч. пер.*

высокое качество данных нередко предписывают также нормативно-правовые документы и отраслевые регламенты. Наконец, некачественные данные влекут за собой и всевозможные прямые убытки. Вот некоторые примеры негативных последствий:

- ошибки в выставленных счетах;
- увеличение числа обращений в службу поддержки клиентов при одновременном снижении способности разрешать возникшие проблемы;
- упущенные возможности и, как следствие, падение оборота и выручки;
- задержка интеграции в процессе слияний и поглощений;
- повышенная уязвимость перед угрозой мошенничества, злоупотреблений и т. п.;
- убытки вследствие ошибочных бизнес-решений, сделанных на основе неверных данных;
- потеря бизнеса и/или клиентуры из-за неспособности подтвердить свою репутацию и/или кредитоспособность.

Однако высокое качество данных — не самоцель, а средство обеспечения организационного успеха. Достоверные данные не только снижают риски и издержки, но и повышают эффективность. Работая с надежными данными, сотрудники более оперативно и согласованно находят ответы на текущие вопросы и тратят меньше времени на поиск нужной информации и оценку ее пригодности, что оставляет им больше времени на глубокое осмысление данных с целью взвешенного принятия решений и качественного обслуживания клиентов.

1.2 Цели и принципы

Программы качества данных преследуют следующие цели.

- Выработка управляемого подхода к обеспечению соответствия данных нуждам их потребителей.
- Определение стандартов и спецификаций механизмов контроля качества данных как составной части жизненного цикла данных.
- Определение и внедрение процессов измерения, мониторинга и учета уровня качества данных.
- Выявление и поддержка использования возможностей по повышению качества данных посредством внесения изменений в системы и процессы, а также осуществление деятельности по проведению измеримых улучшений качества данных на основе требований их потребителей.

Программы качества данных должны быть ориентированы на соблюдение следующих принципов.

- **Критичность.** Чем критичнее данные для организации и ее клиентов, тем больше внимания должно им уделяться в рамках программы качества данных. Приоритизация мероприятий по совершенствованию данных производится на основе учета критичности данных и уровня риска, возникающего в случае использования некорректных данных.

- **Управление жизненным циклом.** Качество данных должно контролироваться и обеспечиваться с момента их создания или получения вплоть до ликвидации. Все промежуточные процессы обработки данных внутри систем и переноса данных из системы в систему, естественно, также должны контролироваться программой качества данных (то есть каждое звено в цепи передачи данных должно гарантированно обеспечивать высокое качество выходных данных).
- **Предупреждение.** Основное внимание должно уделяться предупреждению ошибок; исправление ошибочных записей — мера обязательная, но вторичная по отношению к профилактике их появления.
- **Устранение первопричин.** Из вышесказанного следует, что проблемы с качеством данных необходимо искать и устранять на уровне корневых причин, а не симптомов. Поскольку исходные ошибки часто отыскиваются на уровне алгоритмов, процессов или архитектуры систем, для повышения качества данных нередко приходится изменять технологические процессы, перепрограммировать приложения или перенастраивать системы.
- **Руководство.** Деятельность по руководству данными должна быть направлена на создание данных высокого качества, а меры, предпринимаемые в рамках программы качества данных, — содействовать формированию и поддержанию управляемой среды данных.
- **Управление на основе стандартов (standards-driven).** Все заинтересованные стороны на протяжении жизненного цикла данных предъявляют те или иные требования к их качеству. По мере возможности все эти требования должны обобщаться на уровне четко определенных стандартов и соответствующих им измеряемых показателей, позволяющих оценивать качество данных.
- **Объективность измерений и прозрачность.** Уровни качества данных должны измеряться объективно и согласованно, а методология измерения и оценки показателей качества — доводиться до сведения всех заинтересованных сторон, поскольку их мнение о качестве данных является особенно важным.
- **Встраивание в бизнес-процессы.** Владельцы бизнес-процессов должны следить за качеством данных, создаваемых в ходе этих процессов. Они же отвечают и за обеспечение соблюдения стандартов качества в рамках своих процессов.
- **Систематический контроль.** Владельцы информационных систем несут ответственность за систематический контроль соблюдения требований программы качества данных.
- **Включение в соглашения об уровне обслуживания.** Вопросы контроля качества данных и управления проблемными ситуациями должны быть отражены в соглашениях об уровне обслуживания (Service Level Agreements, SLA).

1.3 Основные понятия и концепции

1.3.1 Качество данных

Термин *качество данных* (*Data Quality, DQ*) распространяется как на характеристики, связанные с высоким качеством данных, так и на процессы измерения или повышения качества данных.

Следует разделять эти два варианта использования термина и пояснять, что понимается под данными высокого качества[1].

Данные можно считать высококачественными в той мере, в которой они соответствуют потребностям и ожиданиям потребителей. То есть данные обладают высоким или низким качеством, если они, соответственно, пригодны или непригодны к использованию по назначению. Следовательно, качество данных зависит от контекста и потребностей потребителей данных.

Одна из трудностей управления качеством данных заключается в том, что ожидания в отношении качества данных не всегда известны. Бывает, что потребители просто неспособны их четко сформулировать. А порой случается и так: люди, отвечающие за управление данными, вовсе не отдают себе отчета в том, что к этим данным могут быть применимы какие-то специфические требования. Однако для того, чтобы данные были надежными и достоверными, профессионалам в области управления данными нужно сделать всё возможное для наилучшего понимания требований клиентов к качеству данных и способов измерения степени соответствия данных этим требованиям. И делаться это должно в режиме постоянного обсуждения, поскольку требования к данным и качеству данных меняются не менее динамично, чем потребности и приоритеты бизнеса, зависящие, в свою очередь, от не менее переменчивых внешних сил и условий.

1.3.2 Критически важные данные

В большинстве организаций имеется множество данных различного уровня значимости. Ключевой принцип управления качеством данных — фокусировать усилия на улучшении ситуации с данными, которые имеют важнейшее значение для организации и ее клиентов. Это придает программе целенаправленность и способность напрямую и ощутимо влиять на удовлетворение потребностей бизнеса.

Оставив за скобками специфичные для различных отраслей признаки критичности данных, выделим общие для всех организаций характеристики, относящиеся к категории критически важных. К таковым относятся данные, требующиеся для:

- соблюдения нормативно-правовых требований;
- бухгалтерского учета и финансовой отчетности;
- управления бизнес-правилами;
- осуществления текущей деятельности;
- планирования бизнес-стратегии, в частности анализа конъюнктуры.

Основные данные (*master data*) относятся к критически важным по определению. Оценку остальных наборов или отдельных элементов данных на предмет их критичности можно

[1] В настоящей редакции DAMA-DMBOK2 мы постарались избегать употребления словосочетания «качество данных» без указания контекста. Использование таких словосочетаний, как, например, *данные высокого качества* или *данные низкого качества*, *работы по обеспечению качества данных* или *средства контроля качества данных*, способствует однозначности трактовки.

проводить, исходя из процессов-потребителей; характера отчетов, в которых фигурируют данные; финансовых, юридических и репутационных рисков, возникающих в случае проблем с качеством данных[1].

1.3.3 Измерения качества данных

Измерениями качества данных (*data quality dimension*) называют измеримые свойства или характеристики данных, находящиеся в прямой связи с их качеством. Термин *измерение* сразу же приводит к ассоциативной аналогии с мерами свойств физических тел (длина, ширина, высота и т. д.). Измерения качества данных служат также источником терминологии, используемой для определения требований к качеству данных. Их же можно использовать для описания результатов как первичной оценки, так и текущих измерений качества данных. Для оценки качества данных организации нужно определить такие измерения, которые одновременно важны для бизнес-процессов (и потому заслуживают рассмотрения) и поддаются объективной оценке. Измерения также служат базисной системой координат при определении правил оценки, которые, в свою очередь, напрямую соотносятся с потенциальными рисками, присущими критически важным процессам.

Например, неверные данные в поле e-mail клиента лишают нас способности информировать потенциальных покупателей о новой продукции и специальных предложениях, что приведет к снижению объемов продаж. Следовательно, нужно фиксировать данные о доле действительных адресов e-mail в профилях клиентов и совершенствовать работу клиентской службы до тех пор, пока не будет обеспечен уровень пригодности адресов e-mail на уровне, к примеру, не ниже 98% от общего числа профилей клиентов.

Многие ведущие теоретики качества данных предлагают в своих публикациях весьма тщательно проработанные наборы измерений[2]. Ниже описаны три наиболее авторитетных подхода, позволяющие более глубоко понять как значимость высококачественных данных, так и спектр возможностей для измерения и оценки качества данных.

В 1996 году Ричард Уанг и Дайана Стронг предложили модель оценки качества данных по следующим пятнадцати параметрам (измерениям) их восприятия потребителями[3].

- Качество данных как таковых:
  - точность;
  - объективность;

[1] Подробное описание подхода к обоснованию критичности данных см.: Jugulum (2014), главы 6 и 7.

[2] Помимо примеров, детально описанных ниже, имеется и множество других, предлагаемых в различных научных публикациях. Детальное обсуждение проблем определения параметров качества данных см. в работах: Loshin (2001), Olson (2003), McGilvray (2008), Sebastian-Coleman (2013); сравнительный анализ различных измерений см.: Myers (2013).

[3] Wang, R. Y., & Strong, D. M. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12 (4), 5, 1996. — *Примеч. пер.*

- убедительность;
- репутация источника.

- Контекстуальное качество данных:
  - полезность;
  - релевантность;
  - актуальность;
  - полнота;
  - достаточность объема данных.
- Репрезентативность:
  - интерпретируемость;
  - понятность;
  - логичность представления;
  - лаконичность представления.
- Наличие подтверждения качества данных:
  - доступность;
  - безопасность доступа и защита от несанкционированного доступа.

В том же 1996 году Томас Редман в книге «*Качество данных в информационную эру*»[1] сформулировал альтернативный набор параметров качества данных на основе структуры данных[2]. Редман определяет элемент данных как «триплетное представление»: значение атрибута (из области его значений) внутри сущности. Измерение при этом может относиться к различным аспектам данных: к модели (сущностям и атрибутам) или значениям данных. Далее Редман вводит понятие «измерение представления», определяя его как набор правил записи элементов данных. В общей сложности в трех этих общих категориях (модель данных, значения и представление данных) он описывает более двадцати измерений.

Модель данных

- Контент:
  - релевантность данных;
  - возможность получения значений;
  - четкость определений.
- Уровень детализации:
  - разбивка на атрибуты;
  - точность областей определения атрибутов.

[1] Thomas C. Redman. *Data Quality for the Information Age*. Artech House, 1996.

[2] В переработанное издание книги, опубликованное под названием «Качество данных: путеводитель» (Redman, 2001), вошла значительно доработанная и расширенная версия. — *Примеч. пер.*

- Состав:
  - естественность: смысл в том, чтобы каждому атрибуту соответствовал простой и понятный прототип в реальном мире, а также в выполнении правила «один атрибут — одно фактическое свойство объекта»;
  - однозначность идентификации: каждый объект модели должен явственным образом отличаться от любого другого объекта;
  - однородность;
  - минимально необходимая достаточность.
- Согласованность:
  - семантическая согласованность компонентов модели;
  - структурная согласованность объектов по всем типам сущностей.
- Реагирование на изменения:
  - устойчивость;
  - гибкость;

Значения данных

- Точность.
- Полнота.
- Актуальность.
- Непротиворечивость.

Представления данных

- Адекватность.
- Интерпретируемость.
- Переносимость.
- Соответствие формату.
- Гибкость формата.
- Допустимость пустых значений.
- Эффективное использование объемов памяти хранилищ.
- Соответствие физических экземпляров данных установленным форматам.

Редман отмечает, что согласованность сущностей, значений и представлений может достигаться путем определения и наложения ограничений. Виды ограничений при этом зависят от уровня согласования и типа согласуемых структур.

Ларри Инглиш в книге «*Улучшение хранилищ данных и качества деловой информации*» (English, 1999) подразделяет измерения качества данных на две обобщенные категории — неотъемлемые

и утилитарные качества[1]. Неотъемлемые характеристики не зависят от использования данных и являются константами. Утилитарные же качества ассоциируются с представлениями данных и зависят от того, каким образом они используются.

- **Неотъемлемые** характеристики качества:
  - соответствие определениям;
  - полнота значений;
  - обоснованность или соответствие бизнес-правилам;
  - соответствие данным в источнике;
  - соответствие действительности;
  - точность;
  - отсутствие дублирования;
  - эквивалентность избыточных или рассредоточенных элементов данных;
  - синхронизированность избыточных или рассредоточенных элементов данных.
- **Утилитарные** характеристики качества:
  - доступность;
  - своевременность обновления до актуального состояния;
  - контекстуальная ясность;
  - пригодность к использованию;
  - целостность истории происхождения;
  - корректность или соответствие фактам.

В 2013 году британским отделением Ассоциации управления данными (DAMA UK) опубликован аналитический доклад с описанием шести ключевых измерений качества данных.

- **Полнота**: отношение фактически имеющегося в хранилище объема данных к потенциально доступному (0–100%).
- **Уникальность**: ни одному реально существующему экземпляру предмета описания (объекта) не должно соответствовать более одной записи в рамках идентификации описываемых предметов/объектов.
- **Актуальность**: степень отражения данными реального положения вещей на текущий момент времени.
- **Годность** определяется синтаксическим соответствием данных определениям (по формату, типу, диапазонам значений и т. п.).

[1] Расширенные и уточненные параметры качества данных по Инглишу представлены в книге «Качество информации в прикладном понимании» (English, 2009).

- **Соответствие:** Степень соответствия данных реальным объектам или событиям, которые ими описываются.
- **Согласованность**: отсутствие противоречий между различными представлениями одного и того же (согласно определениям) предмета или сущности.

В том же аналитическом докладе DAMA UK описываются и другие важные характеристики, влияющие на качество данных. Они не называются в докладе измерениями, но схожи с контекстными и репрезентативными характеристиками данных согласно модели Уанга — Стронг и с утилитарными характеристиками данных по Инглишу. Дополнительные параметры оценки качества данных, согласно этой модели, включают следующее.

- **Полезность**: насколько понятны, доходчивы, релевантно определены, доступны и точны данные?
- **Своевременность реагирования** (в дополнение к актуальности): поддерживается ли возможность оперативного изменения данных без потери стабильности?
- **Гибкость**: насколько данные совместимы и сопоставимы с другими данными? Допускают ли группировку, классификацию и перепрофилирование? Достаточно ли просты в обращении?
- **Надежность**: организованы ли процессы руководства данными и обеспечения безопасности данных? Какова репутация данных, чем или как она подтверждается или удостоверяется?
- **Ценность**: имеется ли экономическое обоснование с анализом рентабельности или окупаемости затрат на управление данными? Оптимально ли используются данные? Всё ли в порядке с защитой персональных, личных и конфиденциальных данных? Не допускается ли предприятием каких-то неправомерных действий или нарушений? Соответствует ли его деятельность корпоративному имиджу?

Единой универсальной классификации измерений качества данных до сих пор не выработано, однако вышеописанные формулировки содержат общие идеи. Измерения включают часть характеристик, оцениваемых по вполне объективно измеримым показателям (полнота, действительность, соответствие формату и т. п.), и часть, которая в значительной степени зависит от контекста или субъективной интерпретации (полезность, надежность источника, репутация и т. п.). Какие бы названия измерений ни использовались, основными аспектами качества данных являются: полнота (отсутствие пробелов); правильность (корректность, точность, достоверность); непротиворечивость (согласованность, целостность, уникальность), актуальность (своевременность обновления или реагирования); доступность; возможность использования (годность); безопасность (защищенность). Таблица 29 содержит набор общепринятых измерений качества данных с определениями и описаниями подходов к их измерению.

Таблица 29. Общепринятые измерения качества данных

Измерение	Определение и описание
Актуальность	Под актуальностью данных понимают совокупность характеристик, относящихся к соблюдению сроков или графиков их получения, синхронизации или обновления. Показатели актуальности при этом нужно определять исходя из ожидаемой волатильности данных в источниках. Как часто они обновляются? Делается это по расписанию или по особым случаям? Синхронизация, то есть соответствие данных текущей версии в информационном источнике, — единственный полностью объективный показатель актуальности. Некоторые данные относительно статичны — например, многие справочные данные, значения которых могут не меняться годами (коды стран, почтовые индексы и т. п.). А вот волатильные данные сохраняют актуальность весьма недолго. Некоторые из них, например биржевые котировки на финансовых веб-страницах, вообще обновляются в режиме, близком к реальному времени, и тут уже потребители данных должны понимать всю меру риска, проистекающего от запаздывания обновлений. В рабочее время, пока биржевые рынки открыты, подобные оперативные данные обновляются с крайне высокой частотой. После закрытия торгов данные остаются неизменными до начала следующего рабочего дня, но актуальности не утрачивают, поскольку рынок не активен. Время запаздывания или задержки обновления данных в хранилищах, доступных пользователям, относительно момента фактического появления или изменения данных, — еще один важнейший показатель актуальности. Например, при пакетной обработке обновлений в ночные часы задержка обновления на момент возобновления доступа пользователей к данным в 08:00 утра может составлять от нуля для данных, сгенерированных непосредственно в ночное технологическое окно, до 24 часов для первых данных, поступивших в систему сутками ранее (см. главу 8).
Консистентность/ Допустимость	Данные проверяются на соответствие значений элементов областям или множествам допустимых значений, которые могут определяться как наборы (например, через справочные таблицы) или интервалы допустимых значений, или же через проверочные правила. Области допустимых значений определяются с учетом типа, формата, разрядности и точности/погрешности измерения ожидаемых на входе величин. Также могут устанавливаться допустимые сроки годности данных — например, поступающих с RFID (радиочастотных идентификаторов) или с каких-нибудь датчиков научно-измерительной аппаратуры. Все данные подлежат валидации на предмет соответствия установленным областям допустимых значений. Важно помнить, что допустимость значений данных не гарантирует их точности или корректности в каждой конкретной записи.
Полнота	Все ли требующиеся данные наличествуют? Полнота может измеряться на уровне наборов данных, записей или столбцов. Все ли ожидаемые записи присутствуют в наборе данных? Корректно ли заполнены поля записей? (Требования к полноте могут зависеть от статуса записи.) Заполнены ли значениями все обязательные столбцы/атрибуты? (В случае наличия условий, требующих заполнения необязательных столбцов, проверяется полнота данных и по ним.) Применяйте к набору данных дифференцированные правила проверки полноты данных: i) в обязательных для заполнения полях; ii) в условно обязательных полях; и iii) в необязательных полях; плюс iv) отсутствие данных в полях неприменимых атрибутов. Измерения на уровне набора данных могут требовать сравнения с источником записей или историческими уровнями полноты данных.

Измерение	Определение и описание
Разумность	Проверка данных на разумность сводится к выявлению наборов или кластеров данных, выходящих за рамки здравого смысла. Например, распределение продаж по географическим регионам должно хотя бы приблизительно соответствовать накопленным нами знаниям о структуре потребительского спроса в этих регионах. Контроль разумности вводных может быть формализован различными способами. Например, в вышеприведенном примере можно сравнивать поступающие данные с контрольными наборами данных для каждого региона или же с предыдущими экземплярами статистики по каждому региону (например, объемами продаж за предыдущий квартал или за тот же период в другие годы). Некая доля субъективизма при оценке разумности данных неизбежна. Если не уверены, проработайте совместно с потребителями данных ожидаемые ими показатели и возьмите их за основу определения реалистичных пределов разумного. После первых серий измерений их результаты можно будет использовать для уточнения и корректировки допусков, равно как и для отслеживания изменения тенденций (см. раздел 4.5).
Согласованность	Речь может идти о проверке: непротиворечивости значений внутри набора данных; отсутствия расхождений в значениях между наборами данных; корректности определения связей по значениям между всеми наборами данных. Сюда же могут включаться и требования к размеру и составу наборов данных, которыми обмениваются различные системы, и динамические параметры согласования данных по времени. Могут определяться требования согласованности: i) различных подмножеств значений атрибутов внутри одной и той же записи (внутренняя непротиворечивость записи); ii) между множествами значений атрибутов из различных наборов и записей (согласованность записей); iii) между множествами значений одних и тех же атрибутов одной и той же записи в различные моменты времени (хронологическая согласованность). Наконец, под согласованностью может пониматься и последовательное использование одного и того же формата для регистрации данных одинаковой структуры. Тут очень важно еще и четко понимать разницу между согласованностью или непротиворечивостью (это логические категории) и точностью, достоверностью, корректностью и прочими техническими характеристиками данных. Согласующиеся внутри наборов и по всем наборам характеристики данных могут послужить прекрасным базисом для стандартизации. Стандартизацией данных называют приведение контента и формата вводных данных в соответствие с едиными правилами. Стандартизация формы и содержания данных, в свою очередь, способствует повышению эффективности и согласованности результатов сравнительного анализа данных. Встраивайте во все каналы обмена данными наборы правил проверки согласованности, обеспечивающие как непротиворечивость значений одних и тех же или связанных атрибутов данных в различных записях или сообщениях, так и согласованность между собой всех значений отдельно взятого атрибута (в частности, посредством определения области или списка допустимых значений). Например, вполне разумно ожидать, что число транзакций в сутки должно находиться в пределах скользящего среднего за предшествующие 30 календарных дней $\pm$ 5% (или $\pm$ 3σ).
Соответствие	Степень близости данных к «реальности». Измерить ее бывает крайне трудно, поскольку о фактических характеристиках описываемого объекта обычно можно судить только по собранным или полученным данным. Объективную оценку точности организация может получить разве что посредством полного воспроизведения всего процесса сбора данных или методом ручной проверки и подтверждения точности всех записей. Большинство показателей точности рассчитываются методом сравнения полученных данных с некими эталонными данными из источника, считающегося заведомо достоверным и содержащего достаточно точные данные (например, из системы регистрации), или надежного источника справочных данных (например, Dun & Bradstreet).

Измерение	Определение и описание
Уникальность / Отсутствие дублирования	Уникальность данных в наборе выражается в отсутствии объектов-клонов или иных тождественных друг другу сущностей. Утверждать, что все сущностные объекты в наборе данных уникальны, можно лишь при обеспечении выполнения условия «одно значение ключа ↔ один и только один объект» в наборе данных. Соответственно, проверяется уникальность тестированием структуры ключей (см. главу 5).
Целостность	Под целостностью или связностью данных в широком смысле понимают полноту, точность и согласованность совокупности имеющихся данных. Однако в узком прикладном понимании контроля качества данных под целостностью обычно понимают либо целостность данных на уровне ссылок (наличие во всех парах логически связанных объектов данных общего для обоих объектов ссылочного ключа), либо внутреннюю связность набора данных, то есть отсутствие в нем логических пустот (недостающих элементов). Наборы данных, не соответствующие критериям внутренней целостности, считаются поврежденными или сохраненными с потерей данных. В наборах данных, лишенных ссылочной целостности, присутствуют либо записи со ссылочными ключами, ведущими в никуда (так называемые «сироты»), либо неоднократно повторяющиеся идентичные строки («дубликаты»), искажающие результаты расчета средних и особенно суммарных значений. Статистику «сирот» и «дубликатов» в различных наборах данных можно отслеживать как в абсолютном (по числу дефектных записей), так и в процентном выражении.

Рисунок 92 соотносит различные измерения качества данных с соответствующими им понятиями и концепциями. Стрелками обозначены перекрестные ссылки, что служит дополнительной иллюстрацией отсутствия устоявшихся соглашений относительно определения стандартных наборов измерений. Например, измерение «соответствие» отвечает понятиям «соответствие реальному миру» и «соответствие доверенному источнику», но также связано с понятиями, отвечающими измерению «допустимость», в частности с понятием «контроль допустимости значений.

1.3.4 Роль метаданных в обеспечении качества данных

Метаданные — незаменимый инструмент управления качеством данных. Качество данных определяется степенью их соответствия нуждам потребителей, а метаданные описывают, что именно отображают данные, — то есть без них невозможно соотнести данные с конкретными нуждами. Наличие надежного процесса определения данных через метаданные также позволяет организации формализовывать и документировать стандарты и требования, на соответствие которым будут проверяться данные в рамках контроля качества. Иными словами, качество данных определяется их соответствием ожиданиям, а ожидания разъясняются через метаданные.

Хорошее управление метаданными также способствует оптимизации усилий, направленных на повышение качества данных в масштабах организации. Репозиторий метаданных вполне может использоваться и для централизованного складирования результатов измерений показателей качества данных, и для обеспечения совместного доступа к ним всех подразделений организации, — и это будет отличной отправной точкой для команды качества данных по направлению к выработке общеорганизационного консенсуса относительно приоритетов и драйверов улучшений (см. главу 12).

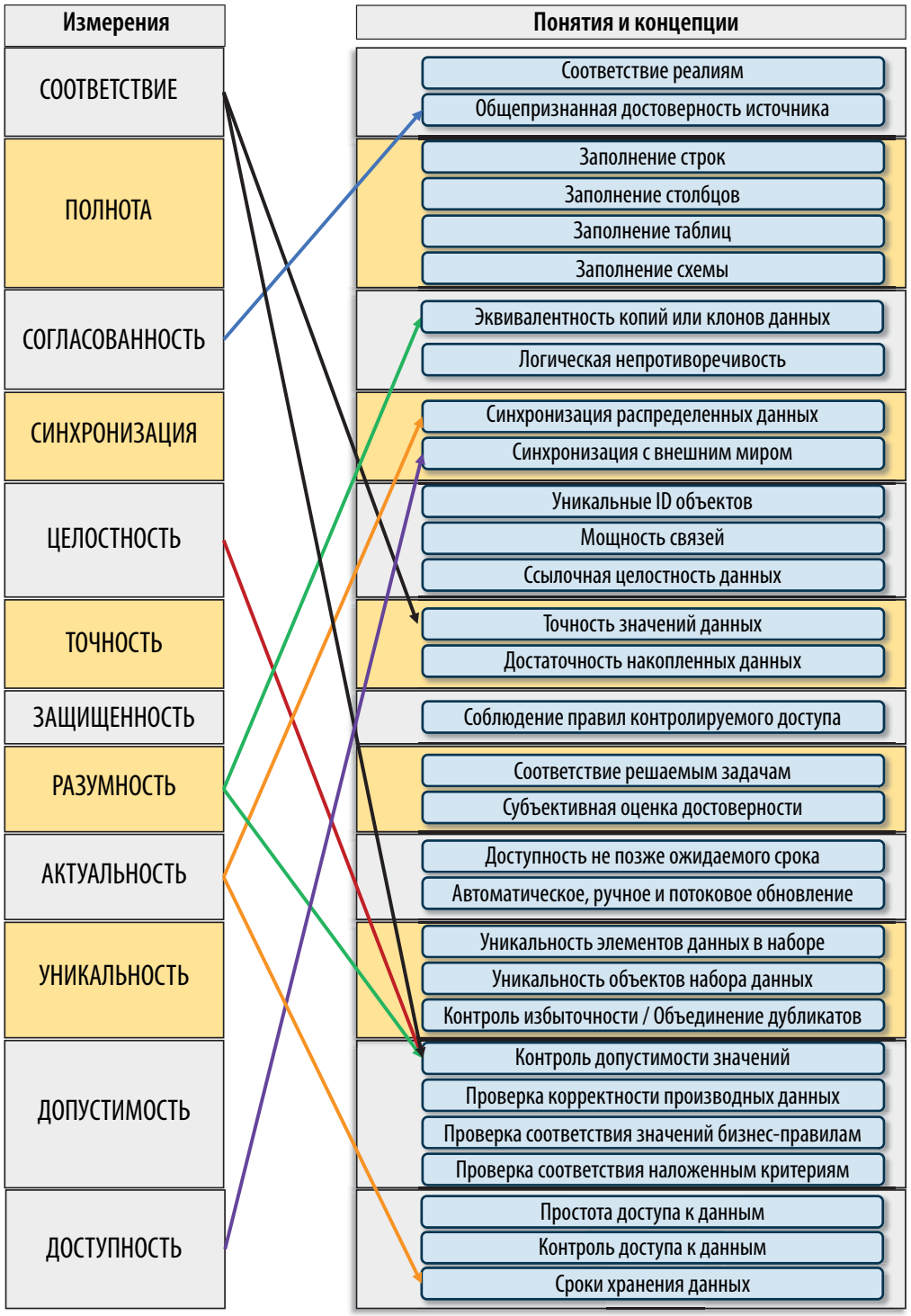


Рисунок 92. Взаимосвязи между параметрами качества данных[1]

[1] Позаимствовано с доработками из Myers (2013) с разрешения правообладателя.

1.3.5 Стандарты ISO в области качества данных

На стадии разработки находится международный стандарт качества данных ISO 8000, призванный обеспечить возможность обмена сложными по структуре данными в независимых от приложений форматах. Во введении к ISO 8000 заявлено: «Способность создавать, собирать, хранить, обслуживать, передавать, обрабатывать и представлять данные для поддержки бизнес-процессов своевременно и наименее затратно требует как понимания определяющих характеристик качества данных, так и способности измерять, управлять и учитывать качество данных».

ISO 8000 определяет лишь универсальные характеристики, которые могут быть проверены в любой организации в цепи поставки данных с целью объективной оценки соблюдения ею стандартов качества данных, устанавливаемых ISO 8000[1].

Первая по хронологии публикации часть ISO 8000-110:2009 посвящена синтаксису, семантике кодирования и соблюдению спецификаций основных данных[2]. К 2016 году введены в действие части 100 (обзор и введение), 120 (происхождение), 130 (точность), 140 (полнота) и 150 (рамочная модель управления качеством) и другие части стандарта качества основных данных. Работа продолжается[3].

Соответствующими стандартам качества ISO считаются «переносимые данные, удовлетворяющие предъявляемым требованиям»[4]. Стандарты качества данных ISO разрабатываются в рамках полномасштабных усилий по обеспечению межплатформенной переносимости данных с целью сохранения. «Переносимыми» считаются данные, читаемые без помощи приложений, в которых они были созданы. Данные, для считывания и использования которых требуется лицензионное ПО, к таковым не относятся, поскольку в некоторых ситуациях организация, по какой-либо причине не сумевшая продлить лицензию, лишается доступа к данным, созданным с помощью соответствующего лицензионного ПО.

Обеспечение соответствия данных «предъявляемым требованиям» невозможно без четкой и недвусмысленной формулировки самих требований. Технические аспекты ISO 8000 определяются через стандарт ISO 22745[5], регламентирующий порядок определения, структуру основных и порядок обмена ими. Стандарт ISO 22745 содержит также примеры на XML и определяет формат обмена кодированными данными[6]. ISO 22745 предусматривает создание переносимых данных посредством представления данных с помощью открытых технических словарей, совместимых с ISO 22745, таких как словарь eOTD Ассоциации управления кодами для электронной торговли (ECCMA).

Основное назначение ISO 8000 — помочь организациям научиться отличать качественные данные от некачественных, запрашивать только данные, соответствующие стандартам качества,

[1] http://bit.ly/2ttdiZJ

[2] См.: ГОСТ Р ИСО 8000-110-2011. — *Примеч. пер.*

[3] http://bit.ly/2sANGdi

[4] http://bit.ly/2rV1oWC

[5] См.: ГОСТ Р ИСО 22745-20-2013. — *Примеч. пер.*

[6] http://bit.ly/2rUZyoz

и проверять поступающие данные на предмет соответствия стандартам. Если стандарты соблюдены, качество данных можно подтверждать автоматически с помощью компьютерной программы.

Параллельно с разработкой стандартов качества данных и средств проверки их соблюдения (ISO 8000-1x) ведется работа и над группой стандартов управления качеством данных ISO 8000-6x. Эта группа стандартов определяет структуру и организацию управления качеством данных в масштабе организации. Рамочной структурой управления качеством данных организации, определяемой стандартом ISO 8000-61:2016, предусмотрены следующие основные направления работ, которые будут детально описаны в последующих частях стандарта серии ISO 8000-6x:

- планирование качества данных;
- контроль качества данных;
- обеспечение качества данных;
- повышение качества данных.

1.3.6 Жизненный цикл повышения качества данных

Большинство методологических подходов к повышению качества данных позаимствованы из теории управления качеством технологического производства[1]. В рамках такой парадигмы любые данные считаются, грубо говоря, конечным продуктом комплекса технологических процессов по переработке информационного сырья. Процесс создания данных может быть простым и одношаговым (сбор или получение), а может быть многоэтапным и включать целый ряд последовательных информационно-технологических операций: сбор данных, включение и накопление в хранилище, обобщение в витрине данных и т. д. и т. п. На каждом этапе данные и их качество подвергаются риску: при сборе возможны ошибки; при передаче из системы в систему — потери, дублирования или искажения; при интеграции и накоплении, анализе или обобщении — методологические ошибки и технические проблемы и т. д. Для повышения качества данных необходимо располагать возможностью оценки соответствия выходных данных ожиданиям, которые определяются, с одной стороны, фактическим содержанием входных данных, а с другой — требованиями к технологическим процессам. Поскольку выходные данные отдельно взятого процесса служат исходными данными для других процессов, требования по обеспечению качества данных должны определяться на уровне всей цепочки передачи данных и согласованным образом предъявляться ко всем ИТ-процессам, задействованным в их переработке.

Общий подход к повышению качества данных должен предусматривать реализацию классического цикла Шухарта — Деминга (см. рис. 93) в той или иной его вариации[2]. Будучи основанным

[1] См.: Wang (1998), English (1999), Redman (2001), Loshin (2001) и McGilvray (2008); обзор литературы по теории данных как продукта см.: Pierce (2004).

[2] Понятие «цикл PDCA» (*сокр.* от *англ.* Plan-Do-Check-Act) первоначально предложено Уолтером Шухартом (*англ.* Walter Shewhart, 1891–1967) и популяризовано У. Эдвардсом Демингом (*англ.* W. Edwards, 1990–1993), предпочитавшим вариацию «цикл PDSA» (*сокр.* от *англ.* Plan-Do-Study-Adjust). Цикл DMAIC (*сокр.* от *англ.* define, measure, analyze, improve, control) и производная от него концепция «шести сигм» — вариации на ту же тему. См.: http://bit.ly/1lelyBK

на методологии точных наук, этот четырехфазный цикл задает модель решения задачи методом последовательных приближений: *планирование → реализация → контроль → доработка → планирование →…*

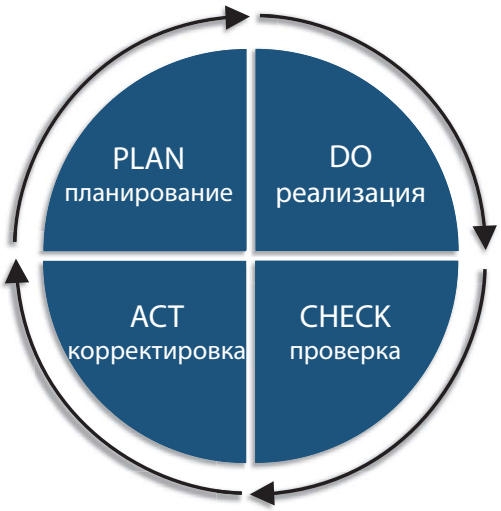


Рисунок 93. Цикл Шухарта — Деминга

Усовершенствования внедряются через строго определенную последовательность шагов. Применительно к программе качества данных это подразумевает следующий алгоритм действий: состояние данных подлежит контролю на предмет соответствия стандартам; если стандарты не соблюдены, требуется доработка, которая начинается с поиска и выявления корневых причин несоответствия данных стандартам с переходом на фазы планирования и реализации мер по устранению первопричин несоответствий, которые могут быть обусловлены технологическими, методологическими, организационными и человеческим факторами. По завершении внесения всех необходимых исправлений и работы над ошибками система управления качеством данных продолжает функционировать в режиме мониторинга систем и контроля текущих данных на предмет выявления возможных новых нарушений стандартов.

Внедрение цикла управления качеством данных для набора данных, который ранее не отслеживался в рамках вышеописанной модели непрерывного совершенствования, начинается с выявления данных, не соответствующих стандартам и/или нуждам потребителей, и проблемных данных и/или процессов, препятствующих успешному решению стоящих перед бизнесом задач. Таким образом, данные нужно проверять на соответствие не только стандартам качества по всем ключевым параметрам, но и всем известным бизнес-требованиям. Далее нужно устанавливать корневые причины несоответствий, чтобы все заинтересованные стороны имели возможность объективно и взвешенно оценить как затратность устранения недоработок, так и уровень риска в случае их сохранения. Эта часть работы обычно осуществляется совместно с *распорядителями данных* и иными заинтересованными лицами.

На стадии *планирования* команда качества данных составляет список текущих задач и проблем, сортирует их по масштабности и приоритетности, оценивает и сравнивает различные варианты решений. План должен строиться на прочном фундаменте анализа корневых причин. Без знания первопричин и последствий имеющихся проблем невозможны ни анализ полезности или эффективности затрат, ни определение приоритетов, а без этого ни о каком планировании говорить не приходится.

На стадии *реализации* команда качества данных руководит работами по устранению корневых причин имеющихся проблем, параллельно планируя показатели и средства последующего контрольного мониторинга данных. В тех случаях, когда корневые причины носят нетехнический характер, команда качества данных совместно с владельцами процессов прорабатывают возможные процедурные изменения и порядок их осуществления. В случае проблем технического характера команда качества данных совместно с соответствующими инженерно-техническими службами обеспечивают надлежащую реализацию требующихся технических изменений и проверяют полученные результаты на предмет возможных ошибок.

На стадии *проверки* осуществляется активный мониторинг качества данных по заданным параметрам соответствия требованиям. До тех пор, пока данные стабильно укладываются в контрольные допуски, дополнительных действий не требуется, а процессы считаются контролируемыми и соответствующими бизнес-требованиям. Но как только обнаруживается снижение качества данных ниже допустимого порогового уровня, необходимо принимать дополнительные меры по возвращению ситуации к норме.

Стадия *корректировки* включает работы по оперативному устранению текущих проблем с данными по мере их выявления системами контроля качества. Как только объем или характер текущих проблем выходят за рамки таких возможностей, цикл возобновляется и начинается поиск первопричин, а затем — проработка возможных решений.

Непрерывность обеспечения качества данных достигается за счет перезапуска цикла управления качеством данных в случае возникновения любой из перечисленных ниже ситуаций:

- выход текущих результатов измерений контрольных показателей за пределы допусков;
- появление новых наборов данных;
- изменение действующих или появление дополнительных требований к имеющимся наборам данных;
- изменение бизнес-правил, стандартов или ожиданий.

Сделать наборы данных правильными изначально — дешевле, чем исправлять неправильные наборы данных. Встроить процессы управления качеством данных в процессы оперативного управления данными с самого начала — на порядок дешевле, чем последующее исправление. Обеспечивать стабильно высокое качество данных на протяжении всего их жизненного цикла — менее рискованно, чем пытаться повышать качество данных в рамках существующих процессов. К тому же и по организации такие перестройки на ходу бьют достаточно тяжело. Определение критериев качества

данных до начала планирования нового процесса или системы — признак зрелости *организации в области управления данными* и отличное средство укрепления административной дисциплины и налаживания плодотворного сотрудничества между функциональными подразделениями.

1.3.7 Бизнес-правила обеспечения качества данных

Бизнес-правила описывают порядок выполнения внутренних операций с целью обеспечения успешного результата без нарушения накладываемых внешних требований. Соответственно, бизнес-правила обеспечения качества данных (бизнес-правила качества данных) описывают порядок обеспечения пригодности и полезности данных, имеющихся в распоряжении организации. Эти правила можно привести в соответствие с измерениями качества и использовать для описания требований, предъявляемых к данным. Например, бизнес-правило «коды стран должны вводиться в стандартном двухбуквенном формате ISO 3166-1 alpha-2» может быть реализовано посредством поля с раскрывающимся списком названий стран в пользовательском интерфейсе и справочной таблицы соответствия кодов названиям, а контроль соблюдения правила можно реализовать с помощью счетчика количества записей с недействительными значениями в полях Код страны.

В целом, бизнес-правила обычно конфигурируются через настройки программного обеспечения или защищенные шаблоны документов с полями ввода с ограничениями по допустимым значениям. Ниже приведены примеры простейших видов бизнес-правил качества данных.

- **Единообразная трактовка определений данных** во всех процессах, реализованных и используемых в организации, может достигаться алгоритмическим согласованием значений в вычисляемых полях, правилами, устанавливающими ограничения по времени или месту, или правилами, ставящими возможность свертки таблицы в зависимость от статуса корректности значений в полях данных в редактируемых записях.
- **Наличие обязательных значений / Полнота записи**: правилами определяются поля, допускающие и не допускающие отсутствие значения / неопределенное значение.
- **Соблюдение формата**: правила проверки соответствия структуры значения, присваиваемого элементу данных, установленному стандарту формата (например, № телефона или адреса e-mail).
- **Множество допустимых значений**: правило проверки наличия введенного значения в таблице допустимых значений соответствующего элемента данных (например, коды стран по ISO 3166-1 alpha-2).
- **Диапазон допустимых значений**: правила ограничения вводимого значения диапазонами числовых, лексикографических или временны́х значений (например, $0 \leq n \leq 100$).
- **Эквивалентность представлений**: различные элементы данных одного набора могут содержать эквивалентные значения в разных представлениях областей допустимых значений. Опять же, удобно проиллюстрировать правила отождествления на примере кодов стран: если один и тот же элемент в разных представлениях описывается значениями из разных справочных таблиц (например, ISO 3166-1 alpha-2, ISO 3166-1 numeric, название), правила должны контролировать, во всех ли представлениях фигурирует одна и та же страна, например: 'RU' ≡ '643' ≡ 'Russia'.

- **Логическое соответствие**: правила проверки взаимной непротиворечивости двух или более значений, между которыми существуют наложенные условные ограничения. Например, Почтовый индекс должен соответствовать Населенному пункту.
- **Контрольная проверка**: правила, позволяющие выявлять ошибки в указанных значениях путем сверки с системой учета или иным верифицированным источником (например, проверка даты продажи по номеру накладной в журнале учета).
- **Проверка уникальности**: правила, определяющие объекты, которые могут существовать лишь в единственном экземпляре, или устанавливающие принцип «одна запись ↔ один реальный объект».
- **Проверка соблюдения временны́х параметров**: правила, определяющие стандартные характеристики актуальности, доступности, обновления данных и т. п.

Правила остальных видов могут включать статистические функции, применяемые к наборам записей (см. раздел 4.5). Например:

- проверка допустимости числа записей в файле;
- проверка допустимости среднего значения по множеству транзакций;
- проверка допустимости дисперсии значений по множеству транзакций.

Пороговые значения определяются и уточняются по мере накопления статистики.

1.3.8 Распространенные причины проблем с качеством данных

Проблемы с качеством могут возникнуть на любой фазе жизненного цикла данных, начиная с создания и заканчивая окончательным удалением. Анализируя корневые причины, аналитикам следует не упускать из виду ни одного потенциального источника проблем, включая ошибки ввода и обработки данных, архитектуру и структуру систем, реализацию автоматизированных процессов и ручные вмешательства в их работу. Многие проблемы носят комплексный характер и бывают обусловлены целым рядом причин и факторов (особенно если люди раз за разом изыскивают обходные пути и «наколенные решения» вместо поиска первопричин). Приведенные ниже описания распространенных причин служат одновременно и подсказками, как не допустить возникновения проблем. Для этого нужно: совершенствовать интерфейсы; включать проверку правил качества данных в процедуры обработки данных; уделять первоочередное внимание обеспечению качества данных на стадии проектирования систем; жестко контролировать любые ручные вмешательства в автоматизированные процессы.

1.3.8.1 ПРОБЛЕМЫ ВСЛЕДСТВИЕ НЕДОСТАТКА ЛИДЕРСТВА

Многие почему-то полагают, будто большинство проблем с данными становятся следствием ошибок при вводе. Более продвинутые сотрудники сознают, что неполно или плохо проработанные бизнес- и технические процессы чреваты куда большими ошибками и проблемами, чем

ввод неверных данных. Однако здравый смысл подсказывает, а исследования подтверждают, что многие проблемы с качеством данных возникают от отсутствия приверженности организации деятельности по обеспечению качества данных, а отсутствие приверженности — от недостатка лидерства как в форме руководства качеством данных, так и в форме управления.

Каждая организация располагает ценными с точки зрения обеспечения ее текущей деятельности информационными активами. Действительно, работа любой организации так или иначе зависит от ее возможностей по совместному использованию информации. Невзирая на это, очень мало организаций управляют своими данными как ценным активом, и еще меньше делают это с должной тщательностью. В большинстве организаций рассогласованность данных (по структуре, формату, значениям и применению) представляет собой куда более серьезную проблему, чем ошибки в данных как таковые, поскольку служит труднопреодолимым препятствием на пути интеграции данных. Одна из причин, по которым программы руководства данными уделяют так много внимания определению терминов и выработке общего языка, как раз и состоит в том, что единообразие терминологии — необходимая отправная точка на пути к обеспечению согласованного управления данными и информационными системами предприятия.

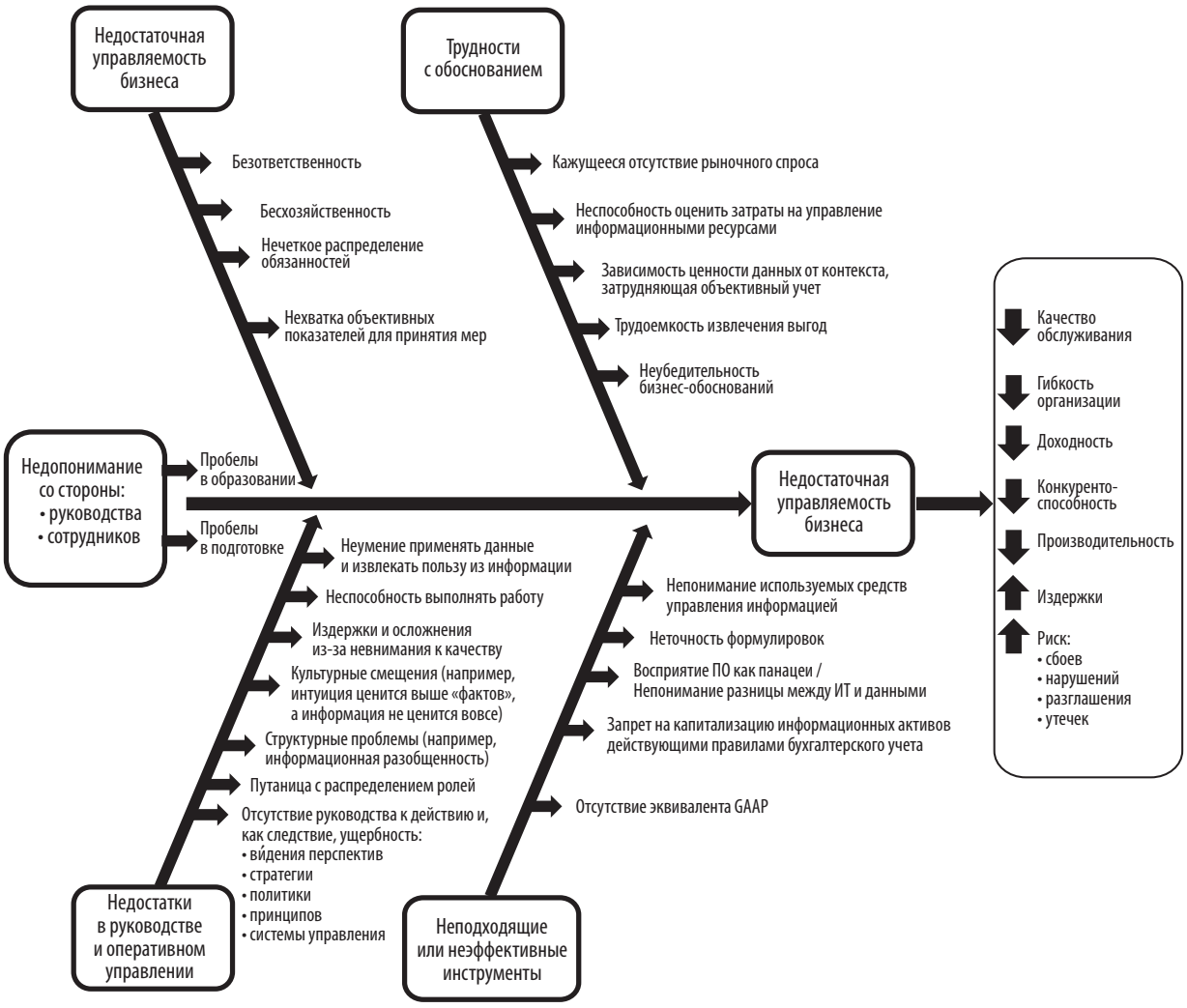
Многие программы руководства данными создаются исходя исключительно из требований по обеспечению нормативно-правового соответствия, а не из соображений реализации потенциальных возможностей по извлечению ценности из данных как актива. Недопонимание руководителями важности управления данными как активом приводит к недостаточной приверженности этой деятельности внутри организации, включая деятельность по управлению качеством данных (Evans and Price, 2012). Ниже представлены основные препятствия для осуществления деятельности по управлению данными как активом организации (см. рис. 94).

Эффективному управлению качеством данных препятствуют следующие факторы:

- недопонимание значения управления качеством данных руководством и сотрудниками;
- недостаточное руководство бизнесом (business governance);
- недостаток лидерства и управления;
- трудности с обоснованием необходимости совершенствования управления качеством данных;
- использование неподходящих или неэффективных инструментов измерения ценности данных и показателей качества данных.

Всё вышеперечисленное негативно сказывается на качестве обслуживания и ожиданиях клиентов, производительности, моральном климате, эффективности работы, обороте и конкурентоспособности организации и к тому же приводит к неоправданному росту затрат на оперативное управление и усугублению рисков[1] (см. главу 11).

[1] Классификация препятствий на пути эффективного управления качеством данных позаимствована с доработками из *The Leader's Data Manifesto* (https://dataleaders.org/).



Препятствия, мешающие компаниям управлять имеющейся информацией как бизнес-активом: распространенные корневые причины
© 2017 dataleaders.org

Рисунок 94. Препятствия для осуществления деятельности по управлению данными как активом[1]

1.3.8.2 ПРОБЛЕМЫ, ВОЗНИКАЮЩИЕ НА СТАДИИ ВВОДА ДАННЫХ

- **Плохо спроектированные интерфейсы ввода данных** могут повлечь самые серьезные проблемы с качеством. В случае отсутствия поддержки функций автоматического исправления или хотя бы контроля данных на входе в систему операторы ввода данных начнут экономить время на заполнении необязательных полей и замене значений по умолчанию на фактические в обязательных полях.

[1] Диаграмма разработана Д. Макгилврей (Danette McGilvray), Дж. Прайсом (James Price) и Т. Редманом (Tom Redman). Используется с разрешения правообладателя (https://dataleaders.org/).

- **Слишком длинные и/или неупорядоченные списки**: столь простой, казалось бы, элемент интерфейса ввода данных, как раскрывающийся список, чреват ошибками ввода, особенно если значений в нем много, а тем более если они еще и не рассортированы.
- **Переназначение полей**: в некоторых организациях практикуют использование старых полей ввода по новому назначению, сообразуясь с изменившимися потребностями бизнеса, вместо того чтобы вносить изменения в модель данных и пользовательский интерфейс. Результатом такой практики часто становится путаница в данных.
- **Низкий уровень подготовки**: незнание специфики процессов чревато вводом неверных данных даже и при наличии средств контроля и редактирования. Если операторы понятия не имеют об ожидаемых значениях и потенциальных последствиях ошибок ввода, а также получают сдельную оплату за объем введенных данных или премируются за скорость, а не за безошибочность ввода, то данные будут вводиться быстро, но некачественно.
- **Изменения в бизнес-процессах** неизбежны и часто сопровождаются изменением действующих и/или введением новых бизнес-правил требований к качеству данных, включая точность ввода. Однако изменения бизнес-правил не всегда находят своевременное и полное отражение в конфигурационных настройках систем. Отсюда и неизбежные ошибки ввода вследствие того, что интерфейс не перенастроен под новые или изменившиеся требования. Кроме того, возможен и ущерб качеству данных вследствие несогласованного перевода на новые правила различных подсистем и интерфейсов.
- **Рассогласованность при выполнении бизнес-процессов**: данные, создаваемые различными процессами, неизбежно будут рассогласованными, если отсутствует надлежащее согласование между самими процессами при их выполнении. Причины же рассогласованного выполнения процессов могут быть обусловлены как недостаточной подготовкой операторов, так и проблемами с документацией или реализацией изменившихся требований.

1.3.8.3 ПРОБЛЕМЫ, ВОЗНИКАЮЩИЕ НА СТАДИИ ОБРАБОТКИ ДАННЫХ

- **Неверные представления об источниках данных** неизбежно приводят к проблемам в производственной среде вследствие ошибок получения или обработки, несоответствия структуры и форматов, неполной или устаревшей системной документации, несоответствия или незнания протоколов передачи и т. п. (подобное случается, к примеру, когда эксперты в предметных областях увольняются, не передав задокументированные знания преемникам). Аналогичные проблемы возникают и при объединении систем — например, при слияниях и поглощениях, когда у новых владельцев имеются лишь самые смутные представления о связях между исходными системами. При необходимости интеграции множества систем — источников данных и/или потоков данных, выдаваемых этими системами, всегда есть риск упустить из виду те или иные детали, особенно в тех случаях, когда далеко не всё известно о различных источниках данных, а время поджимает.
- **Устаревшие бизнес-правила** подлежат своевременному выявлению и обновлению или замене. Если предусмотрен регулярный автоматический контроль и обновление правил, то сама эта

процедура также подлежит мониторингу и обновлению, иначе могут возникнуть проблемы, обусловленные пропущенными изменениями и/или ложноположительными срабатываниями.

- **Изменения в структуре данных**, поступающих из систем-источников, порой происходят вовсе без предварительного уведомления потребителей (пользователей или систем), или же уведомление поступает слишком поздно, чтобы можно было успеть как следует подготовиться к учету изменений. Результатами становятся недопустимые значения, неподдерживаемые форматы или иные характеристики входных данных, препятствующие их приему, загрузке и обработке. Но хуже всего ситуации, когда изменения внешне проходят незамеченными, на результатах последующих процессов сказываются негативно, а выявляется это лишь по накоплении критических объемов брака.

1.3.8.4 ПРОБЛЕМЫ, ОБУСЛОВЛЕННЫЕ СИСТЕМНЫМИ ПРОЕКТНЫМИ РЕШЕНИЯМИ

- **Нарушение ссылочной целостности данных**: без обеспечения ссылочной целостности о высоком качестве данных не может быть и речи ни на уровне отдельного приложения, ни на уровне системы управления данными предприятия. Отключение проверки целостности ссылок (например, ради повышения производительности или ускорения отклика системы) рано или поздно повлечет за собой серьезные проблемы с качеством данных, включая:
  - дублирование и, как следствие, нарушение принципа уникальности данных;
  - строки, включенные в одни отчеты и не включенные в другие, приводят к тому, что одни и те же вычисления могут заканчиваться различными результатами;
  - невозможность обновления данных из-за изменившихся требований к ссылочной целостности;
  - неточность данных из-за присвоения отсутствующим элементам значений по умолчанию.
- **Несоблюдение требований уникальности**: появление непредусмотренных копий экземпляров данных в файлах или таблицах. Если проверки на уникальность отключены или урезаны (например, ради повышения производительности СУБД), суммарные и усредненные величины будут рассчитаны некорректно.
- **Ошибки в алгоритмах и настройках программ**: если мэппинг данных, схемы движения, формулы преобразования или иные правила и алгоритмы обработки данных заданы с ошибками или пробелами, неизбежно возникнут проблемы с качеством данных, причем самые разнообразные — от неверно рассчитанных значений до отправки данных не по адресу, присвоения недействительных ключей и ошибочного определения ссылочных или логических связей.
- **Некорректная модель данных**: если модель данных построена на непроверенных гипотезах, которые не соответствуют действительности, неизбежно возникнут проблемы с качеством данных. Возможные последствия — утеря данных вследствие выхода фактических значений за пределы области допустимых значений, несоответствие разрядности, неверное присвоение идентификаторов или ключей.
- **Переназначение полей**: использование старых полей с новыми целями, чтобы не тратить времени на изменение модели данных или кодов программ, может привести к путанице

с наборами значений, неверной трактовке данных, а потенциально и к нарушению целостности структуры данных (например, из-за ошибочно присвоенных ключей).

- **Хронологические несоответствия**: при отсутствии консолидированного словаря данных различные системы могут использовать несовпадающие форматы даты/времени, что чревато рассогласованиями данных по времени и утерей данных при синхронизации между системами-источниками.
- **Недостаточно организованное управление основными данными** может привести к выбору ненадежных источников и, как следствие, хроническим проблемам с качеством данных, которые к тому же еще и крайне трудно диагностировать до тех пор, пока не возникнет сомнений в достоверности источника.
- **Дублирование данных** становится еще одним следствием плохого управления данными. Непредусмотренные дубли возникают в основном по двум причинам.
  - **Единственный источник — множественные локальные экземпляры**: например, экземпляры данных (экземпляры сущности) об одном и том же клиенте присутствуют в нескольких идентичных или похожих таблицах одной и той же базы данных. Через некоторое время определить, в какой именно таблице содержатся актуальные данные, без тщательного разбирательства, требующего знания специфики систем, становится невозможным.
  - **Множественные источники — единственный экземпляр**: например, данные о покупателе поступают в центральное хранилище из множества систем в торговых точках. Обработка этих данных может вестись с использованием нескольких экземпляров, создаваемых в разных областях временного хранения. Проблемы в таком случае возникают при объединении данных в постоянной области, если четко не прописаны правила приоритетности источников.

1.3.8.5 ПОВТОРНЫЕ ПРОБЛЕМЫ ВСЛЕДСТВИЕ НЕПРОДУМАННОГО РЕШЕНИЯ ПЕРВОНАЧАЛЬНЫХ ПРОБЛЕМ

Ручное обновление сведений непосредственно в рабочей базе данных, а не с помощью определения бизнес-правил для приложений или процессов, — занятие рискованное. Наспех написанные скрипты или вручную вводимые команды, как правило, используются как экстренная мера в надежде в авральном порядке «исправить» данные в чрезвычайной ситуации, возникшей, например, в результате злонамеренного вброса ложных данных, взлома защиты, внутреннего мошенничества, хакерской атаки или иных обстоятельств, повлекших сбой нормального исполнения бизнес-процессов.

Как и любой другой непроверенный код, такие исправления сопряжены с высоким риском возникновения вторичных ошибок из-за непредвиденных последствий и побочных эффектов, например: изменение данных, которые менять не требовалось; повреждение данных; сохранение первоначальных проблем в части данных (например, исторических). Большинство таких исправлений к тому же изменяют данные необратимым образом, записывая новые значения поверх старых, а не добавлением исправленных строк.

Такие изменения обычно невозможно исправить простым восстановлением из резервной копии, поскольку в журналы они не заносятся, то есть восстановление данных до исходного состояния возможно только из резервной копии всей базы данных. Следовательно, подобные

инициативы должны быть строго наказуемы, поскольку они чреваты нарушениями информационной безопасности и сбоями в работе, на устранение которых понадобится куда больше времени, чем ушло бы на надлежащее исправление исходной проблемы. Все изменения должны вноситься в соответствии с утвержденным бизнес-процессом по управлению изменениями.

1.3.9 Профилирование данных

Профилирование данных (*data profiling*) — это форма анализа с целью исследования данных и оценки качества. Профилирование использует статистические методы с целью выявления истинной структуры, контента и качества массива данных (Olson, 2003). Программный комплекс профилирования предоставляет статистику, позволяющую аналитикам выявлять закономерности в контенте и структуре данных. Например:

- **Счетчик пустых значений** позволяет выявлять столбцы, где таковые присутствуют, и проверять допустимость подобной ситуации согласно определениям и/или правилам.
- **Max/Min значения** проверяются на их допустимость и позволяют выявлять такие аномалии, как, например, отрицательные значения в полях, смысл которых таковых не предусматривает.
- **Max/Min длина** позволяет выявлять поля со значениями, выходящими за пределы допуска по числу знаков.
- **Частотное распределение** значений в отдельных столбцах позволяет оценивать разумность данных (например, с точки зрения распределения транзакций по кодам стран или часто и редко встречающихся значений), а также процент записей, где сохранены значения по умолчанию.
- **Проверка соответствия типа и формата** позволяет выявлять несоответствия и вести статистику некондиционных данных, в том числе и по типам ошибок ввода (например, неверное число знаков после запятой, недопустимые пробелы, выборочные значения и т. п.).

К профилированию относится и перекрестный анализ столбцов, позволяющий выявлять пересечения или дублирование данных в различных столбцах, а также скрытые зависимости, и перекрестный анализ таблиц на предмет выявления пересекающихся множеств значений и определения связей через внешние ключи. Большинство средств профилирования данных оснащено инструментами углубленного анализа выявленных статистических закономерностей.

Результаты профилирования должны оцениваться профессиональными аналитиками на предмет соответствия данных правилам и бизнес-требованиям. Хороший аналитик способен использовать результаты профилирования также и для проверки известных и выявления новых взаимосвязей и характеристик данных, скрытых корреляций и закономерностей внутри и между наборами данных, включая бизнес-правила и ограничения допустимости. Профилирование часто используется на стадии анализа данных, требующегося для проектов (в частности, интеграционных; см. главу 8), или для оценки текущего состояния данных, которые планируется улучшить. Результаты профилирования данных могут использоваться также и для выявления

возможностей для повышения качества не только самих данных, но и метаданных (Olson, 2003; Maydanchik, 2007).

Будучи хорошим средством изучения данных, профилирование, однако, является лишь первым шагом к повышению качества данных. Оно помогает организациям выявлять проблемы, а для решения проблем требуются уже другие аналитические приемы, включая анализ бизнес-процессов, генеалогии данных и глубинных причин выявленных проблем.

1.3.10 Повышение качества данных в процессе обработки

В то время как первоочередное внимание в рамках усилий по обеспечению качества данных часто уделяется предотвращению ошибок, имеется немало возможностей и для повышения качества данных в процессе их обработки (см. главу 8).

1.3.10.1 ОЧИСТКА ДАННЫХ

Очистка (или *исправление*) *данных* (*data cleansing* или *data scrubbing*) заключается в их преобразовании с целью приведения в соответствие с требованиями стандартов или правилами определения допустимых значений. Очистка включает выявление данных с ошибочными значениями и исправление этих ошибок.

Исправление данных посредством очистки на регулярной основе — занятие дорогостоящее и рискованное. В идеале потребность в очистке должна неуклонно снижаться по мере выявления и устранения корневых причин появления некачественных данных. Снижению затрат на очистку данных способствуют следующие меры:

- внедрение механизмов защиты от ошибок ввода;
- исправление данных в системах-источниках;
- совершенствование бизнес-процессов, в рамках которых создаются вводные данные.

Тем не менее в определенных ситуациях от сплошного контроля и исправления вводных избавиться нереалистично, а в некоторых случаях такой фильтр на промежуточном участке потока данных в системе обходится дешевле любого альтернативного решения.

1.3.10.2 УЛУЧШЕНИЕ КАЧЕСТВА ДАННЫХ

Улучшение или *обогащение данных* (*data enhancement* или *data enrichment*) состоит в добавлении к набору данных атрибутов, способствующих повышению качества и пригодности данных для использования по назначению. Иногда улучшения достигаются и посредством интеграции внутриорганизационных наборов данных, а также дополнением их внешними наборами, включая коммерческие (см. главу 10). Примеры улучшений:

- **Метки Даты/Времени**: документирование даты/времени создания, последнего изменения и/или срока годности различных элементов данных весьма способствует их упорядочиванию

и отслеживанию истории событий. В случае выявления проблем с данными метки времени могут оказаться крайне полезными с точки зрения выявления корневых причин, поскольку позволяют аналитикам определять четкую хронологию событий перед возникновением проблем.

- **Аудит данных**: по журналам или файлам данных аудита можно отслеживать происхождение и историю преобразования различных элементов данных, а это бывает очень важно и с точки зрения выявления первопричин проблем, и с точки зрения валидации самих данных.
- **Справочные словари**: словари специфической для бизнеса терминологии, а также онтологии и глоссарии укрепляют и расширяют понимание смысла и контекста описываемых явлений, тем самым помогая лучше управлять данными.
- **Контекстная информация**: добавление данных о контексте (местонахождении, среде, методах доступа и т. п.) помогает находить и помечать данные, нуждающиеся в детальном анализе.
- **Географическая информация** может совершенствоваться посредством стандартизации адресов и добавления геолокационных кодов, индексов, районов, карт местности, точных координат (широты/долготы), схем проезда и т. д. и т. п.
- **Демографическая и эконометрическая информация**: данные о потребителях очень выигрывают от обогащения закодированными сведениями о поле, возрасте, семейном положении, уровне доходов, этнической принадлежности и т. п., а данные о корпоративных клиентах — от сведений о годовом обороте, числе сотрудников, площади занимаемых помещений и т. п.
- **Психологические и поведенческие профили** используются для сегментирования целевых групп по привычкам, образу жизни, вкусам, предпочитаемым брендам, членству в организациях, предпочитаемым видам досуга, способам и маршрутам поездок на работу, времени посещения магазинов и т. д. и т. п.
- **Данные об оценке объектов недвижимости и движимых активов** полезны в любых отраслях.

1.3.10.3 СИНТАКСИЧЕСКИЙ АНАЛИЗ И ФОРМАТИРОВАНИЕ ДАННЫХ

Синтаксический анализ (*data parsing*) — это процесс анализа предопределенных правил организации данных с целью определения содержания или значений. Он позволяет аналитикам определять наборы шаблонов, которые могут быть использованы для того, чтобы отличить годные данные от некондиционных. При совпадении данных с определенными шаблонами могут автоматически инициироваться те или иные действия.

В частности, синтаксический анализ позволяет присваивать характеристики значениям элементов данных, относящихся к экземпляру сущности, чтобы затем по этим характеристикам можно было определить потенциальные источники дополнительных полезных данных. Например, если среди значений атрибута ‘Name’ (имя) обнаруживаются помимо фамилий/имен еще и названия компаний, делается вывод, что в данном случае в общий список попали физические и юридические лица, и включаются правила сортировки записей на две категории по соответствующему признаку. Используйте такой подход в любой ситуации, когда имеется возможность распределить значения данных по семантическим иерархиям: например, посредством разбивки категории «комплектующие» на подкатегории «детали», «узлы» и «модули».

Многие проблемы с качеством данных возникают из-за неоднозначности трактовок близких значений, что затрудняет отнесение данных к одной из смежных категорий. В такой ситуации можно извлечь и объединить данные всех смежных категорий и заново их рассортировать, применив уточненные стандартные критерии представления и получив таким образом достоверную картину. При выявлении в структуре исходных данных недопустимого значения приложение может попытаться исправить его на допустимое согласно набору правил. Стандартизацию лучше обеспечивать с помощью карт приведения структуры исходных данных в соответствие со структурой целевого представления, включая необходимые преобразования, пересчеты и переформатирования.

Например, предположим, что нам нужно привести к единому формату номера телефонов из разнородных источников. Где-то номера указаны без пробелов, где-то с пробелами, где-то через дефисы, а в некоторых представлениях присутствуют еще и номера с литерами. Человек безошибочно распознает и наберет телефонный номер в любом из этих форматов. А вот для того, чтобы удостовериться в том, что номера указаны точно и без ошибок (или хотя бы допустимы), или выявить дублирующие друг друга контактные телефоны, в то время как действует правило «одно лицо — один номер», требуется синтаксический анализ значений с точным определением сегментов (код доступа в сеть, код страны, код города/оператора, номер абонента) и последующим приведением всех номеров к единому стандартному формату.

Другой пример — стандартизация записи имен клиентов. Хорошее средство семантического анализа должно уметь распознавать и сличать на предмет совпадений тысячи тысяч всевозможных сочетаний и перестановок, вариантов и вариаций различных компонентов, включая фамилию, имя (в том числе полное и варианты уменьшительного), отчество или второе имя (если есть), инициалы, звания, должности, титулы, обозначения и характеристики, а затем встраивать полученные в результате анализа обобщенные компоненты в каноническое представление, которое смогут использовать другие службы данных.

Способность человеческого мозга распознавать знакомые образы и структуры позволяет нам с легкостью относить формально разнородные значения к одной и той же абстрактной категории данных; людям интуитивно понятно, что 8(495) и +7 495 — всего лишь разные представления одного и того же префикса телефонного номера, поскольку они к этому привыкли. Аналитику же приходится педантично описывать все возможные структуры форматов данных, которые только могут обнаружиться в полях столбцов с весьма общими заголовками наподобие Контактное лицо, Описание продукта и т. п. Современные средства контроля качества данных поддерживают не только синтаксический анализ значений на уровне, достаточном для четкого выявления всех рассогласованных по формату представлений одних и тех же данных, но нередко и приведение всех однотипных данных к единому стандартизированному формату, что весьма упрощает последующий сравнительный анализ и исправление. После накопления достаточной статистики средства структурно-семантического анализа данных позволяют в значительной мере автоматизировать распознавание и приведение к стандартной форме осмысленных компонентов значений на входе в систему.

1.3.10.4 ПРЕОБРАЗОВАНИЕ И СТАНДАРТИЗАЦИЯ ДАННЫХ

При преобразовании данных к стандартному виду применяются правила обработки, позволяющие перевести их в формат, который может прочитать целевая система. Однако «читаемость» данных не гарантирует приемлемости их значений. Правила обработки и проверки должны применяться непосредственно в интеграционном потоке данных или встраиваться в специальные отдельно используемые инструменты.

Любые преобразования данных должны иметь встроенные механизмы стандартизации. При выработке правил переноса данных из системы в систему строго придерживайтесь спецификаций мэппинга. Выявленные средствами синтаксического анализа нестандартные компоненты в структуре данных подлежат реструктурированию, исправлению и прочим изменениям согласно действующим правилам с целью приведения их в соответствие с установленными стандартами. Фактически стандартизация является частным случаем трансформации данных, но только с использованием правил, определенных не произвольным образом, а с учетом всей суммы накопленных знаний о контексте, лингвистике и идиоматике, подкрепленных многократной проверкой этих правил на предмет их соответствия реальному положению вещей специалистами по разработке правил или поставщиками инструментов (см. главу 3).

2. ПРОВОДИМЫЕ РАБОТЫ

2.1 Определение данных высокого качества

Некачественные данные часто узнаваемы с первого взгляда. Намного сложнее дать четкое определение высококачественных данных. Потребители либо вовсе теряются, затрудняясь сформулировать критерии и признаки качества данных, либо отделываются общими фразами: «данные должны соответствовать действительности», «нам нужны точные цифры» и т. п. Но даже из таких ответов можно сделать вывод, что в потребительском понимании главным критерием высокого качества данных служит их пригодность к использованию по назначению. Перед вводом в действие программы качества данных полезно получить как можно более точное и детализированное понимание нужд бизнеса, сложившейся терминологии и болевых точек организации, чтобы изначально имелся консенсус относительно базовых стимулов и приоритетов в сфере повышения качества данных. Набор стандартных вопросов к целевой группе потребителей данных, по ответам на которые можно составить достаточно полное и точное представление о текущем состоянии и готовности организации к внедрению модели качества данных, основанной на принципах непрерывного совершенствования жизненного цикла данных, в самой общей формулировке включает следующие принципиальные вопросы.

- Что именно вы, как ответственное лицо, понимаете под «высококачественными данными»?
- Как сказывается низкое качество данных на ведении и стратегии вашего бизнеса?

- Какие новые стратегические возможности откроются перед вашим бизнесом с повышением качества данных?
- Какие стимулы к повышению качества данных являются приоритетными?
- Определены ли допуски погрешностей данных? Если да, то каковы предельно допустимые отклонения?
- Какие структуры руководства, обеспечивающие поддержку повышения качества данных, существуют?
- Какие дополнительные структуры руководства могут понадобиться?

Помимо вышеперечисленных вопросов, для получения исчерпывающей картины текущего состояния качества данных в организации необходимо подойти к проблеме всесторонне и рассмотреть ее под различными углами. Для этого нужно:

- понять стратегию и цели бизнеса;
- уточнить у заинтересованных лиц все болевые точки, риски и бизнес-стимулы;
- срежиссировать комплексную экспертизу данных методами профилирования и статистического анализа;
- задокументировать зависимости между данными в бизнес-процессах;
- задокументировать техническую архитектуру и системную поддержку бизнес-процессов.

Подобная экспертиза иногда позволяет выявить целый ряд возможностей для значительных улучшений, а приоритетные из их числа затем определяются по потенциальной пользе от их реализации с точки зрения организации. Используя вводные, полученные от заинтересованных лиц, включая распорядителей данных и экспертов в предметных областях бизнеса и ИТ, команда качества данных окончательно определяет смысл понятия «качество данных» и приоритетные направления программы.

2.2 Определение стратегии качества данных

Для повышения качества данных требуется стратегия, определяющая как работу, которую нужно проделать, так и способы ее практического выполнения. Приоритеты программы качества данных должны согласовываться с бизнес-стратегией. Принятие на вооружение готовой или разработка собственной рамочной структуры и методологии программы качества данных помогает согласованно планировать стратегию и тактику действий, обеспечивая при этом еще и средства объективного измерения достигнутого прогресса и результатов. Рамочная структура должна предусматривать методы, обеспечивающие:

- понимание и приоритизацию бизнес-нужд;
- выявление критически важных данных в привязке к бизнес-нуждам;

- определение бизнес-правил и стандартов качества данных, соответствующих бизнес-требованиям;
- определение и измерение показателей соответствия данных ожиданиям;
- доведение полученных заключений до сведения заинтересованных лиц и сбор отзывов;
- приоритизацию проблем и управление их разрешением;
- выявление и приоритизацию возможностей для совершенствования;
- измерение, мониторинг и учет показателей качества данных;
- управление метаданными, получаемыми в рамках процессов управления качеством;
- интеграцию механизмов программы качества данных в бизнес-процессы и технологические процессы.

Рамочная структура должна также описывать организационные аспекты программы качества данных и порядок использования инструментальных средств, обеспечивающий максимальную отдачу. Как уже упоминалось во вводной части настоящей главы, для повышения качества данных команда программы качества данных должна привлекать к деятельности по обеспечению качества данных сотрудников бизнес- и технологических подразделений с целью выявления критических проблем, выработки практических рекомендаций, разработки и внедрения операционных процессов, необходимых для реализации концепции непрерывного управления качеством данных. Часто такая команда входит в состав организационной системы управления данными (Data Management Organization). Аналитики качества данных должны тесно сотрудничать с распорядителями данных на всех уровнях. Также у них должны иметься рычаги влияния на политику организации, в частности политику определения бизнес-процессов и развития информационных систем. Однако наличие такой команды само по себе не служит гарантией разрешения всех проблем с качеством данных, испытываемых организацией. Работа по обеспечению качества и приверженность высокому качеству данных должны стать неотъемлемой частью повседневных практик организации. Стратегия качества данных должна также предусматривать распространение передовых методов (см. главу 17).

2.3 Определение критически важных данных и бизнес-правил

Не все данные одинаково важны. Основные усилия по управлению качеством данных должны быть направлены на важнейшие для организации данные: те, повышение качества которых принесет максимальную отдачу организации и ее клиентам. В качестве приоритетных могут выбираться различные критерии ценности данных — обязательность для соблюдения установленных внешних требований, финансовая значимость, прямое влияние на потребителей и т. п. Часто усилия по повышению качества данных начинаются с проработки основных данных, которые по определению являются важнейшими для любой организации. Результатом анализа значимости становится упорядоченный список приоритетных данных, которым команда качества данных и руководствуется.

Определив критически важные данные, аналитики качества данных должны выявить бизнес-правила, описывающие или подразумевающие требования, предъявляемые к качественным

характеристикам этих данных. Важно помнить, что многие бизнес-правила явным образом нигде не документируются, поскольку их соблюдение считается само собой разумеющимся. Поэтому для того, чтобы добраться до бизнес-правил как таковых, может потребоваться проведение реверс-инжиниринга на основе анализа существующих бизнес-процессов, рабочих процедур, регламентов, политик, стандартов, системных настроек, триггеров и процедур присвоения кодов статуса на уровне ПО, — и всё это должно дополняться простыми соображениями здравого смысла. Например, если маркетинговая компания хочет ориентироваться на целевую группу, определяемую демографическими характеристиками, то потенциальные показатели качества данных могут определяться уровнем соответствия аудитории демографическим параметрам, таким как пол, возраст, уровень доходов семьи и т. п.

Большинство бизнес-правил так или иначе относятся к порядку сбора или создания данных, в то время как показатели качества данных призваны оценивать степень их пригодности к использованию по назначению. Однако оба эти понятия — создание и использование данных — также взаимосвязаны. Желание использовать данные обусловлено не только тем, что они отражают, но и тем, как и откуда эти данные получены. Например, чтобы разобраться со статистикой продаж за указанный квартал, организации нужно знать не только сумму выручки, но и располагать достоверными данными о структуре продаж (объемы проданных товаров по наименованиям, каналам сбыта, доле постоянных/новых покупателей и т. п.).

Все возможные способы использования тех или иных данных выяснить, как правило, нереалистично, зато вполне можно понять процессы и правила сбора или создания данных. Измеримые характеристики годности данных должны разрабатываться в привязке к известным способам их использования и в проекции на оси параметров качества данных, то есть описывать их полноту, соответствие, допустимость, целостность и т. д. Параметры качества позволяют аналитикам как определять правила (например: «поле X обязательно для заполнения»), так и описывать полученные результаты и выводы (например: «поле не заполнено в 3% записей; полнота данных = 97%»).

На уровне полей или столбцов правила могут определяться достаточно просто и прямолинейно. Правила полноты определяют, является ли поле обязательным и, если не является, дополнительные условия, при которых требуется его заполнение. Правила допустимости задаются посредством определения множества или диапазона допустимых значений и в некоторых случаях дополняются условными ограничениями, определяемыми через связи между полями. Например, значение почтового индекса должно быть не только допустимым само по себе, но и не противоречить коду региона. Следует также определять и правила, действующие на уровне набора данных. Например, каждому клиенту должен соответствовать допустимый почтовый адрес.

Определение правил качества данных — задача трудная по той причине, что большинство людей не приучено к осмыслению данных на языке правил. Поэтому, возможно, потребуется подбираться к формулировке правил окольными путями, задавая заинтересованным лицам вопросы не о самих правилах, а о требуемых характеристиках данных на входе и выходе того или иного бизнес-процесса. Полезно также расспросить сотрудников о болевых точках; о том, чем оборачивается отсутствие или ошибочность какого-то элемента данных; как они выявляют проблемы, по

каким признакам распознают дефектные данные, и т. д. и т. п. Важно помнить, что всех негласных правил вам всё равно не выявить, но и неполного набора обычно вполне достаточно для первичной оценки состояния данных. Выявление и уточнение действующих правил — процесс непрерывный и бесконечный. Один из лучших способов подобраться вплотную к точным формулировкам правил — распространить результаты проведенной экспертизы. Ознакомившись с ними, заинтересованные лица получают возможность взглянуть на свои данные под новым углом, и это может помочь им четче сформулировать правила, которыми они руководствуются, предъявляя те или иные требования к данным.

2.4 Проведение первичной оценки качества данных

По завершении выявления критических потребностей бизнеса и данных, необходимых для их удовлетворения, приходит черед важнейшего этапа первичной экспертизы качества данных, который состоит в оценке фактически имеющихся данных путем всестороннего изучения их содержания и взаимосвязей между элементами на предмет сравнения реального состояния данных с ранее выявленными бизнес-правилами и ожиданиями. Первоначально аналитики выявляют массу всяческих неожиданностей — не отраженные в документах и моделях связи и зависимости в наборах данных, скрытые правила, избыточные, несогласованные и противоречивые данные и т. п. — наряду с данными, полностью соответствующими правилам. После этого аналитикам качества данных совместно с распорядителями данных, экспертами в предметных областях и потребителями данных нужно рассортировать выявленные проблемы по категориям и в порядке приоритетности.

Цель первичной экспертизы качества данных — узнать о них всё необходимое для составления плана первоочередных мер по устранению самых вопиющих недостатков. Начинать лучше с небольшого, сфокусированного усилия, обеспечивающего быстрый результат и служащего подтверждением правильности концепции и демонстрацией того, как именно устроен и работает процесс повышения качества данных. Вот основные этапы первичной экспертизы данных:

- Определение целей, стимулов и направлений работы.
- Выявление данных, подлежащих оценке (для начала лучше сфокусироваться на небольшом наборе важнейших данных, а в некоторых случаях и на отдельно взятом ключевом элементе данных или конкретной проблеме с качеством данных).
- Выявление областей применения и основных потребителей данных.
- Выявление известных рисков, с которыми сопряжено использование оцениваемых данных, включая потенциальные пагубные последствия проблем для организации.
- Проверка данных на соответствие действующим и предлагаемым правилам.
- Документирование уровней несоответствия и типов проблем.
- Повторный углубленный анализ на основе первичных заключений с целью:
  - количественной оценки выявленных проблем;
  - приоритизации проблем по степени негативного влияния на бизнес;
  - проработки гипотез относительно корневых причин проблем с данными.

- Согласование приоритетов с руководством, экспертами и потребителями данных.
- Планирование первоочередных мер на основе полученных заключений.
  - Устранение выявленных проблем, в идеале вместе с корневыми причинами.
  - Меры по совершенствованию процессов с целью профилактики рецидивов.
  - Механизмы текущего контроля и учета.

2.5 Выявление и приоритизация потенциальных улучшений

Следующей после доказательства работоспособности процесса повышения качества целью становится его стратегическое применение. Для этого требуется выявить потенциальные усовершенствования и расставить приоритеты. Для выявления проблем и возможностей может потребоваться полномасштабное профилирование более крупных наборов данных, чем на первичном этапе. Или же — в качестве альтернативы — можно использовать и такие средства, как опрос заинтересованных сторон относительно испытываемых ими проблем с данными с последующим анализом источников проблем и их последствий для бизнеса. В любом случае на завершающей стадии расстановки приоритетов потребуется взвешенное сочетание объективного анализа данных с результатами обсуждений имеющихся проблем и пожеланий с заинтересованными сторонами.

Этапы полномасштабного профилирования и анализа данных в целом не отличаются от описанных применительно к первичной оценке небольшой выборки данных и включают: определение целей, изучение практик использования данных и связанных с ними рисков, оценку степени соответствия реализованных процессов правилам, документирование результатов и подтверждение их экспертами, использование полученной информации для внесения приоритетных исправлений и планирования дальнейших усилий по повышению качества данных. Однако на пути полномасштабного профилирования иногда возникают труднопреодолимые технические препятствия. В таких случаях требуются немалые усилия для того, чтобы скоординировать работу аналитической группы и получить обобщенные результаты, которые смогут послужить хорошей основой для понимания ситуации и определения возможностей с целью выработки единого эффективного плана действий, если таковой возможен. Крупномасштабное профилирование, как и первичное в ограниченном масштабе, должно вестись с прицелом на оптимизацию лишь критически важных данных.

Важно помнить, что профилирование данных — только первый этап анализа качества данных. Оно позволяет установить наличие проблем и очертить их круг, но не способно выявить ни их корневые причины, ни масштабы негативных последствий для бизнеса. Анализ последствий — отдельное направление исследований, требующее учета мнений всех заинтересованных сторон, участвующих в цепи передачи данных. Поэтому при планировании крупномасштабного профилирования уделите достаточно времени распространению результатов, приоритизации проблем и определению круга вопросов, требующих углубленного анализа.

2.6 Определение целей повышения качества данных

Знания, полученные в ходе предварительных экспертных оценок, служат базисом для формулировки конкретных целей программы качества данных. Планируемые улучшения могут принимать

самые различные формы — от элементарных (например, внедрение средств исправления ошибок ввода данных) до самых что ни на есть комплексных и позволяющих устранять корневые причины низкого качества данных. Планы исправлений и улучшений должны включать как точечные решения самых насущных проблем ценой минимальных затрат, так и долгосрочные стратегические изменения, направленные на решение имеющихся корневых проблем, а главное — реализацию механизмов защиты от возникновения новых корневых проблем на месте устраненных.

Не упускайте из виду и множество препятствий на пути к совершенствованию качества данных: системные ограничения, свойство данных устаревать, востребованность нуждающихся в доводке данных рабочими процессами, сложность ландшафта ИТ-среды, неподатливость организационной культуры и т. п. Во избежание осложнений, обусловленных подобными факторами и способных торпедировать программу качества данных как таковую, устанавливайте конкретные и реализуемые в краткосрочной перспективе цели, основанные на объективном и последовательном понимании ценности данных для бизнеса и наглядно показывающие финансовую отдачу от вложений в повышение их качества.

Например, целью ставится повышение показателя числа профилей клиентов с полными данными с 90% до 95% за счет совершенствования процессов и внесения изменений в конфигурационные настройки системы. Понятно, что для подтверждения улучшения ситуации требуется всего лишь сравнение исходных показателей полноты данных с текущими. А вот для доказательства реальной ценности наличия полных данных о клиентах полезна статистика иного рода: снижение числа жалоб/претензий, ускорение обработки заявок и т. п. Регистрируйте подобные показатели — и проще будет объяснить ценность усилий, направленных на повышение качества данных. Никому в руководстве нет дела до полноты и точности заполнения полей, если нет понимания, что неполнота и неточность приносят убытки. Иными словами, инвестиции в качество данных должны окупаться. Поэтому при выявлении проблем первым делом озаботьтесь обоснованием окупаемости затрат на их устранение с учетом следующих факторов:

- критичность (значимость, приоритетность) данных, затрагиваемых проблемой;
- объемы (доля) проблемных данных;
- период датирования (период) проблемных данных;
- число и тип бизнес-процессов, использующих проблемные данные;
- число клиентов, поставщиков, сотрудников и т. п., затрагиваемых проблемой;
- риски, проистекающие из проблемы;
- затраты на устранение корневых причин;
- затраты на потенциальные временные решения.

При экспертном анализе проблем, особенно таких, корневые причины которых установлены и требуют чисто технических решений, всегда стремитесь изыскивать не локальные, а системные решения, позволяющие предотвратить рецидивы проблем. Профилактика, в целом, обходится дешевле лечения, а иногда и на много порядков (см. главу 11).

2.7 Разработка и внедрение операционных процедур обеспечения качества данных

Многие программы качества данных начинаются с серии проектов по устранению выявленных недостатков. Для устойчивого обеспечения качества данных программа должна предусматривать план, позволяющий команде качества данных управлять правилами и стандартами качества данных, осуществлять мониторинг их соблюдения, выявлять и решать возникающие проблемы, а также предоставлять регулярную отчетность.

Для поддержки деятельности по обеспечению качества данных аналитики качества данных и распорядители данных должны, кроме того, привлекаться к документированию стандартов данных и бизнес-правил, а также к разработке требований к качеству данных, предъявляемых сторонним поставщикам.

2.7.1 Управление правилами качества данных

В процессе профилирования и анализа данных вскрываются действующие в организации гласные и негласные бизнес-правила и вырабатываются стандартизированные правила обеспечения качества данных. По достижении программой качества данных определенной ступени зрелости нужно приступать к встраиванию механизмов выявления и документирования правил в общий процесс развития и совершенствования информационной системы организации. Своевременное и явное определение правил позволяет:

- четко определять качественные характеристики данных;
- предъявлять требования к системным правкам и защитным механизмам, призванным не допустить порчи данных;
- формулировать требования к качеству данных поставщиков и третьих сторон;
- обосновывать измеримые показатели качества данных в системах мониторинга и отчетности.

Строго говоря, правила и стандарты качества данных — это критически важная категория метаданных. Следовательно, для эффективного управления стандартами и правилами необходимо обеспечивать соблюдение, по сути, тех же требований, которые предъявляются к управлению метаданными (см. главу 12). Сами же правила должны соответствовать следующим требованиям.

- **Согласованное документирование.** Определите и утвердите стандарты и шаблоны документации правил, чтобы все они были представлены в одном и том же формате и не допускали смысловых разночтений.
- **Определение в терминах измерений качества данных.** Измерения качества (см. раздел 1.3.3) помогают людям понимать, что именно оценивается. Согласованное использование измерений способствует эффективной оценке и решению возникших проблем.
- **Привязка к бизнеса-эффектам.** Измерения качества помогают понимать характер проблем, но не являются самоцелью. Поэтому стандарты и правила должны определяться с прямым

указанием на их роль в обеспечении успешной работы организации. Показатели, не связанные с бизнес-процессами, не стоят того, чтобы их оценивать.

- **Подтверждение результатами анализа данных.** Аналитики качества данных не должны определять правила на основе догадок. Все правила подлежат предварительному тестированию на реальных данных. Во многих случаях правила помогают выявлять проблемы с данными. Но бывает и так, что анализ результатов тестирования правил показывает непригодность или неполноту самих правил.
- **Подтверждение экспертами в предметных областях.** Цель правил — описать, как должны выглядеть данные. Зачастую, не зная деталей реализации рабочих процессов изнутри организации или подразделения, судить о корректности применения правил к предмету, описываемому данными, и, как следствие, к самим данным весьма затруднительно, и лучше проконсультироваться с экспертами в предметных областях, которые либо подтвердят годность правил, либо разъяснят аналитикам истинный смысл полученных ими результатов.
- **Доступность всем потребителям данных.** Все потребители данных должны иметь доступ ко всем задокументированным правилам, применимым к этим данным. Это нужно и для лучшего понимания данных пользователями, и для обеспечения полноты их применения, и для выявления возможных противоречий или неполноты самих правил. Кроме того, у потребителей данных должна иметься возможность получать разъяснения точного смысла и порядка применения правил и оставлять отзывы и предложения.

2.7.2 Измерение и мониторинг показателей качества данных

Операционные процедуры управления качеством данных основаны на возможности измерения и мониторинга текущих показателей качества. Организация измерения и мониторинга нужна по двум одинаково важным причинам:

- необходимость информирования потребителей данных об уровне их качества;
- необходимость управления информационными рисками, обусловленными изменениями в бизнес-процессах и технологических процессах.

Некоторые измеримые показатели позволяют оценивать деятельность по обоим указанным направлениям. Разработка измеримых показателей должна вестись на основе результатов экспертного анализа данных и выявленных корневых причин имеющихся проблем. Метрики, предназначенные для информирования потребителей, относятся к критическим элементам данных и связям, которые в случае нарушений оказывают прямое негативное влияние на бизнес-процессы. Метрики, предназначенные для минимизации риска, используются для мониторинга данных и отношений, которые имеют проблемы в анамнезе или не исключают их возникновения в перспективе. Например, если данные сохраняются после применения к ним набора правил ETL (извлечения, преобразования, загрузки), а сами правила ETL зависят от бизнес-процессов и могут меняться

в случае каких-либо изменений в этих бизнес-процессах, необходим мониторинг соответствующих данных с целью своевременного выявления изменений в процессах по изменениям в данных.

Важно не забывать об имевшихся в прошлом проблемах и во избежание рецидивов включать их симптомы в критерии выявления рисков. Например, если в прошлом неоднократно возникали проблемы с производными данными, получаемыми путем расчетов по сложным формулам, значит, все формулы расчетов таких данных подлежат проверке на корректность во избежание повторения подобных ошибок в будущем. В большинстве случаев целесообразно придерживаться следующего правила: выявив и устранив проблему с какой-либо функцией обработки данных, поставьте под наблюдение с помощью контрольных показателей все аналогичные функции, чтобы в будущем подобные проблемы в случае их возникновения выявлялись автоматически.

Результаты измерений могут фиксироваться на двух уровнях — детализации исполнения отдельных правил и суммарных результатов (статистики) исполнения всех правил. Каждое правило должно иметь стандартный, целевой или пороговый показатель для сравнения. Обычно для этого используется статистическая функция расчета доли или процента корректных или некорректных данных, например:

$$\text{ПригодныеДанные}\,(r) = \frac{(\text{ПровереноЗаписей}\,(r) - \text{ОтбракованоЗаписей}\,(r))}{\text{ПровереноЗаписей}\,(r)} \times 100\%$$

$$\text{НепригодныеДанные}\,(r) = \frac{\text{ОтбракованоЗаписей}\,(r)}{\text{ПровереноЗаписей}\,(r)} \times 100\%\,,$$

где r — правило проверки. Например, если по результатам проверки 10 000 записей на предмет соблюдения бизнес-правила r выявлено 560 записей, не соответствующих r, то в рассматриваемом примере мы получим результаты: ПригодныеДанные (r) = 9440/10 000 = 94,4%, НепригодныеДанные (r) = 560/10 000 = 5,6%.

Табличное представление определений и результатов измерений показателей качества данных, подобное приведенному в примере (см. табл. 30), помогает встраивать различные метрики и индикаторы качества в отчеты, получать сводные показатели и пояснять важные для понимания моменты. Отчет может носить более формализованный характер, если адресован проектам, отвечающим за устранение выявленных проблем. Отфильтрованные отчеты могут быть полезны распорядителям данных или бизнес-аналитикам, выявляющим тенденции и факторы влияния. Таблица 30 содержит примеры правил, сконструированных подобным образом. Там, где это применимо и уместно, результаты могут приводиться в двух выражениях — позитивном (% данных, соответствующих правилам и ожиданиям) и негативном (% данных, не соответствующих правилам).

Таблица 30. Примеры метрик качества данных

Измерение и бизнес-правило	Замеры	Метрики	Индикатор статуса
Полнота **Бизнес-правило 1: заполнение обязательных полей**	Доля записей с заполненными обязательными полями от общего числа записей	Результат деления числа записей с заполненными обязательными полями на общее число записей в таблице или базе данных, умноженный на 100 для перевода в процентное выражение	**Неприемлемо:** соблюдение < 80% несоблюдение > 20%
Пример 1: заполнение поля Индекс в таблице почтовых адресов	**Заполнено:** 700 000 **Не заполнено**: 300 000 **Итого:** 1 000 000	**Соблюдение (%):** 700 000/1 000 000 × 100% = 70% **Несоблюдение (%):** 300 000/1 000 000 × 100% = 30%	**Результат в примере:** неприемлемо
Уникальность **Бизнес-правило 2: каждый экземпляр объекта должен быть представлен в таблице единственной записью**	Доля записей в таблице, у которых выявлен хотя бы один дубликат	Результат деления числа записей с дубликатами, умноженный на 100 для перевода в процентное выражение	**Неприемлемо:** > 0%
Пример 2: в справочном списке Почтовые индексы каждый индекс должен быть представлен одной и только одной записью	**Записей с дубликатами:** 10 000 **Всего записей:** 1 000 000	10 000/1 000 000 × 100 = 1,0% почтовых индексов представлены в двух и более строках	**Результат в примере:** неприемлемо
Актуальность **Бизнес-правило 3: записи должны обновляться в установленные сроки**	Доля транзакций, не отражаемых как завершенные из-за невозможности получить обновленные данные от службы синхронизации	Число не завершенных в срок транзакций, отнесенное к числу попыток завершения транзакций, умноженное на 100 для перевода в процентное выражение	**Неприемлемо:** соблюдение < 99% несоблюдение > 1%
Пример 3: запись о завершении биржевой транзакции должна поступать не позднее 5 минут после транзакции	**Незавершенных транзакций:** 2000 **Попыток завершения транзакций:** 1 000 000	**Соблюдено:** (1 000 000 – 2000) / 1 000 000 × 100 = 99,8% записей о завершении транзакций поступило в срок **Не соблюдено**: 2000/1 000 000 × 100 = 0,20% записей о завершении транзакций поступило в срок	**Результат в примере:** приемлемо
Допустимость **Бизнес-правило 4: если значение в поле X = V1, значение в поле Y должно = V1′**	Доля записей, где правило соблюдено	Отношение числа записей, где правило соблюдено к общему числу записей со значением X = V1, умноженное на 100 для перевода в процентное выражение	**Неприемлемо:** < 100%

Измерение и бизнес-правило	Замеры	Метрики	Индикатор статуса
Пример 4: счета выставляются при отгрузке товара	Число записей с кодами статусов **отгружено = да** И **счет = да**: 999 000 Число записей с кодами статусов **отгружено = да** ИЛИ **счет = да**: 1 000 000	**Соблюдено:** 999 000/1 000 000 × 100 = 99,9% **Не соблюдено**: (1 000 000 – 999 000)/1 000 000 × 100 = 0,10% записей содержат только отметку об отгрузке товара *или* выставлении счета, что является нарушением правила	**Результат в примере:** неприемлемо

Правила проверки качества данных — основа операционного управления качеством данных. Поддержка правил может интегрироваться в сервисы приложений или сервисы данных, которые поддерживают жизненный цикл данных, либо посредством такого коммерческого ПО, как инструменты контроля качества данных, средства реализации правил, средства мониторинга и отчетности, либо с помощью приложений собственной разработки или созданных по заказу.

Обеспечивайте непрерывный мониторинг всех потоков данных посредством включения в них промежуточных процессов оперативного контроля и измерения параметров качества. Автоматизированный мониторинг соответствия данных правилам качества можно проводить как в процессе потоковой передачи (обработка в потоке), так и с помощью пакетной обработки данных. Измерения могут вестись на трех уровнях детализации данных: значений элементов или полей; экземпляров сущностей или табличных записей; наборов данных. Таблица 31 описывает основные методы сбора информации о результатах измерения характеристик данных. Обработка в потоке обычно используется для проверки качества при создании данных или при их перемещении от одной фазы обработки к другой, а пакетная — для проверки массивов данных, накопленных в файлах или таблицах баз данных, перед их интеграцией в хранилище. Для массивов данных методы обработки в потоке, как правило, не подходят, поскольку из-за необходимости проверки соблюдения правил, касающихся статистических параметров, требуется обработка сразу всего массива.

Таблица 31. Методы мониторинга качества данных

Уровень детализации	Обработка в потоке	Пакетная обработка
Элемент/Поле	Проверка ввода/изменений в приложениях Службы валидации значений данных Специальные программируемые приложения	Прямые запросы Средства профилирования и/или анализа данных
Экземпляр/ Запись	Проверка ввода/изменений в приложениях Службы валидации записей данных Специальные программируемые приложения	Прямые запросы Средства профилирования и/или анализа данных
Набор/Массив	Средства проверки при передаче данных из процесса в процесс (между стадиями обработки)	Прямые запросы Средства профилирования и/или анализа данных

Учет результатов измерения и мониторинга как в операционных процедурах, так и в отчетности обеспечивает возможность непрерывного мониторинга уровней качества данных с целью обратной связи и совершенствования процессов генерирования и/или сбора данных.

2.7.3 Разработка операционных процедур выявления и устранения проблемных вопросов

Средства текущего мониторинга позволяют лишь сигнализировать о возникновении проблем, а оперативно разбираться с их причинами и изыскивать эффективные способы исправления ситуации должна команда качества данных. Для этого входящие в ее состав специалисты должны располагать тщательно проработанными процедурами, регламентирующими порядок их действий по следующим направлениям.

- **Диагностика проблемного вопроса** состоит в решении следующих объективных задач: анализ симптомов сбоя; отслеживание происхождения проблемных данных вплоть до первоисточника; выявление проблемы и места ее возникновения; установление корневых причин. Соответствующий процедурный регламент должен описывать, как именно команда качества данных:
  - анализирует проблемы с данными в контексте объективно имеющейся информации о прохождении проблемными данными этапов обработки в ИТ-системах и локализует участок или процесс, где возникла ошибка;
  - проверяет, не было ли внесено в эксплуатационную среду каких-либо изменений, которые могли спровоцировать или повлечь ошибки в работе системы;
  - оценивает, не было ли возникновение проблем с качеством данных обусловлено вмешательством сторонних факторов — например, сбоев в работе смежных систем или процессов;
  - определяет, не является ли проблема прямым следствием получения некондиционных данных из внешних источников.
  - ВАЖНО: к выявлению и анализу корневых причин должны привлекаться эксперты в предметных областях как из технических, так и из бизнес-подразделений. Команда качества данных должна координировать совместную работу, но конечный успех зависит от слаженности взаимодействия между всеми функциональными подразделениями.
- **Формулировка перечня возможных корректирующих мероприятий:** в зависимости от результатов диагностики перечень альтернативных вариантов разрешения проблемной ситуации может предусматривать:
  - устранение нетехнических корневых причин (слабостей в подготовке и руководстве, нечеткого распределения обязанностей и полномочий и т. п.);
  - изменение ИТ-систем с целью устранения корневых причин технического характера;
  - доработку средств и механизмов контроля во избежание рецидивов;
  - введение дополнительных проверок и средств мониторинга;
  - прямую корректировку ошибочных данных;

- игнорирование ошибок, если они не оказывают существенного влияния на результаты или если их исправление обходится дороже потенциального ущерба.

- **Разрешение проблем**: выявив доступные варианты, команда качества данных должна обсудить их с владельцами бизнес-данных и совместно с ними определить наилучший вариант разрешения проблемной ситуации. Для этого этапа должны быть проработаны процедуры, детально описывающие порядок анализа следующего круга вопросов:
  - сравнительная оценка альтернативных вариантов по критериям эффективности и/или полезности затрат;
  - выбор, доработка и рекомендация оптимального варианта;
  - выработка и предоставление плана реализации решения;
  - реализация принятого решения, включая внедрение необходимых доработок.

Решения, принимаемые в процессе устранения проблемы, должны фиксироваться в системе отслеживания инцидентов. Данные из этой системы (если в ней обеспечен должный уровень их ведения) позволяют получить глубокое представление о наиболее распространенных причинах проблем и стоимости их устранения. Не забывайте указывать все подробности, включая описание проблемного вопроса, его корневые причины, варианты исправления и окончательное решение.

В системе отслеживания собираются данные обо всех аспектах обработки проблемных данных до и после реализации решения, включая распределение рабочих обязанностей, масштабы, объемы и частоту возникновения проблем, время, прошедшее с момента их возникновения до первичного выявления, диагностики, утверждения плана решения и окончательного устранения. Эти метрики служат ценным исходным материалом для глубокого осмысления текущей организации рабочих процессов, загрузки систем, распределения ресурсов, а также важными контрольными точками состояния данных, по которым можно отслеживать важные события при планировании будущих усилий по совершенствованию оперативного управления качеством данных.

Данные отслеживания инцидентов полезны и для потребителей информации. Располагая исправленными данными и зная о внесении исправлений, можно принимать ответственные решения без опасения, — напротив, с осознанием того, что ошибки устранены, а данные исправлены, и с полным пониманием того, как именно всё это было сделано. Одно это служит веской причиной для документирования методов изменения правил учета данных и обоснований изменений. Такая документация должна быть доступна не только потребителям данных, но и разработчикам программного обеспечения, чтобы у них была возможность учитывать полученные уроки при разработке кодов будущих продуктов. Тем, кто реализует изменения, их необходимость бывает очевидной, но историю изменений вести нужно всё равно, хотя бы ради того, чтобы она не изгладилась из памяти и оставалась доступной будущим поколениям потребителей данных. Отслеживание инцидентов с качеством данных требует от сотрудников навыков классификации, регистрации и мониторинга; при необходимости можно предусмотреть дополнительную

профессиональную переподготовку. Кроме того, для обеспечения эффективного отслеживания необходимо соблюдение следующих условий.

- **Стандартизация проблемных вопросов качества данных и мер по их устранению.** Терминология, используемая для описания проблем с данными, может сильно варьироваться в зависимости от отрасли или специализации, поэтому полезно разработать глоссарий стандартных терминов и понятий. Наличие глоссария существенно упростит не только классификацию проблем и решений, но и ведение отчетности. Кроме того, стандартизация упрощает количественную оценку объемов проблем и работ, выявление тенденций, схем и зависимостей во взаимодействиях между системами и/или людьми, а также всесторонний учет результативности деятельности по обеспечению качества данных. При выявлении новых обстоятельств или корневых причин проблему всегда можно переквалифицировать из одного класса в другой.
- **Определение процесса назначения ответственных**. Операционные процедуры должны предусматривать назначение аналитиков, непосредственно отвечающих за расследование инцидентов и диагностику причин, и специалистов, прорабатывающих альтернативные решения проблем. При назначении ответственных не лишним будет справляться в системе регистрации инцидентов, у кого из сотрудников имеется опыт решения аналогичных проблем в прошлом.
- **Управление процедурой эскалации проблемных вопросов.** Управление проблемными вопросами требует наличия четко определенной системы эскалации вопросов в зависимости от их масштаба, тяжести, запущенности или остроты. Порядок передачи проблем по инстанциям должен быть отражен в соглашении об уровне обслуживания (SLA) и реализован технически с использованием системы отслеживания инцидентов (что способствует повышению оперативности и эффективности разрешения проблем с данными).
- **Управление потоком работ по разрешению проблемных вопросов.** SLA в области качества данных определяет круг задач мониторинга, контроля и разрешения проблемных вопросов, каждой из которых соответствует четко определенный поток работ. Поддержка управления этими потоками может быть оказана со стороны системы отслеживания инцидентов, предоставляющей информацию о ходе работ.

2.7.4 Определение соглашений об уровне обслуживания в области качества данных

Соглашение об уровне обслуживания в области качества данных определяет предъявляемые организацией требования к срокам разрешения проблемных вопросов качества данных для каждой из обслуживаемых систем. Кроме того, в SLA включается график проведения плановых проверок качества данных, способствующих своевременному выявлению и устранению проблем на ранней стадии их проявления, а со временем и снижению частоты возникновения проблем. Помимо выявления и анализа корневых причин, от операционных процедур обеспечения качества данных требуется еще и предоставление схемы их устранения в установленные соглашением сроки. Регулярные проверки качества данных, а также процедуры мониторинга и оперативного устранения

неполадок создают предпосылки для устранения проблем с качеством данных раньше, чем они приведут к негативным результатами для бизнеса. В SLA должны быть учтены следующие аспекты операционного контроля качества данных:

- элементы данных, подпадающие под действие соглашения;
- последствия ошибок и сбоев в данных для бизнеса;
- измерения качества, соответствующие каждому элементу данных;
- ожидаемые показатели качества каждого элемента по установленным для него параметрам в каждом приложении или системе в цепи прохождения данных по бизнес-процессам;
- методы измерения ожидаемых показателей;
- пороги допустимых отклонений по каждому измеримому показателю;
- должностные лица, которые должны уведомляться о случаях выхода данных за пороговые пределы допустимых отклонений параметров качества данных;
- временны́е графики и крайние сроки устранения или исправления проблем;
- стратегия передачи решений проблем на высший уровень руководства;
- премирование/штрафы.

SLA также фиксирует распределение ролей и обязанностей в рамках операционных процедур обеспечения качества данных. Кроме того, процедуры предусматривают формирование отчетности о соблюдении установленных бизнес-правил и мониторинг работы персонала, отвечающего за оперативную реакцию на инциденты с качеством данных. Распорядители данных и сотрудники, отвечающие за поддержку качества, обеспечивая необходимый уровень обслуживания, должны учитывать накладываемые на них ограничения (в связи с выполнением требований SLA) и включать деятельность по обеспечению качества в свои индивидуальные планы работ.

Для тех случаев, когда проблемные вопросы не разрешаются в установленные сроки, должна быть предусмотрена процедура вынесения обсуждения причин несоблюдения SLA на высшие уровни как руководства, так и управления данными. Соглашением определяются крайние сроки передачи уведомления, указываются вышестоящие руководители, которым по цепочке докладывается о ситуации, а также условия, при которых необходима эскалация. Имея набор правил качества данных, методики проверки и измерения показателей их соблюдения, допустимые пороги погрешностей, согласованные с бизнес-клиентами, и соглашения об уровнях обслуживания, команда качества данных располагает всем необходимым для мониторинга соблюдения параметров качества данных на уровне, удовлетворяющем бизнес, равно как и для объективной оценки собственной деятельности по выполнению процедур, предусмотренных на случай возникновения ошибок в данных.

Отчеты по выполнению SLA формируются на регулярной основе в соответствии с графиком, определяемым требованиями бизнеса и подразделений, осуществляющих операционную деятельность. В тех случаях, когда SLA предусматривает премирование/штрафы по итогам отчетных периодов, особое внимание в отчетах уделяется статистическим показателям и тенденциям.

2.7.5 Разработка системы отчетности о качестве данных

Работа по оценке качества данных и управления проблемными вопросами, связанными с данными, не принесет желаемого результата, если не будет дополняться отчетами, из которых потребители данных смогут получать исчерпывающее представление об их текущем состоянии. При разработке системы отчетности особое внимание следует уделить отражению следующих аспектов:

- ведомость оценки качества в целом, по различным категориям показателей и с разной глубиной детализации различных категорий, с ориентацией на различные целевые группы потребителей данных;
- тенденции изменения качества данных со временем с пояснениями методик оценки и смысла тенденций (позитивный или негативный);
- показатели, предусмотренные SLA, позволяющие судить, в частности, о своевременности диагностики причин и оперативности устранения проблем с качеством данных обслуживающим персоналом;
- показатели оценки управления проблемными вопросами в области данных, включая данные мониторинга состояния текущих проблемных вопросов;
- соблюдение командой качества данных установленных политик руководства данными;
- соблюдение ИТ-персоналом и бизнес-подразделениями установленных политик в области качества данных;
- позитивные результаты, достигнутые благодаря реализации проектов по повышению качества данных.

Метрики, используемые в отчетности, должны максимально соответствовать показателям качества данных, определенным в SLA, поскольку это способствует согласованности целей, преследуемых командой качества данных и ее бизнес-клиентами. Кроме того, программа качества данных должна отчитываться о положительных результатах, достигнутых в рамках проектов по повышению качества, и делать это лучше в терминах бизнеса, чтобы постоянно напоминать организации о том, что качество данных оказывает прямое влияние на ее клиентов.

3. ИНСТРУМЕНТЫ

Инструменты должны выбираться с учетом системной архитектуры и планируемых настроек еще на фазе планирования программы качества данных предприятия. Программное обеспечение для управления качеством данных обычно поставляется с готовым набором начальных настроек правил, но организациям нужно обязательно разрабатывать и задавать для каждого инструмента собственные правила, учитывающие специфику контекста и требующихся действий.

3.1 Инструменты профилирования данных

Инструменты профилирования данных позволяют собирать высокоуровневую статистику, дающую аналитикам возможность выявлять закономерности и тенденции и проводить первичную оценку различных параметров качества данных. Некоторые из них подходят и для текущего мониторинга, но особую важность средства профилирования данных имеют для анализа больших массивов данных с целью выявления проблем. Особенно хорошо с этим справляются инструменты профилирования, оснащенные средствами визуализации (см. главы 5 и 8, а также раздел 1.3.9 настоящей главы).

3.2 Инструменты формирования запросов к данным

Профилирование данных — лишь первый шаг по пути анализа данных, который позволяет выявлять потенциальные проблемы. После этого команде качества данных нужно детально разобраться с глубинными причинами проблем и определить закономерности, которые выведут на их источники. С этой целью можно, например, формировать аналитические запросы, позволяющие оценить другие аспекты качества данных, такие как уникальность и целостность.

3.3 Инструменты моделирования данных и средства ETL

Инструменты моделирования данных и средства реализации процессов извлечения, преобразования и загрузки (ETL) оказывают прямое влияние на качество данных. Если использовать их, обладая точным представлением о данных, они позволят существенно повысить качество. Применение же этих средств при отсутствии достаточных знаний о данных, которые предполагается обрабатывать, может привести к обратному эффекту. Участники команды качества данных должны совместно с разработчиками обеспечить минимизацию риска причинения ущерба данным и при этом постараться наиболее полно реализовать потенциал предоставляемых возможностей по моделированию и обработке данных с целью повышения их качества (см. главы 5, 8 и 11).

3.4 Шаблоны правил качества данных

Шаблоны правил качества помогают аналитикам фиксировать требования к данным. Они также служат мостом к взаимопониманию между сотрудниками бизнес-подразделений и технических служб. Согласованные формулировки правил упрощают перевод бизнес-потребностей в программный код, который может быть встроен в модуль обработки правил, анализатор данных в составе инструмента профилирования или средство интеграции. Шаблон может иметь несколько секций, предназначенных для определения бизнес-правил различных типов.

3.5 Репозитории метаданных

Как отмечалось в разделе 1.3.4, определения качества данных формулируются с помощью метаданных и сами становятся ценными метаданными. Команда качества данных должна тесно сотрудничать с коллегами из команды управления метаданными, чтобы гарантировать доступ потребителей данных к требованиям по качеству данных, правилам, результатам измерений и документации, описывающей различные проблемы.

4. МЕТОДЫ

4.1 Превентивные меры

Лучшая гарантия создания данных высокого качества — не допускать проникновения в организацию некачественных данных. Превентивные меры избавляют, как минимум, от риска повторения известных ошибок. Кроме того, проверять данные после того, как они успели попасть в среду эксплуатации, поздно: качество уже пострадало. Подходы к профилактике появления некачественных данных включают следующее.

- **Контроль на входе.** Создайте правила отбраковки, исключающие ввод или поступление некондиционных данных в систему.
- **Подготовка персонала, осуществляющего производство данных.** Нужно гарантировать, что сотрудники, работающие с информационными системами, понимают степень значимости данных, поставляемых ими пользователям и другим системам. Внедрите систему стимулирования или аттестационные оценки, учитывающие точность и полноту данных, а не только скорость их ввода.
- **Определение и обеспечение соблюдения правил.** Создайте подобие межсетевого экрана, который содержит сводную таблицу всех бизнес-правил качества данных и проверяет качество данных перед их использованием приложениями — например, центральным хранилищем данных. Такой фильтр данных может, например, автоматически проверять уровень качества данных, обработанных приложением, и, если он окажется ниже установленного значения, направлять уведомление аналитику данных о возникновении проблемы.
- **Контроль качества данных из внешних источников.** Изучите реализованные у поставщика данных процессы с целью проверки структур данных, используемых определений, а также происхождения и источников данных. Такая практика позволит оценить степень интегрируемости внешних данных с вашими системами и данными, а также отсечь попадание в вашу организацию данных сомнительного происхождения и/или качества, не говоря уже о данных, использование которых не было санкционировано правообладателем.
- **Внедрение практики руководства и распоряжения данными.** Убедитесь, что роли, обязанности и полномочия четко определены, и строго следите за соблюдением правил привлечения к работам, принятия решений и распределения ответственности за эффективное управление данными и информационными ресурсами (McGilvray, 2008). Совместно с распорядителями данных проведите полную ревизию процессов и механизмов генерирования, отправки и получения данных.
- **Формализованный контроль изменений.** Необходимо обеспечить обязательное предварительное тестирование всех изменений в хранимых данных до их переноса в среду эксплуатации. Во избежание внесения прямых изменений в обход нормальных рабочих процессов реализуйте все необходимые процедуры проверки.

4.2 Корректирующие меры

Корректирующие меры предусматриваются на случай возникновения и выявления проблем. Диагностика несоответствий данных критериям качества должна вестись систематически, а выявленные проблемы искореняться на уровне первопричин с целью минимизации издержек и рисков, которыми чревато регулярное исправление рецидивов. Решать проблему по месту ее возникновения — самая лучшая практика управления качеством данных. Обычно она подразумевает применение таких превентивных мер, которые устранят не просто причины выявленных проблем с качеством данных, но и саму возможность их повторного возникновения.

Общепринятыми являются три способа исправления данных.

- **Автоматизированное исправление** включает стандартизацию на основе правил, нормализацию и собственно исправление значений; при этом исправленные значения рассчитываются или генерируются и вносятся в поля данных также автоматически, безо всякого вмешательства человека. Пример: программа автоматизированного исправления почтовых адресов, отправляющая выявленные некорректные адреса в модуль стандартизации, который и приводит их в соответствие с нормами, используя правила, алгоритмы синтаксического анализа и стандартизации, а также справочные таблицы. Автоматическое исправление возможно только в средах с детально проработанными стандартами, едиными правилами и хорошо известной структурой распространенных ошибок. Объемы автоматических исправлений в таких системах со временем могут снижаться, если в среде реализована обратная связь с системами выше по потоку и туда отправляются сведения о выявленных ошибках и исправлениях.
- **Полуавтоматическое исправление (с ручным подтверждением)** отличается от первого подхода тем, что после автоматизированного исправления данные проходят этап ручной проверки и подтверждения перед сохранением. Можно настроить правила исправления адресов и фамилий/имен с определенным уровнем разрешающей способности при распознавании, чтобы исправления, в целом, вносились автоматически, но с присвоением некой оценки степени уверенности в их корректности. Исправления с оценкой уверенности выше пороговой могут сохраняться без проверки человеком, а остальные отправляются на утверждение распорядителю данных. Изучая структуру прошедших и не прошедших утверждение автоматических исправлений, по мере необходимости корректируйте правила и пороговое значение. Среды, где административный надзор требуется в силу чувствительности части данных в наборах (например, в системах MDM), служат хорошим примером показания к применению полуавтоматического исправления с ручным подтверждением.
- **Ручное исправление.** Случаются ситуации, когда полностью ручное исправление ошибок в данных — единственный доступный вариант либо по причине отсутствия технических средств автоматизации, либо по причине чрезвычайной чувствительности или важности данных, не допускающих внесения в них каких-либо правок без надзора уполномоченных лиц. В таких случаях для внесения исправлений в ручном режиме лучше предусмотреть

специальный интерфейс с элементами управления и полями редактирования, а также контрольный журнал регистрации всех правок. Вариант с отправкой исправленных записей сразу же в среду эксплуатации в таких ситуациях использовать чрезвычайно рискованно. Постарайтесь его избегать.

4.3 Программные модули проверки и аудита качества

Создавайте общедоступные повторно используемые программные модули многоцелевого назначения для поддержки регулярно повторяющихся процессов проверки и аудита качества данных и включайте их в библиотеку для разработчиков. В случае внесения изменений в функциональность модуля обновится и функциональность всех использующих его приложений. Подобные модули значительно упрощают сопровождение систем. Хорошо спроектированные блоки программного кода способствуют предотвращению множества проблем с качеством данных. Не менее важно и то, что такой подход обеспечивает согласованность выполняемых процессов. Там, где законодательством или отраслевыми регламентами предусмотрена обязательная отчетность, соответствующая строго определенному набору показателей качества, часто требуется еще и документированное подтверждение происхождения предъявляемых надзорным органам результатов. Модули проверки качества пригодны и для этого. Для предоставления в общий доступ данных со спорными параметрами качества или, напротив, высоко котирующихся данных используйте модули с полями примечаний для описания качественных характеристик и рейтингами достоверности.

4.4 Эффективные метрики качества данных

Критически важным компонентом управления качеством данных является разработка метрик, информирующих потребителей о характеристиках качества, которые наиболее важны для оценки степени пригодности данных к использованию. Измеримых параметров всегда имеется в избытке, но далеко не все из них актуальны и стоят времени и труда, затрачиваемых на их измерение и учет. Разрабатывая метрики, аналитикам качества данных следует учитывать следующие характеристики.

- **Измеримость.** Параметры качества должны быть измеримыми. К примеру, та же «актуальность» данных остается абстрактным и никак не проверяемым свойством, если отсутствуют четкие критерии определения степени актуальности информации. Даже столь очевидная характеристика, как «полнота» данных, также нуждается в определении объективной меры. Ожидаемые результаты должны поддаваться количественному определению в рамках дискретного диапазона значений.
- **Значимость для бизнеса.** Из множества доступных для измерения параметров далеко не все переводятся в полезные для бизнеса метрики. Прежде всего, результаты измерений должны интересовать потребителей данных. Ценность метрики с точки зрения бизнеса будет весьма сомнительной, если измеряемая величина никак не привязана ни к одному аспекту

бизнес-операций или производительности. Каждая метрика качества данных должна так или иначе отражать влияние данных на ключевые показатели бизнеса.

- **Приемлемость для бизнеса.** Измерения качества данных задают рамки бизнес-требований к качеству данных. Определение количественных показателей, увязанных с измерениями качества, позволяет предъявить потребителям самые веские доказательства соответствия данных всем предъявляемым требованиям. Соответствие должно определяться пороговыми уровнями приемлемости. Если оценка данных по какому-либо параметру не ниже пороговой, данные приемлемы для бизнеса. Если ниже, они не соответствуют предъявляемым требованиям.
- **Ответственность/Распоряжение.** Метрики должны быть понятны ключевым заинтересованным лицам (владельцам и распорядителям данных) и одобрены ими. Они должны оперативно уведомляться о выходе значений параметров качества за допустимые пределы, поскольку это означает, что данные перестали соответствовать ожиданиям. При этом владелец данных несет ответственность за сложившуюся ситуацию, а распорядитель — за принятие мер по исправлению.
- **Контролируемость.** Метрики должны отражать аспекты бизнеса, поддающиеся контролю. Иными словами, при выходе значения измеряемого параметра за пределы установленного допуска должна инициироваться процедура улучшения данных. Если же метрика не обеспечивает контроля ситуации, то она, возможно, является излишней.
- **Отслеживание тенденций.** Метрики дают организации возможность оценивать изменения качества данных с течением времени. Отслеживание изменений позволяет команде качества данных проводить мониторинг соблюдения условий SLA и соглашений о совместном использовании данных, а также наглядно подтверждать эффективность принимаемых мер по обеспечению надлежащего качества данных и услуг по их предоставлению. После стабилизации процессов работы с данными можно переходить к применению методов статистического управления процессами. Они позволяют не только выявлять текущие тенденции, но и составлять незаменимые в любом бизнесе прогнозы на будущее.

4.5 Статистическое управление процессами

Статистическое управление процессами (Statistical Process Control, SPC) — разработанный в 1920-х годах метод технического контроля качества промышленной продукции по вводным, промежуточным и выходным параметрам технологических процессов[1]. Впоследствии алгоритм SPC получил широкое распространение в самых разных отраслях и входит в стандартный набор методов управления качеством — в том числе и качеством данных[2]. Применительно к качеству данных процесс определяется просто как последовательность исполняемых операций (шагов) по

[1] Концепцию SPC (*сокр.* от *англ.* Statistical Process Control), основанную на использовании описанных ниже контрольных карт, разработал в 1924 г. создатель теории непрерывного управления качеством Уолтер Шухарт. — *Примеч. пер.*

[2] См.: Redman (1996 и 2001), Loshin (2000), Sebastian-Coleman (2013), Jugulum (2014).

преобразованию входных данных в выходные. Основополагающий постулат SPC: согласованный процесс обработки согласованных входных данных дает согласованные результаты (данные) на выходе. В рамках применения метода измеряется некое текущее усредненное значение (например, среднее арифметическое, медиана, среднестатистическое) и некий показатель разброса вокруг него (например, диапазон, дисперсия, среднеквадратичное отклонение). Для них определяются допуски.

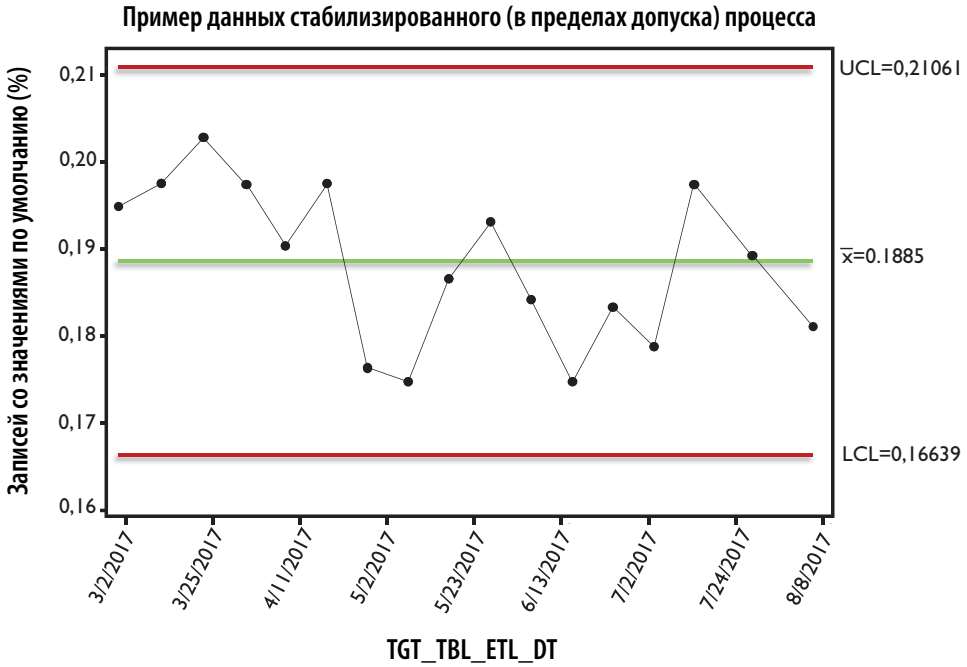


Рисунок 95. Контрольная карта Шухарта

Основной инструмент SPC — контрольная карта (см. рис. 95), которая представляет собой не что иное, как график динамического ряда контрольных значений с рассчитанным по ним средним значением ($\bar{x}$) и предельно допустимыми (контрольными) отклонениями. В рамках стабильного процесса выход результатов измерений за пределы контрольного допуска сигнализирует об особом случае.

SPC позволяет отличать предсказуемые результаты от непредсказуемых по степени отклонения внутренних показателей процессов. Отклонения в процессе подразделяются на обусловленные общими причинами, заложенными в самом процессе, и особыми причинами, привнесенными извне и потому непредсказуемыми или возникающими спорадически. При отсутствии особых причин среди источников разброса значений система считается устойчивой и статистически контролируемой, что и позволяет вычислять среднее значение и диапазон стандартных отклонений, по которым затем и выявляются аномальные изменения.

Применение SPC к измерениям показателей качества данных подразумевает, что данные рассматриваются как продукт некоторого процесса. Иногда процесс создания данных описывается

элементарно (например, человек заполняет поля формы ввода). Встречаются, однако, и крайне сложные процессы (например, набор алгоритмов накопления клинико-статистических данных с целью сравнительного анализа эффективности различных клинических протоколов). Но главный принцип всегда один: при единообразном вводе и единообразной обработке идентичных данных должен всякий раз получаться идентичный результат. Если же вводные данные или исполняемые процессы изменяются, то изменяются и результаты. Все эти компоненты поддаются измерению и оценке. Результаты измерений можно использовать для выявления статистически значимых отклонений — особых случаев. Знание причин, приводящих к особым случаям, помогает минимизировать риски, связанные со сбором и обработкой данных.

Статистическое управление процессами позволяет выявлять сбои и совершенствовать процессы. Но прежде всего нужно накопить статистически значимый ряд первичных измерений, выявить особые случаи и устранить их причины. После этого процесс принимает стабилизированный характер и ставится под статистический контроль. Для этого нужно предусмотреть текущие контрольные измерения с целью своевременного выявления сверхнормативных отклонений. Ранняя диагностика проблем упрощает расследование их корневых причин. Результаты измерений могут использоваться также и для изыскания способов сужения диапазона разброса значений вследствие общих причин, что, в свою очередь, повысит эффективность процессов.

4.6 Выявление и анализ корневых причин

Корневой причиной проблемы называется некий фактор, устранение которого приводит к исчезновению проблемы как таковой. Отсюда же и выражение «искоренить проблему». Анализ корневых причин заключается во всестороннем изучении всевозможных факторов, вносящих вклад в наблюдаемые проблемы, с целью докопаться до первопричин их возникновения и идентифицировать условие или явление, устранение которого снимет проблему как таковую.

Поясним на примере из области управления качеством данных. Скажем, имеется процесс, на вход которого поступает файл с данными, который передается клиентом раз в месяц. Результаты контрольных измерений показывают, что в январе, апреле, июле и октябре качество данных стабильно ниже, чем в других месяцах. Проверяем сроки поступления данных и выясняем, что в декабре, марте, июне и сентябре файлы поступали 30-го числа, а в остальные месяцы — 25-го числа. Смотрим дальше и выясняем, что файлы передаются из бухгалтерии, которая как раз в конце этих месяцев занята закрытием квартальной финансовой отчетности, и ее сотрудникам в эти дни попросту не до проверки качества данных, включаемых в файлы, предназначаемые для передачи, поскольку отчетность — главный приоритет. Отсюда вывод: корневая причина низкого качества данных, поступающих в конце квартала, — в том, что составляются они в последний момент и в спешке вследствие конфликта приоритетов. Возможное решение — сместить график передачи ежемесячных данных на середину месяца.

Распространенные приемы поиска корневых причин включают анализ по Парето (правило 80/20), причинно-следственную диаграмму Исикавы («рыбий скелет»), логистический анализ, анализ процессов и метод «пяти почему» (McGilvray, 2008).

5. РЕКОМЕНДАЦИИ ПО ВНЕДРЕНИЮ

Повышение качества данных в масштабах организации — задача непростая, даже если мероприятия по ее решению предусмотрены в программе руководства данными и поддержаны руководителями верхнего уровня. В академических кругах давно перешла в разряд классики дискуссия на тему: «Как лучше внедрять программу управления качеством — сверху вниз или снизу вверх?» На практике обычно лучше всего работает гибридный подход: сверху вниз идут спонсорская поддержка, координация и ресурсы, а снизу вверх — выявление проблем, инициативы и пошаговые достижения.

Повышение качества данных требует изменения мышления и поведения людей в отношении данных. А изменение культуры никогда не бывает простым. Оно требует планирования, обучения и закрепления знаний (см. главу 17). В то время как специфика культурного перевоспитания зависит от конкретной организации, на практике большинству программ качества данных так или иначе приходится планировать следующие аспекты внедрения.

- **Метрики для оценки ценности данных и ущерба, причиняемого данными низкого качества.** Один из способов повышения уровня осознания необходимости управления качеством данных в масштабах организации — через наглядную демонстрацию (с использованием соответствующих метрик) ценности данных и окупаемости вложения в их совершенствование. Подобные метрики (не путать с показателями качества данных) служат лучшим обоснованием необходимости финансирования программы качества данных и меняют в лучшую сторону отношение к данным и поведение как рядовых сотрудников, так и руководства (см. главу 11).
- **Операционная модель взаимодействий между ИТ- и бизнес-подразделениями.** Деловым людям не нужно лишний раз рассказывать о важности и ценности данных: они и сами прекрасно понимают, какие данные им жизненно необходимы и как много они значат для бизнеса. Зато администраторы баз данных и другие специалисты по ИТ лучше разбираются в том, где и как хранятся и обрабатываются данные, — и кому, как не им, переводить бизнес-определения и требования к качеству данных на язык запросов и кодов, позволяющих выявлять некондиционные записи (см. главу 11)?
- **Изменения в порядке выполнения проектов.** Общее руководство проектами должно гарантировать включение в каждый проект и финансирование мер по обеспечению качества данных (таких, как профилирование и анализ данных, определение требований к качеству, механизмов предупреждения, устранения и исправления проблем, контрольные измерения и показатели и т. п.). Благоразумие диктует необходимость планирования требований к качеству данных и средств предупреждения и раннего выявления проблем на начальной стадии любого проекта.
- **Изменения в бизнес-процессах.** Качество данных повышается за счет совершенствования процессов, в рамках которых они создаются. Рабочей группе программы качества данных

нужно дать возможность оценивать не только технические, но и бизнес-процессы, влияющие на качество данных, и рекомендовать необходимые изменения.

- **Финансирование проектов по исправлению и улучшению данных.** В некоторых организациях допускают серьезную ошибку, не планируя никаких мер по исправлению имеющихся данных даже на фоне понимания наличия проблем. Сами собой данные в норму не приходят. Формальный анализ полезности затрат на проекты исправления и совершенствования данных служит лучшим способом взвесить все плюсы и минусы и определить приоритетные направления работы по повышению качества данных.
- **Финансирование деятельности по обеспечению качества данных.** Обеспечение стабильно высокого качества данных требует текущего мониторинга данных, учета результатов и оперативного управления решением проблем по мере их выявления.

5.1 Оценка готовности / Оценка рисков

Большинство зависимых от данных организаций имеют широкий выбор возможностей для совершенствования. Но степень поддержки программы качества данных зависит прежде всего от зрелости организации в плане управления данными в целом (см. главу 15). Готовность организации к восприятию и внедрению передовых практик управления качеством данных можно оценить по следующим признакам и характеристикам.

- **Приверженность руководства управлению данными как стратегическим активом.** Прежде чем запрашивать финансирование программы качества данных, стоит задаться вопросом, хорошо ли высшее руководство понимает роль данных в организации в целом. В какой степени наверху осознают ценность данных для достижения стратегических целей? С какими рисками ассоциируются некачественные данные в представлении руководства? Насколько оно осведомлено о преимуществах грамотного распоряжения данными? Наконец, насколько оптимистичны прогнозы относительно возможности реального изменения организационной культуры в сторону всеобщей поддержки усилий по повышению качества данных?
- **Текущий уровень понимания организацией качества своих данных.** Большинству организаций для того, чтобы отважиться на поиски приключений, подстерегающих на пути к повышению качества данных, для начала требуется в полной мере осознать, что множество текущих затруднений и болевых точек напрямую обусловлены низким качеством данных и прямым образом указывают на необходимость совершенствования программы качества данных. Тут крайне важен разбор ошибок и их последствий. Нужно наглядно показывать: вот некачественные данные, а вот прямые и косвенные издержки, которые из-за них несет организация. Анализ болевых точек к тому же помогает определять и приоритетные проекты по совершенствованию управления качеством данных.
- **Объективная оценка состояния данных** служит первым шагом на пути выявления болевых точек и потребностей в усовершенствованиях. Для измерения и описания данных можно использовать методы профилирования и статистического анализа, а также численной оценки

известных проблем и болевых точек. Не зная реального состояния данных, команде программы качества данных затруднительно будет определить приоритетные задачи по их совершенствованию.

- **Риски, связанные с созданием, обработкой или использованием данных.** Выявление потенциальных проблем с данными и обусловленных ими угроз благополучию организации закладывает основу управления рисками. В организации, знать не желающей о подобных рисках, заручиться поддержкой программы качества данных бывает проблематично.
- **Культурная и техническая готовность к масштабируемому мониторингу качества данных.** На качестве данных могут негативно сказываться как бизнес-процессы, так и технические процессы. Для повышения качества требуется сотрудничество между бизнес- и ИТ-подразделениями. Если конструктивные рабочие отношения между ними не налажены, на прогресс рассчитывать затруднительно.

Заключения по результатам оценки готовности помогут определиться с отправными точками и темпами изменений. Они же могут послужить системой координат для прокладывания дорожных карт к поставленным целям. Если в организации налицо крепкая поддержка усилий по повышению качества данных и хорошее знание собственных данных, возможен вариант безотлагательного запуска полномасштабной стратегической программы качества данных. Если же в организации имеют слабое представление о реальном состоянии данных, то для начала, вероятно, лучше сосредоточить усилия на выработке такого понимания и лишь после осознания необходимости совершенствования приступать к выработке полноценной стратегии.

5.2 Организационные и культурные изменения

Качество данных повышается не за счет коллекционирования программных средств и умных книг, а благодаря выработке у всех руководителей и сотрудников особого образа мышления, который помогает не забывать о качестве данных и нуждах потребителей в процессе рутинной работы. Для привития соответствующего отношения к качеству данных часто требуются значительные усилия по изменению сложившейся культуры, немыслимые без дальновидного лидерства (см. главу 17).

Первым шагом становится разъяснительная работа относительно важности роли данных для организации. Все сотрудники должны действовать ответственно, не замалчивать возникающие вопросы по поводу качества данных и обеспечивать высокое качество производимой ими самими информации, предназначенной для использования другими людьми. Каждый, кто прикасается к данным, способен повлиять на их качество. Обеспечение качества данных — предмет заботы всей организации, а не только команды качества данных или ИТ.

Ценить каждого новообретенного и сохраненного клиента учат любого менеджера; таким же образом научите сотрудников ценить и данные: пусть знают, какими издержками чреваты некачественные данные и что именно приводит к снижению их качества. Например, если данные о клиенте неполны, могут возникнуть ошибки с комплектацией или спецификацией отгружаемого

товара и, как следствие, прямые и косвенные убытки. В лучшем случае последует возврат товара, но возможны и жалобы, и рекламации, и долгие объяснения, приводящие к непродуктивному расходу времени операторов службы работы с клиентами, а потенциально — судебные иски и репутационный ущерб. Неполнота данных о клиентах из-за отсутствия четких правил должна беспокоить всех и каждого в организации, поскольку может обернуться несоблюдением требований и стандартов кем угодно и повредить всем и каждому.

В конечном счете сотрудникам нужно научиться новому образу мышления и действий, чтобы производить высококачественные данные и управлять данными таким образом, чтобы гарантировать высокое качество работы. А это требует методичной подготовки и регулярного закрепления полученных знаний по следующему кругу вопросов:

- основные и наиболее распространенные причины проблем с данными;
- экосистемы данных организации и причины, требующие системного подхода к повышению качества данных на уровне предприятия;
- последствия низкого качества данных;
- причины, по которым повышение качества данных должно вестись на непрерывной основе (а не носить характер разовых вмешательств);
- «язык данных» как средство формулировки стратегии, отчетности, ожиданий и показателей удовлетворенности клиентов;

В программу подготовки должны входить также вводные разъяснения к любым готовящимся изменениям процессов, включая наглядные и доказательные подтверждения их позитивного влияния на качество данных.

6. РУКОВОДСТВО КАЧЕСТВОМ ДАННЫХ

Эффективность программы качества данных значительно повышается, если она встроена в программу руководства данными. Впрочем, зачастую руководство данными в масштабах организации как раз и внедряется с целью устранения проблем с качеством (см. главу 3). В любом случае интеграция усилий по обеспечению качества данных с деятельностью по руководству данными помогает команде программы качества данных налаживать сотрудничество с широким спектром заинтересованных и согласующих сторон, куда входят:

- подразделения ИБ и управления рисками, которые помогают выявлять риски и уязвимости в отношении данных;
- разработчики бизнес-процессов и сотрудники центров обучения персонала, способствующие реализации усовершенствований;

- распорядители и владельцы данных, которые используются в операционных и бизнес-процессах, способные помочь с определением критически важных данных, стандартов качества и требований, приоритетных для разрешения проблем, и т. п.

Команды, входящие в состав организационной системы руководства данными, могут оказать содействие в реализации программы качества данных по следующим вопросам:

- определение и утверждение приоритетов;
- выявление круга лиц, заинтересованных в качестве данных, и обеспечение возможности доступа к ним с целью согласования всех необходимых решений и действий;
- разработка и сопровождение стандартов качества данных;
- ведение отчетности с результатами измерения параметров качества данных в масштабах организации;
- общие рекомендации и наставления по мотивации сотрудников;
- установление средств коммуникации и механизмов обмена опытом;
- разработка политики и правил управления качеством данных, включая меры по обеспечению их соблюдения;
- мониторинг и отчетность;
- распространение/публикация результатов проверок, выявление и согласование возможностей для дальнейшего совершенствования;
- разрешение разногласий и конфликтов интересов; определение целей.

6.1 Политика в области качества данных

Усилия по обеспечению качества данных должны поддерживаться четко сформулированной политикой руководства данными, а сами, в свою очередь, способствовать проведению этой политики в жизнь посредством разработки и реализации правил и процедур в области качества данных. Например, политикой руководства данными может предусматриваться периодический аудит качества данных организации и предписываться соблюдение соответствующих стандартов, правил и рекомендаций. Все области знаний по управлению данными требуют наличия определенной политики, но правила программы качества данных в этом отношении стоят особняком, поскольку имеют обычно важнейшее значение для обеспечения соблюдения организацией требований действующего законодательства и надзорных органов. Политика в области качества данных должна включать следующие разделы.

- Назначение, сфера действия и применения политики качества данных.
- Термины и определения.
- Функции и сфера ответственности программы качества данных.
- Функции и сфера ответственности других заинтересованных сторон.
- Отчетность.

- Реализация политики, включая меры по минимизации риска, профилактике нарушений, надзору, обеспечению ИБ и защиты данных.

6.2 Метрики

Значительная часть работы команды качества данных заключается в измерении, расчете и документировании всевозможных показателей качества. Можно выделить следующие высокоуровневые категории метрик для оценки качества данных.

- **Окупаемость**: суммарные затраты на работы по совершенствованию качества данных в сопоставлении с оцениваемым экономическим эффектом от повышения качества данных.
- **Уровни качества**: показатели уровней наличия некондиционных данных или нарушений требований в различных наборах данных.
- **Тенденции изменения качества данных**:
  - динамика повышения качества данных по отчетным периодам в сравнении с целевыми показателями или допусками;
  - статистика нарушений или инцидентов с качеством данных по отчетным периодам.
- **Метрики для оценки управления проблемными вопросами**:
  - статистика числа выявленных проблемных вопросов по категориям данных / параметрам качества;
  - статистика проблемных вопросов с качеством данных по бизнес-функциям с разбивкой по статусам (решена, решается, изучается, передана на высший уровень);
  - статистика проблемных вопросов по уровням приоритетности/серьезности/тяжести последствий;
  - сроки разрешения проблем.
- **Соблюдение нормативов по уровням обслуживания**: подразделения и ответственные сотрудники, вовлеченные в деятельность по обеспечению качества данных, выполняемые и реализованные проекты по экспертизе качества данных, соответствие процесса управления качеством данных предъявляемым требованиям (в целом).
- **План реализации программы качества данных**: текущее состояние и дорожная карта мероприятий по развитию.

7. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Batini, Carlo, and Monica Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*. Springer, 2006. Print.

Brackett, Michael H. *Data Resource Quality: Turning Bad Habits into Good Practices*. Addison-Wesley, 2000. Print.

Deming, W. Edwards. *Out of the Crisis*. The MIT Press, 2000. Print.

English, Larry. *Improving Data Warehouse and Business Information Quality: Methods For Reducing Costs And Increasing Profits*. John Wiley and Sons, 1999. Print.

English, Larry. *Information Quality Applied: Best Practices for Improving Business Information, Processes, and Systems*. Wiley Publishing, 2009. Print.

Evans, Nina and Price, James. «Barriers to the Effective Deployment of Information Assets: An Executive Management Perspective». *Interdisciplinary Journal of Information, Knowledge, and Management*. Volume 7, 2012. Accessed from http://bit.ly/2sVwvG4

Fisher, Craig, Eitel Lauría, Shobha Chengalur-Smith and Richard Wang. *Introduction to Information Quality*. M. I. T. Information Quality Program Publications, 2006. Print. Advances in Information Quality Book Ser.

Gottesdiener, Ellen. *Requirements by Collaboration: Workshops for Defining Needs*. Addison-Wesley Professional, 2002. Print.

Hass, Kathleen B. and Rosemary Hossenlopp. *Unearthing Business Requirements: Elicitation Tools and Techniques*. Management Concepts, Inc, 2007. Print. Business Analysis Essential Library.

Huang, Kuan-Tsae, Yang W. Lee and Richard Y. Wang. *Quality Information and Knowledge*. Prentice Hall, 1999. Print.

Jugulum, Rajesh. *Competing with High Quality Data*. Wiley, 2014. Print.

Lee, Yang W., Leo L. Pipino, James D. Funk and Richard Y. Wang. *Journey to Data Quality*. The MIT Press, 2006. Print.

Loshin, David. *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann, 2001. Print.

Loshin, David. *Master Data Management*. Morgan Kaufmann, 2009. Print.

Maydanchik, Arkady. *Data Quality Assessment*. Technics Publications, LLC, 2007 Print.

McCallum, Ethan. *Bad Data Handbook: Cleaning Up the Data So You Can Get Back to Work*. 1st Edition. O'Reilly, 2012.

McGilvray, Danette. *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. Morgan Kaufmann, 2008. Print.

Myers, Dan. «The Value of Using the Dimensions of Data Quality», *Information Management*, August 2013, http://bit.ly/2tsMYiA

Olson, Jack E. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, 2003. Print.

Redman, Thomas. *Data Quality: The Field Guide*. Digital Press, 2001. Print.

Robertson, Suzanne and James Robertson. *Mastering the Requirements Process: Getting Requirements Right*. 3rd ed. Addison-Wesley Professional, 2012. Print.

Sebastian-Coleman, Laura. *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*. Morgan Kaufmann, 2013. Print. The Morgan Kaufmann Series on Business Intelligence.

Tavares, Rossano. Qualidade de Dados em Gerenciamento de Clientes (CRM) e Tecnologia da Informação [Data Quality in Management of Customers and Information Technology]. São Paulo: Catálise, 2006. Print.

Witt, Graham. *Writing Effective Business Rules: A Practical Method*. Morgan Kaufmann, 2012. Print.

Большие данные и наука о данных

1. ВВЕДЕНИЕ

После 2000 года термины *большие данные* (*Big Data*) и *наука о данных* (*Data Science*) стали употребляться даже слишком часто. При этом понимание смысла стоящих за ними понятий во многом утеряно, — по крайней мере, круг устоявшихся определений, относительно которых выработан консенсус, крайне ограничен. Даже само определение «большие» трактуется весьма относительно. Тем не менее за обоими этими понятиями — «большие данные» и «наука о данных» — стоят значительные технологические изменения, благодаря которым человечество имеет возможность генерировать, хранить и анализировать колоссальные объемы данных, и эти объемы продолжают неуклонно расти. Что еще важнее, люди научились использовать такие данные для моделирования, прогнозирования и влияния на поведение, а также получения углубленных представлений о широком спектре важнейших предметов, включая статистику здравоохранения, управления природными ресурсами, экономического развития и т. д.

Термин «большие данные» указывает не только на объем данных, но и на их разнообразие (структурированные и неструктурированные, документы, файлы, аудио- и видеозаписи, потоковые данные и т. д.), а также на скорость, с которой они производятся. Специалистов, которые исследуют данные, строят предиктивные (predictive) и предписывающие (prescriptive) модели, а также модели машинного обучения (machine learning), проводят на их основе анализ и осуществляют внедрение полученных результатов в интересах заинтересованных сторон, стали теперь называть «учеными в области данных» или «учеными по данным» (data scientists).

На самом же деле понятие «наука о данных» используется для обозначения хорошо известной прикладной статистики (applied statistics). Другое дело, что вычислительные мощности, необходимые для выявления статистических закономерностей, сегодня выросли настолько, что способствовали появлению больших данных и реализации технологий их статистико-аналитической обработки. Традиционная бизнес-аналитика (BI) подобна «зеркалу заднего вида» (rear-view mirror), поскольку описывает тенденции, выявленные по результатам изучения структурированных ретроспективных данных. Иногда выявленные закономерности бизнес-аналитики используются и для прогнозирования, но уверенности в надежности таких прогнозов нет и быть не может по определению, поскольку это всего лишь экстраполяции в будущее прошлых тенденций,

которые в любой момент могут измениться. До недавнего времени углубленный анализ колоссальных массивов данных был невозможен по технологическим причинам, и аналитикам приходилось полагаться на ограниченные по размерам статистические выборки или иные средства приблизительной оценки. С ростом вычислительных мощностей ученые научились накапливать и обрабатывать гораздо более объемные массивы данных и применять к ним комплексные методы анализа, позаимствованные из прикладной математики, статистики, информатики, обработки и преобразования сигналов, теории вероятностей, распознавания образов, машинного обучения, моделирования неопределенности, визуализации данных и других прикладных областей знания с целью углубленного изучения и предсказания поведения систем на основе массивов больших данных. Иными словами, наука о данных нашла новые способы анализа данных и извлечения из них ценности.

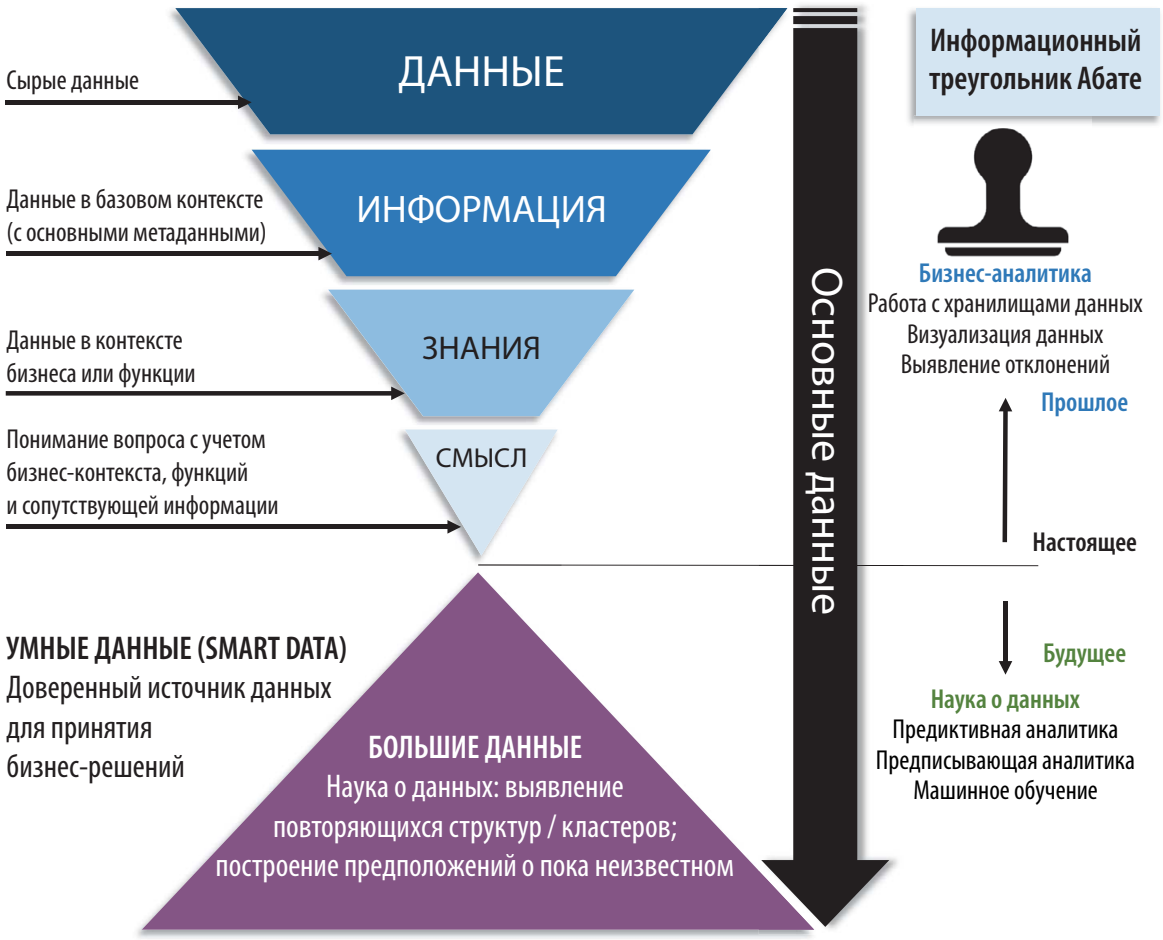


Рисунок 96. Информационный треугольник Абате[1]

[1] Роберт Абате (*англ.-ит.* Robert J. Abate) — американский специалист по архитектурам данных на основе сервисов (SBA) и управлению большими данными, вице-президент нью-йоркской секции DAMA. — *Примеч. пер.*

С привнесением больших данных в среды хранилищ данных и BI (см. главу 11) методы науки о данных стали использоваться для обеспечения возможности смотреть вперед («через лобовое стекло» — windshield). Возможность прогнозирования на основе моделей, в том числе в режиме, близком к реальному времени, с использованием разнородных данных из множества различных источников помогает организациям всё лучше понимать направления своего развития (см. рис. 96).

Однако для использования преимуществ больших данных требуется изменить методы управления данными. Большинство хранилищ данных используют традиционную реляционную модель. Большие данные, как правило, в виде такой модели не представлены. В большинстве хранилищ данных обработка тесно связана с процедурами ETL (извлечение, преобразование, загрузка). В решениях для обработки больших данных (в частности, в так называемых «озерах данных») используется концепция ELT, то есть загрузка и последующее преобразование. Не менее важно и другое: скорость и потоки загрузки в случае сбора больших данных столь велики, что стандартные подходы к критически важным аспектам управления данными — интеграции, управлению метаданными, обеспечению качества данных — становятся неприемлемыми, и возникает необходимость в выработке и реализации принципиально новых решений еще и в этих областях.

1.1 Бизнес-драйверы

Главный драйвер развития в организации работ в области сбора и исследования больших данных — стремление к обнаружению скрытых бизнес-возможностей посредством всесторонней аналитической проработки массивов данных с использованием широкого спектра диверсифицированных алгоритмов. Большие данные побуждают к инновациям, поскольку объемы и разнообразие массивов, доступных для исследования, растут безостановочно, и все эти данные можно использовать для определения моделей прогнозирования нужд потребителей и создания персонализированных презентаций продуктов и услуг. Наука о данных способствует повышению производительности и результативности обработки больших данных. Алгоритмы машинного обучения помогают автоматизировать сложные по структуре и ресурсоемкие комплексы рабочих процессов, способствуя повышению эффективности работы организации, снижая затраты и минимизируя риски.

1.2 Принципы

Большие данные сулят заманчивую перспективу глубокого осмысления реальности под новыми и неожиданными углами зрения, но для этого нужно для начала уметь ими управлять. Из-за большого разнообразия источников и форматов управление большими данными дается на порядок сложнее и требует значительно большей дисциплины по сравнению с управлением реляционными базами данных. Принципы управления большими данными до конца не сформировывались, но главный принцип на сегодняшний день сформулирован предельно четко: управление большими данными требует тщательного управления метаданными, описывающими источники больших данных, чтобы можно было обеспечить полный учет файлов данных, их происхождения, контента и ценности.

Рисунок 97. Контекстная диаграмма: большие данные и наука о данных

1.3 Основные понятия и концепции

1.3.1 Наука о данных

Как уже отмечалось во введении, наука о данных объединяет статистический анализ, машинное обучение, интеграцию и моделирование данных для построения прогнозных моделей и выявления структурных закономерностей в содержании данных.

Иногда науку о данных трактуют более узко, относя к ней только предиктивное моделирование, что не лишено оснований в том смысле, что именно на стадиях моделирования и прогнозирования аналитики больших данных придерживаются естественно-научной методологии в строгом понимании.

Аналитик данных выдвигает гипотезу о возможном наблюдаемом поведении предметов статистического описания еще до начала каких-либо действий. Например, часто бывает, что покупка предмета потребления одной категории с высокой вероятностью влечет за собой покупку предмета потребления другой (пример: покупка жилья влечет за собой покупку мебели). Затем аналитик исследует большие объемы исторических данных с целью проверки справедливости этой гипотезы и определения статистической корреляции между двумя параметрами модели. Если гипотеза подтверждается, а корреляция (показатель обусловленности второго события первым) достаточно высока, модель может стать основной для практического применения в целях прогнозирования поведения или даже использования ее в режиме реального времени — например, для контекстной рекламы.

Разработка решений в науке о данных ведется методом итеративного подключения к модели всё новых и новых источников данных по мере наработки статистически значимых результатов с целью углубления и детализации полученных выводов. Эффективность практического применения методологии науки о данных зависит от следующих факторов.

- **Богатство исходных данных** как признак потенциала выявления в них скрытых закономерностей и тенденций в поведении организаций или потребителей.
- **Сопоставление и анализ информации**: технические приемы, используемые для понимания смыслового наполнения данных и правильного сочетания их наборов с целью выдвижения и проверки гипотез о взаимосвязях и закономерностях.
- **Извлечение и выдача информации**: обработка массивов данных с применением математических моделей и алгоритмов и создание визуальных и иных представлений выходных данных, позволяющих выявлять глубинные закономерности и характеристики поведения.
- **Оформление результатов анализа данных** с целью их распространения.

Таблица 32 сравнивает роль традиционной модели хранилища данных / бизнес-аналитики с моделями прогнозной и предписывающей аналитики, которые можно реализовать в рамках методологии науки о данных.

Таблица 32. Прогресс аналитики

Традиционные средства DW/BI	Наука о данных	
Описание	Предварительное прогнозирование	Предписание
Осмысление прошлого	Понимание настоящего	Предвидение будущего
Анализ истории: что произошло; как и почему это случилось?	Модели прогнозирования: что и с какой вероятностью произойдет?	Сценарный анализ: какая последовательность действий даст желаемые результаты?

1.3.2 Процесс осуществления деятельности в области науки о данных

Рисунок 98 иллюстрирует последовательность итераций в рамках процесса осуществления деятельности в области науки о данных (Data Science process). Результаты на выходе предыдущего этапа этого циклического процесса служат исходными данными для следующего этапа (см. раздел 2).

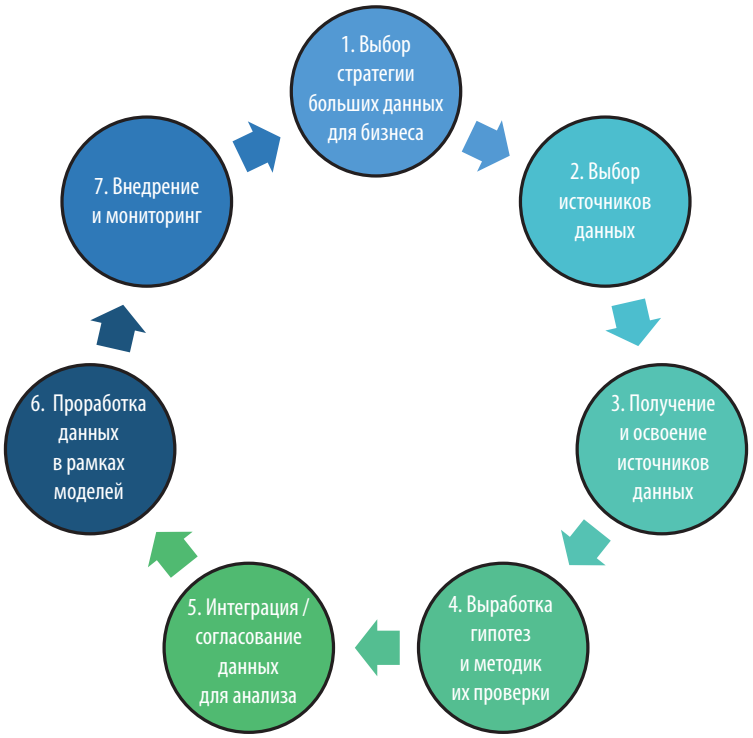


Рисунок 98. Процесс осуществления деятельности в области науки о данных

Наука о данных следует общепринятой методологии познания посредством последовательного приближения к объективной истине через циклы наблюдений, выдвижения и опытной проверки гипотез, накопление результатов экспериментов в рамках предложенной модели и формулирование общих теорий, объясняющих совокупность результатов наблюдений и экспериментов.

В науке о данных этот процесс познания принимает форму наблюдений за данными, создания и оценки годности моделей, объясняющих их поведение.

- **Определение стратегии и потребностей бизнеса в области изучения больших данных.** Сформулируйте требования к желаемым результатам с указанием измеримых материальных выгод от их выполнения.
- **Выбор источников данных.** Идентифицируйте пробелы в имеющейся базе информационных ресурсов и изыщите источники данных, которые позволят заполнить эти пробелы.
- **Получение и освоение источников данных.** Получите все необходимые наборы данных или доступ к их источникам с целью загрузки.
- **Проработка гипотез и методов их проверки средствами науки о данных.** Исследуйте источники данных с помощью средств профилирования, визуализации, статистического анализа и т. п. с целью уточнения требований. Определите алгоритм модели и необходимые типы вводных и выходных данных или смоделируйте несколько альтернативных гипотез и методов анализа (например, сравнительный анализ группировок данных, выявленных посредством кластеризации, и т. п.).
- **Интеграция и согласование данных для анализа.** Годность модели зависит еще и от качества источников данных. Используйте данные из надежных и достоверных источников. По мере необходимости используйте средства интеграции, очистки и доработки данных с целью повышения качества и полезности вводимых наборов.
- **Исследование данных с использованием моделей.** Задействуйте средства статистического анализа и алгоритмы машинного обучения для выявления закономерностей на основе интегрированных данных. Регулярно проверяйте валидность модели и при необходимости вносите коррективы в параметры модели и настройки алгоритмов самообучения, а по мере накопления статистики дорабатывайте и саму модель. Машинное обучение подразумевает многократные прогоны через модель больших массивов реальных данных с целью проверки гипотез и внесения корректив в настройки алгоритмов (например, выявления выпадающих из общего статистического ряда значений). В процессе такой проработки окончательно уточняются и требования. Эволюция модели выверяется по изначально определенным метрикам пригодности/реалистичности результатов. С появлением новых гипотез могут потребоваться дополнительные наборы данных, а по результатам их проверки — новые модели, выходные данные и даже требования.
- **Внедрение и мониторинг.** Модели, которые выдают полезную информацию, можно переносить в производственную среду и использовать для текущего мониторинга ситуации с целью получения данных или, напротив, появления нежелательных тенденций, ставящих под угрозу эффективность текущей бизнес-модели. На этой стадии проекты по изучению данных превращаются в обычные рабочие проекты хранилища данных / бизнес-анализа и в среде хранилища обрастают всеми необходимыми техническими доработками и компонентами (процедурами ETL, DQ, основными данными и т. д.).

1.3.3 Большие данные

На ранней стадии формирования этого понятия большие данные определялись по признаку соответствия трем «V-характеристикам»: Volume — объем, Velocity — скорость, Variety — разнообразие (Laney, 2001). Вместе с широким распространением этой концепции в организациях, стремящихся сполна реализовать потенциал колоссальных массивов слабо структурированной информации, число V-характеристик в мнемоническом правиле определения понятия больших данных удвоилось. В наши дни к ним относят данные со следующими характерными свойствами.

- **Volume — объем** как мера количества данных: большие данные включают миллиарды полей или записей, описывающих тысячи сущностей или элементов.
- **Velocity — скорость** регистрации/генерирования, обработки или распространения: *большие данные* зачастую не только создаются, но и распространяются и даже анализируются в режиме реального времени или близком к нему.
- **Variety/Variability — разнообразие/вариативность** формы или представления: большие данные сохраняются во всевозможных форматах, а их структура зачастую бывает несогласованной не только между наборами, но и внутри отдельно взятых наборов данных.
- **Viscosity — вязкость**: большие данные крайне трудно поддаются как вычленению из общей массы, так и анализу и интеграции с целью практического использования.
- **Volatility — волатильность** как мера непостоянства: большие данные крайне переменчивы, что весьма ограничивает сроки годности полученных с их использованием результатов.
- **Veracity — правдоподобие** по критериям проверки подлинности источника.

Но главной отличительной особенностью больших данных являются колоссальные объемы занимаемой ими памяти: сегодня под большими данными по умолчанию понимают нечто свыше 100 терабайт, а то и петабайты или эксабайты данных. В обычных средах с централизованной архитектурой DW/BI обработка подобных объемов становится весьма проблематичной, поскольку требует ЦОДа с серьезными серверными мощностями и пропускной способностью каналов связи для их загрузки, моделирования, очистки и анализа. Проблему часто решают за счет массивно-параллельной архитектуры обработки данных или сочетания параллельной обработки с распределенными вычислениями и облачными хранилищами. Однако всё это не более чем локальные и временные решения, поскольку проблемы, обусловленные нарастанием объемов и потоков больших данных, имеют гораздо более широкие и далеко идущие последствия. Колоссальные размеры наборов данных требуют от нас изменений в общем подходе к хранению данных, доступу к данным, в концептуальном представлении о данных (в частности, отказа от традиционного мышления категориями структур данных, описываемых реляционными моделями), а также в методах управления данными (Adams, 2009).

Рисунок 99 позволяет составить наглядное представление о расширении спектра данных, которые стали доступны благодаря технологиям сбора и анализа больших данных, и о последствиях этого информационного разнообразия с точки зрения емкости требуемых хранилищ данных.

Рисунок 99. Масштабы задач в области хранения данных[1]

1.3.4 Компоненты архитектуры больших данных

Для правильного выбора, установки и конфигурации средств сбора и анализа *больших данных* требуются опытные специалисты. Необходимо разработать дополнительный комплекс архитектурных решений, его согласование с существующими средствами сбора и анализа данных и обоснование необходимости новых приобретений.

Рисунок 100 описывает концептуальную архитектуру рабочей среды для областей DW/BI и больших данных (о DW/BI подробнее в главе 11). Ключевое различие между средами работы с большими данных и традиционного хранилища заключается в порядке операций: в среде DW/BI реализуется последовательность ETL (извлечение → преобразование → загрузка), а в среде больших данных — алгоритм ELT (извлечение → загрузка → преобразование). Это важнейший момент, поскольку большие данные загружаются до их приведения к совместимому с имеющейся структурой данных виду, что необходимо для интеграции. Во многих случаях интеграции в традиционном смысле приведения к общей модели большим данным и не требуется. Вместо подготовки их к использованию в составе общего комплекса интегрированных данных применяется метод выборочного включения этих данных в процессы, для которых они могут быть полезными (например, в процессе построения модели прогнозирования могут потребоваться какие-то конкретные наборы данных, — только они и будут интегрированы).

[1] *Источник*: Robert Abate / EMC Corporation. Используется с разрешения правообладателей.

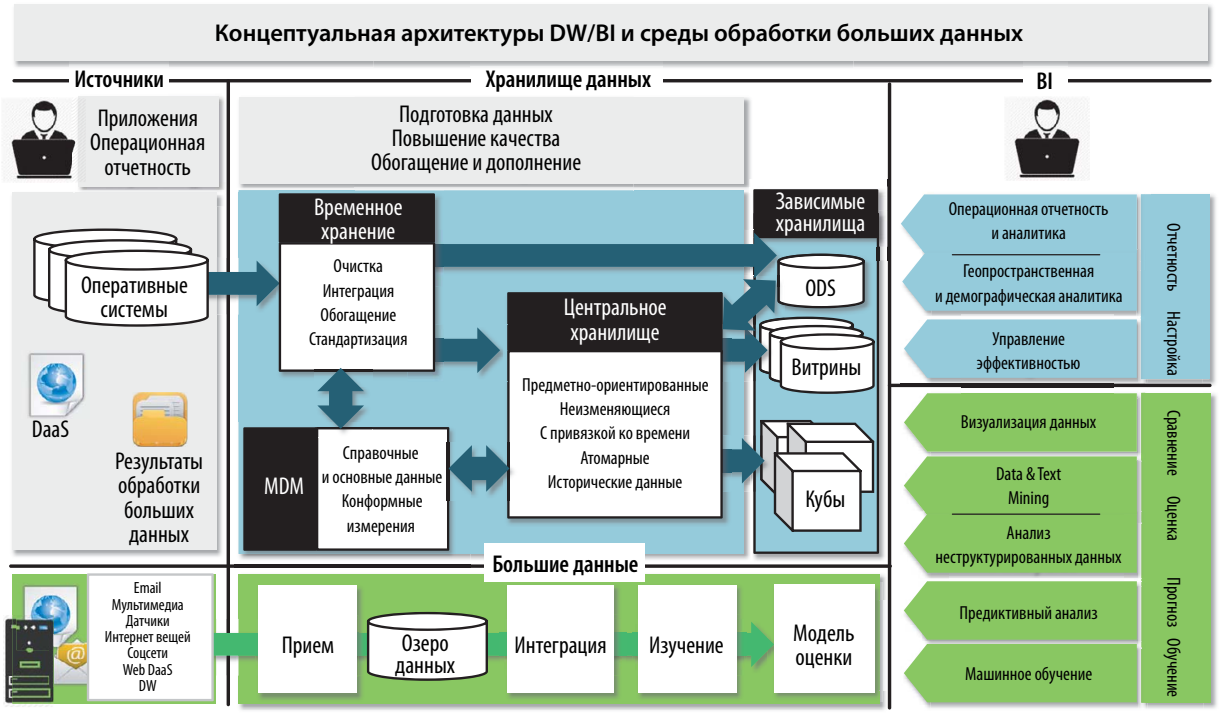


Рисунок 100. Концепция рабочей среды для областей DW/BI и больших данных

Различия между обработкой данных по схемам ETL и ELT столь значительны, что влияют и на систему управления данными. Например, схема ELT позволяет обойтись вовсе без моделирования данных предприятия. Однако подобное упрощение рискует обернуться утерей большей части информативного содержания обрабатываемых данных в процессе беспорядочного освоения. Отсюда вытекает еще и обязательность управляемого сбора метаданных о накапливаемых данных, чтобы со временем не утрачивалось понимание их смысла и назначения.

Далее в настоящем разделе описываются источники больших данных, структура озер данных (хранилище большого объема неструктурированных данных) и средства их реализации, а в разделе 2 — основные направления работ по освоению, интеграции, изучению и оценке результатов анализа больших данных.

1.3.5 Источники больших данных

Значительная часть человеческой деятельности в современном мире осуществляется в электронной форме, а значит, ежедневно накапливаются огромные массивы дополнительной информации, появление которой обусловлено нашими передвижениями по миру, взаимодействиями друг с другом и всевозможными бизнес-транзакциями. Мы непрестанно генерируем большие данные, отправляя электронные письма, высказываясь в соцмедиа, оформляя онлайн-заказы и даже просто играя в сетевые видеоигры. Данные генерируются не только компьютерами и смартфонами, но

и кассовыми терминалами, системами видеонаблюдения, сенсорными датчиками транспортных систем, системами медицинского наблюдения, промышленными, коммунальными, спутниковыми системами, не говоря уже о военной технике. Например, за один регулярный рейс современный гражданский пассажирский самолет генерирует до терабайта данных. Значительную долю больших данных создают всевозможные устройства с интернет-подключением, обменивающиеся информацией с владельцами или между собой, — этот феномен иногда называют интернетом вещей (Internet of Things, IoT).

1.3.6 Озёра данных

Озеро данных — среда накопления массы разнородных по типу и структуре данных, откуда они могут черпаться для сохранения, оценки или анализа. Озёра данных могут создаваться в различных целях. Вот лишь некоторые примеры их функционального назначения:

- среда для работы специалистов по анализу данных методами *науки о данных*;
- центральное хранилище — накопитель сырых данных, иногда с функцией минимальной предварительной обработки;
- альтернативное хранилище детальных архивных версий DW/BI;
- онлайн-архив записей;
- среда для обработки входящих потоковых данных с функцией автоматического распознавания структуры.

Озеро данных может быть реализовано в сложной конфигурации с использованием продвинутых средств обработки данных, включая системы управления хранилищами (например, Hadoop), службы кластеризации, преобразования и интеграции данных. Все подобные обработчики озерных данных специально ориентированы на работу на базе распределенной инфраструктуры хранения данных и имеют аналитическую оснастку, позволяющую собирать данные согласно заданной структурной конфигурации.

Главный риск при использовании озера данных заключается в том, что оно имеет тенденцию к быстрому превращению в болото — грязное, запущенное, вязкое и непрозрачное. Чтобы этого не допустить, нужен учет содержания наполнения озера, а для учета содержания — непрерывная маркировка вводимых данных метаданными прямо на входе, что делает управление метаданными важнейшей задачей сопровождения озера данных. Для того чтобы понять характер связей — хотя бы ассоциативных — между данными в озере, архитекторы и проектировщики часто используют уникальные ключи или иные технические приемы (например, семантические или топонимические модели данных), чтобы аналитики и иные разработчики средств визуализации данных имели хотя бы приблизительное представление о том, что за информация стекается в озеро данных и как ее можно использовать (см. главу 9).

1.3.7 Архитектура на основе сервисов

В последнее время набирает популярность архитектура на основе сервисов (SBA[1]), позволяющая поначалу немедленно выдавать потребителям данные без гарантии их точности или полноты, а параллельно вести доработку данных из того же источника, чтобы затем сохранить их в полном и точном историческом наборе (Abate, Aiken, Burke, 1997). Архитектура SBA представляет собой вариант архитектуры DW с немедленной отправкой данных, поступающих из операционных систем, в хранилище операционных данных (ODS) и одновременной доработкой этих данных в области подготовки с последующей отправкой данных в главное DW, где накапливается история. Архитектура SBA предусматривает наличие трех слоев данных — пакетного, скоростного и слоя выдачи (см. рис. 101).

- **Пакетный слой** реализован в среде озера данных, где ведется полная обработка поступающих данных и хранятся как последние, так и исторические данные.
- **Скоростной слой** содержит только текущие данные, поступающие в режиме реального времени.
- **Слой выдачи** — интерфейс представления сводных данных (текущих и полных).

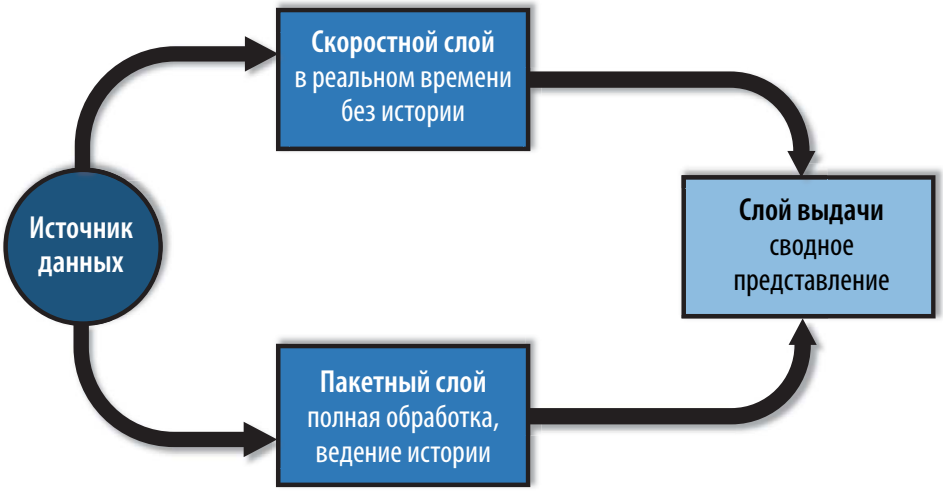


Рисунок 101. Архитектура на основе сервисов (SBA)

Данные загружаются одновременно в пакетный и скоростной слои. Все аналитические вычисления производятся и в скоростном, и в пакетном слоях, что требует, как правило, двух раздельных систем обработки. Решение проблем синхронизации путем подбора оптимального компромисса между полнотой, задержкой, точностью и детализацией сводного представления осуществляется через определение параметров слоя выдачи. Для определения требуемого баланса между, например,

[1] *сокр.* от *англ.* services-based architecture. — *Примеч. пер.*

временем запаздывания и точностью или полнотой отображаемых данных, а также стоимостью и сложностью решения, как правило, используется сравнительная оценка издержек/выгод.

Пакетный слой часто называют также накопительным по времени компонентом, поскольку транзакции туда последовательно дописываются, а скоростной слой — хранилищем операционных данных (ODS), поскольку там представлены лишь последние транзакции (или, если требуется, только правки, дельты, приращения и т. п.). Ресурсоемкость подобной архитектуры оправдывает себя там, где требуется исключить всякую возможность рассинхронизации текущего представления с данными в источнике в текущем слое за счет вынесения строгой обработки данных в исторический слой. Слой выдачи данных или служб данных при такой архитектуре извлекает и сводит данные из обоих слоев, используя *метаданные*. Сервисы данных определяют согласно заданным правилам, из какого слоя какие данные брать для «выдачи» в ответ на те или иные запросы потребителей данных.

1.3.8 Машинное обучение

Машинное обучение исследует методы построения алгоритмов, реализованных в программном обеспечении. Можно рассматривать машинное обучение как синтез методов неконтролируемого самообучения (часто называемых «извлечением информации» — data mining) и методов контролируемого или управляемого обучения, которые имеют глубокие математические корни, в том числе из статистики, комбинаторики и оптимизации систем. Начала формироваться и третья ветвь — так называемое обучение с подкреплением без учителя: задаются целевые параметры, и система упражняется в их соблюдении (пример: автопилот транспортного средства). Программирование машин на быстрое усвоение повторяющихся структур запросов и адаптацию к изменениям наборов данных привело к появлению одноименного раздела «машинное обучение» и в области больших данных, где эта концепция получила совершенно новое применение[1]. Процессы прогоняются, результаты сохраняются, а затем используются при последующих прогонах для уточненной настройки процесса, и такие итерации повторяются до получения результата желаемого уровня точности и детализации.

Машинное обучение занимается структурным построением алгоритмов познания и усвоения знаний. Выделяют три типа таких алгоритмов.

- **Обучение с учителем** основано на применении обобщенных правил (пример: настраиваемый фильтр спама в почтовом приложении).
- **Обучение без учителя** основано на выявлении скрытых паттернов, связей, закономерностей (то есть собственно интеллектуальный анализ данных).
- **Обучение с подкреплением** основано на достижении цели (например, выигрыша шахматной партии).

[1] См., например, «Периодическую таблицу методов визуализации» (http://bit.ly/IX1bvI) — интерактивный путеводитель по различным платформам, доступным разработчикам, теоретикам и практикам обучения машин распознаванию данных.

Статистическое моделирование и машинное обучение используют также для автоматизации нереализуемых или слишком затратных процессов в рамках исследовательских и проектных работ, когда требуется, например, методом проб и ошибок подобрать ключ к огромному набору данных, повторяя цикл экспериментальной обработки, анализа результатов и исправления ошибок. Такой подход позволяет значительно ускорить получение ответа, что и стимулирует организации к инициативам по поиску глубинных закономерностей посредством многократного повторения затратно эффективных процессов. Например, CIVDDD[1] использует машинное обучения и комплексные средства визуализации научных данных с целью оказания помощи государственным органам и миротворческим силам в противостоянии принявшим массовый характер информационным угрозам.

Хотя машинное обучение и использует весьма новые способы получения данных, в этой новой области знания должны соблюдаться все традиционные принципы этичного обращения с данными, в частности и прежде всего — принцип прозрачности. Появились научные доказательства того, что метод обучения нейронных сетей методом глубокого погружения работает. Они учатся и постигают мир. Однако не всегда ясны механизмы их обучаемости. Чем сложнее становятся алгоритмы, лежащие в основе этих процессов, тем менее они прозрачны — и начинают функционировать в режиме «черного ящика». Чем больше переменных учитывают самообучаемые нейронные сети и чем более абстрактными делаются сами эти переменные, тем больше реализуемые ими алгоритмы испытывают пределы возможностей человека понимать и интерпретировать логику машинного мышления (Davenport, 2017). Необходимость обеспечения прозрачности принятия решений по мере дальнейшего совершенствования функциональности неконтролируемого самообучения и его применения во всё более широком спектре ситуаций, вероятно, будет только возрастать (см. главу 2).

1.3.9 Анализ настроений

Мониторинг медиа и анализ текста относятся к автоматизированным методам извлечения аналитической информации из больших массивов неструктурированных и слабо структурированных данных, включая страницы отзывов, соцмедиа, блоги, новостные веб-сайты и т. п. Делается это для того, чтобы понять и обобщить мнения людей и выявить преобладающее в различных социальных группах отношение к брендам, продуктам или услугам, а также любым другим темам или явлениям. Используя алгоритмы обработки естественного языка, синтаксического и лексического разбора предложений или формулировок, средства семантического анализа позволяют выявлять не только доминирующую в высказываниях эмоциональную окраску, но и динамику ее изменения во времени, что открывает возможность предсказывать вероятные сценарии дальнейшего развития событий.

[1] CIVDDD (*сокр.* от *англ.* the Centre for Innovation in Information and Data-Driven Design — «Центр инноваций в информационном проектировании») — субсидируемая межуниверситетская программа по изысканию возможностей для использования средств анализа и визуализации *больших данных* в прикладных информационно-технологических решениях нового поколения, включая новые вычислительные средства, стратегии и интерфейсы представления данных.

Проиллюстрируем этот подход на простейшем примере поиска и подсчета статистики частоты употребления ключевых слов в опубликованных отзывах о продукте. Если в комментарии присутствуют слова «*отличный*», «*восторг*», «*замечательно*» и т. п., вероятно, это позитивный отклик, а присутствие слов «*плохой*», «*дрянь*», «*гадость*» может служить признаком негативного отношения. Распределив отзывы по категориям, можно выяснить преобладающее в целевом сообществе (например, в данной соцсети, блоге и т. п.) отношение. Но, к слову, реальные чувства и эмоции, вызываемые предметом обсуждения, не так легко бывает уловить по причине того, что любое ключевое слово, будучи вырванным из контекста, может быть интерпретировано неверно. Например, слово «*ужасно*» вроде бы указывает на негативное отношение к ресторану, а в отзыве написано: «*Ужасно* вкусно!» А формально позитивную характеристику «*сказочно*» можно найти в возмущенном отзыве: «*Сказочно* нерасторопное обслуживание!» Поэтому семантический анализ эмоциональной окраски должен интерпретировать слова только в контексте. А это уже требует понимания смыслового значения всего отзыва или комментария. Для правильной интерпретации смысла написанного часто требуются функции обработки естественного языка, реализованные на сегодняшний день лишь в суперсистемах уровня IBM Watson.

1.3.10 Интеллектуальный анализ данных и текстов

Интеллектуальным анализом данных (или извлечением информации — data mining) принято называть применение к массивам разнородных данных разнообразных алгоритмов выявления скрытых структурных закономерностей. Интеллектуальный анализ данных постепенно отделился от машинного обучения и сделался отдельной подобластью исследований по созданию искусственного интеллекта. Теория интеллектуального анализа данных формально относится к методологии статистического анализа, известной под названием «обучение без учителя», которая предусматривает применение к набору данных неких алгоритмов изучения, никак не связанных с ожидаемым или желаемым результатом. В то время как стандартные средства генерации запросов и отчетов формулируют вполне конкретные требования к данным, средства интеллектуального анализа данных помогают раскрывать неизвестные ранее взаимосвязи через выявление повторяющихся структур (паттернов). Извлечение данных — ключевое направление работ на этапе первичного изыскания возможностей, поскольку позволяет оперативно идентифицировать поддающиеся изучению элементы обрабатываемого массива данных, выявлять ранее неизвестные и уточнять нечеткие или неклассифицированные связи, закладывая структурную основу классификации элементов изучаемых данных.

В сочетании с семантическим и структурно-лингвистическим анализом текстовой информации интеллектуальный анализ данных позволяет автоматически классифицировать данные по признакам их содержания и интегрировать полученные классификации в онтологии, составляемые по мере накопления данных под общим руководством экспертов в предметной области. Таким образом, появляется возможность анализа электронных текстов в различных средах и форматах без их реструктурирования или конвертирования. Накапливаемые онтологии можно подключать к информационно-поисковым системам, что даст пользователям

и приложениям возможность получать доступ к этим документам через поисковые запросы (см. главу 9).

Извлечение данных и интеллектуальный анализ текстов основаны на использовании ряда стандартных технических приемов, включая описанные ниже.

- **Профилирование** заключается в описании характерных типов поведения людей, групп или организаций и используется для определения признаков нормального поведения с целью выявления серьезных отклонений от нормы — например, в приложениях по отслеживанию мошеннических операций или попыток проникновения в системы. Результаты профилирования служат входными данными для многих компонентов, работающих по принципу самообучения.
- **Сокращение избыточных данных** позволяет заменять исходные, излишне детализированные наборы данных обобщенными, где сохраняются лишь ключевые характеристики или категории, что заметно ускоряет и упрощает обработку и анализ.
- **Ассоциирование** часто встречающихся в связке друг с другом элементов — еще один стандартный алгоритм выявления взаимосвязей, применяемый в интеллектуальном анализе данных. Ассоциативные связи могут использоваться, например, для накопления статистики часто встречающихся наборов элементов, выявления скрытых правил, анализа конъюнктуры локальных рынков. Ну и рекомендательные системы в интернете без использования подобных алгоритмов, понятно, не обходятся.
- **Кластеризация**: группировка элементов в кластеры по признаку близкого сродства или общности неких характеристик упрощает и ускоряет статистический анализ типичных схем и стереотипов поведения. Классический пример кластеризации — сегментация потребительского рынка.
- **Самоорганизующиеся карты** — метод кластерного анализа нейронных сетей, известный также под названием *самоорганизующихся карт Кохонена*[1] или топологически упорядоченных карт. Их использование позволяет снизить размерность пространства оценки без ущерба для результатов аппроксимации. Устранение избыточных пространственных измерений, отметим, по эффективности не уступает изъятию вырожденных переменных из алгебраических уравнений — и решать проще, и результат нагляднее.

1.3.11 Предиктивная аналитика

Предиктивной аналитикой называют подраздел обучения с учителем, в рамках которого пользователи пытаются смоделировать элементы данных и предсказать будущие исходы по оцениваемым вероятностям событий. В методах теории вероятностей и математической статистики прогнозная аналитика, однако, имеет много общего с обучением без учителя в части прописывания, например, предельно допустимых отклонений полученных результатов от предполагаемых, после чего требуется пересмотр гипотез.

[1] Те́уво Ка́леви Ко́хонен (*фин.* Teuvo Kalevi Kohonen, p. 1934) — финский теоретик искусственных нейронных сетей и алгоритмов машинного обучения. Самоорганизующиеся карты — частный случай векторного квантования сети нейронов в так называемом слое Кохонена, где закрепляются алгоритмы, приводящие к успеху. — *Примеч. пер.*

Таким образом, предиктивная аналитика основана на использовании обычных вероятностных (стохастических) моделей обработки вводных данных (включая исторические) для определения вероятности будущих событий (покупок, ценовых изменений и т. п.). При получении информации, выходящей за рамки текущей модели, сама же модель и запрашивает у организации порядок дальнейших действий. Фактором запуска может служить любое событие: заказ в интернет-магазине, текст в новостной ленте, образ в системе распознавания лиц, непредвиденный всплеск спроса на услуги. Пусковым моментом могут являться и внешние факторы. Например, появление негативных материалов о компании в СМИ — верный признак скорого снижения биржевых котировок ее акций. А способность прогнозировать динамику биржевых котировок по новостям — отличное функциональное свойство средств аналитики данных с точки зрения игроков на фондовых рынках.

Зачастую превышение критического порога потока каких-либо характерных данных в режиме реального времени (например, биржевых сделок или обращений в экстренную службу) служит причиной для запуска цепи всевозможных последствий в динамично меняющейся и нестабильной среде. Мониторинг потока событийных данных позволяет устанавливать пороги счетчиков критических событий, определяемых в рамках модели и служащих сигналом для выдачи предупреждения или запуска каких-либо действий.

Запас времени, которое остается в распоряжении у получателей сигнала о прогнозируемом событии до фактического наступления этого события, нередко бывает мизерным (вплоть до долей секунды). Поэтому инвестиции в технологии быстрого реагирования (в частности, резидентные базы данных, широкополосные каналы связи и даже физический перенос ЦОД в непосредственную близость к объекту — источнику данных) оправдываются, если позволяют реально повысить способность к прогнозированию и оперативному реагированию на прогноз.

Простейшая модель прогнозирования — статистическая. Существует множество методик статистического прогнозирования, основанных на выявлении тенденций с экстраполяцией, регрессионном анализе и т. п., но в любом случае требуется сглаживание. Простейший вариант сглаживания данных реализуется путем расчета скользящего среднего или средневзвешенного значения. В специфических случаях могут применяться более сложные техники сглаживания, такие как расчет экспоненциального скользящего среднего, что позволяет управлять коэффициентом сглаживания (фильтрации флуктуаций). Для начала можно применить один из методов регрессионного анализа — метод наименьших квадратов, но в любом случае требуется несколько пробных прогонов для подбора оптимального коэффициента сглаживания. Существуют модели с двумя и более фильтрами экспоненциального сглаживания, позволяющие учитывать, например, недельные колебания на фоне сезонных.

1.3.12 Предписывающая аналитика

Предписывающим анализом называют прогнозный анализ, дополненный определениями корректирующих воздействий на ситуацию с целью изменения конечных результатов, а не ограничивающийся простым их прогнозированием. Таким образом, предписывающая аналитика позволяет

предсказывать, что случится, когда это случится и — главное — по совокупности каких факторов это случится. Будучи способным демонстрировать последствия различных сочетаний решений, предписывающий анализ позволяет моделировать их комбинации с целью максимизации выигрыша или минимизации риска. Методы предписывающего анализа удобны тем, что предусматривают возможность непрерывной подачи на вход скорректированных вводных и перерасчета прогнозов с выдачей скорректированных предписаний. Это повышает и точность прогноза, и результативность предписаний.

1.3.13 Методы анализа неструктурированных данных

Анализ неструктурированных данных основан на сочетании различных методов анализа текстов, ассоциаций, кластеров и прочих вышеописанных методов обучения без учителя, помогающих кодифицировать большие наборы слабо структурированных данных. Могут использоваться и методы обучения с учителем: например, для того чтобы задать направление, ориентацию и наставления машинному мышлению на правильный подход к кодированию выявляемых структурных зависимостей, — и часто лишь человеческое вмешательство позволяет избежать невнятности формулировок или разрешить неоднозначности.

Значение анализа неструктурированных данных возрастает пропорционально нарастанию их доли в мировом информационном пространстве. Бывает, что анализ какого-либо явления просто невозможен без включения в аналитическую модель неструктурированных данных. Однако анализ неструктурированных данных осложняется необходимостью предварительного отделения интересующих исследователей данных от лишних элементов.

Сканирование и тегирование — единственный способ «выуживания» полезных неструктурированных данных из озера, позволяющий отфильтровать их от «воды» и привязать к структурированным данным. Тем не менее тут возникает следующая проблема: какими тегами маркировать данные, не зная заранее их содержания, и как определить условия тегирования? Ответ может быть получен только итерационным путем: по мере выявления реальных условий тегирования уточняются и начинают присваиваться теги, а по мере поглощения и освоения тегированных данных аналитики проверяют правильность условий тегирования, анализируют выловленные данные — и постепенно уточняются и согласуются все условия тегирования и структура тегов, а по мере надобности могут добавляться и новые теги.

1.3.14 Операционная аналитика

Концепция операционной аналитики (она же операционная BI, бизнес-аналитика, потоковая аналитика данных и т. п.) появилась в результате интеграции в операционную деятельность функций анализа данных в режиме реального времени. Средства операционного анализа включают сегментацию пользователей, анализ эмоциональной окраски, геокодирование и другие приемы потоковой обработки данных в целях анализа эффективности маркетинговых кампаний, охвата рынков, популярности продуктов, оптимизации ресурсов, управления рисками и т. д. и т. п.

Операционная аналитика предусматривает встраивание средств слежения в потоки оперативной информации в режиме реального времени, обработку сигналов алгоритмами моделей прогнозирования поведения и запуск автоматических откликов или сигналов тревоги. Разработка модели, триггеров и откликов требует предварительного анализа данных. Проект операционно-аналитического решения должен включать подготовку исторических данных для предварительного задания начальных значений в моделях поведения. Например, в модели розничной торговли требуется оценить типичные наборы взаимодополняющих продуктов в покупательских корзинах. В моделях прогнозирования фондового рынка обычно используются исторические данные о котировках и динамике их изменения. Расчеты пороговых значений запуска отклика на основании предварительно заполненных полей также обычно производятся заранее.

После подтверждения полезности и окупаемости прогностических моделей ретроспективные данные в них начинают дополняться и замещаться текущими (включая поступающие в режиме реального времени и потоковые, структурированные и неструктурированные). Решение должно гарантировать корректную обработку потоков оперативных данных согласно правилам модели, безошибочное срабатывание сигнализаций о выходах измеряемых параметров за пределы допусков и защиту от ложных срабатываний автоматики.

1.3.15 Визуализация данных[1]

Визуализация данных — процесс интерпретации концепций, идей и фактов через наглядные представления, включая фотографии, рисунки, коллажи и всевозможные графики и схемы. Визуализация упрощает понимание иллюстрируемых данных, обеспечивая наглядность и лаконичность их сводного (например, графического) представления. Визуализация позволяет предельно сжато и доходчиво отображать наиболее характерные данные с целью навести зрителей на полезные выводы о скрытых возможностях, рисках или смыслах.

Визуальные представления могут быть как статичными (например, в формате иллюстрированного отчета), так и анимированными, динамично обновляемыми и даже интерактивными, то есть позволяющими конечному пользователю переходить на различные уровни детализации, накладывать фильтры и иным образом упрощать себе визуальный анализ данных. В качестве варианта может предусматриваться и переключение пользователем режима отображения данных в инновационные форматы, такие как интерактивные географические карты и динамические ландшафтные пейзажи данных.

Анализ данных уже давно немыслим без средств визуализации. Все традиционные инструменты бизнес-анализа обязательно включают широкий выбор средств визуального представления

[1] Визуализация данных — динамично развивающаяся область прикладной науки. Принципы визуального представления данных, в целом, основываются на принципах инженерного проектирования (см.: Tufte, 2001; McCandless, 2012). В интернете можно найти множество ресурсов с примерами, как подтверждающими, так и опровергающими справедливость такого представления. См. также «Периодическую таблицу методов визуализации» (http://bit.ly/IX1bvI) и другие ресурсы, опубликованные на сайте швейцарского межуниверситетского проекта визуального ликбеза visualliteracy.org.

данных — таблицы, всевозможные линейные и круговые, плоскостные и объемные, столбчатые и полосчатые графики, гистограммы. С ростом спроса на наглядные данные безостановочно совершенствуются средства их визуализации.

По мере роста зрелости информационной аналитики новые способы визуального отображения данных становятся важным стратегическим преимуществом. Новый взгляд на данные позволяет выявить новые связи и закономерности, а следовательно, и новые возможности для бизнеса. По мере дальнейшего развития и совершенствования средств визуализации организациям придется взращивать такие команды бизнес-аналитиков, которые смогут обеспечивать им конкурентоспособность во всё более компьютерно-управляемом в потоковом режиме мире. И вот тогда бизнес-аналитическими отделами будут крайне востребованы эксперты с навыками визуализации — знатоки данных, художники данных, визионеры данных, — в дополнение к традиционно ценящимся архитекторам и разработчикам моделей данных. Это будет более чем оправданно, если помнить о рисках, проистекающих от искажающих восприятие обманчивых визуальных представлений (см. главу 2).

1.3.16 Объединение данных

Средства получения данных из различных источников и служб позволяют создавать различные агрегированные представления данных для нужд визуализации или анализа. Многие инструменты виртуализации поддерживают агрегирование через функциональность связывания данных из различных источников объединяющими элементами, то есть, по сути, тем же приемом, который традиционно использовался в реляционных моделях для связывания, к примеру, объекта и описания через внешний ключ. Техническая возможность создания различных данных, например, весьма полезна для получения пользовательских представлений и идеально подходит для реализации задач, которые возникают на фазах раскрытия источников или разведки ресурсов данных, позволяя получать быстрые и наглядные результаты. Этот метод может быть применен в веб-приложении, поскольку позволяет организовывать обмен защищенными нарезками, содержащими персональные или конфиденциальные данные, между поставщиками или провайдерами информационных услуг. В сочетании с алгоритмами обучения искусственного интеллекта такие агрегированные представления помогают выявлять интернет-сервисы, оснащенные интерфейсами с поддержкой обработки естественного языка.

2. ПРОВОДИМЫЕ РАБОТЫ

2.1 Стратегическое планирование потребностей бизнеса в больших данных

Стратегия организации в отношении сбора и анализа больших данных должна выстраиваться в согласовании с общей информационной стратегией бизнеса и являться ее неотъемлемой частью. Стратегическое планирование потребностей бизнеса в больших данных должно учитывать следующие критерии.

- **Какие проблемы пытается решить организация? Для каких целей нужны результаты анализа больших данных?** Одно из преимуществ науки о данных — возможность взглянуть на организацию под новым углом, зафиксировать отправную точку и оценить перспективы дальнейшего развития. Организация может определить, что данные нужны ей для понимания бизнеса и бизнес-среды, доказательства ценности планируемых новых продуктов, исследования или изобретения новых подходов к ведению бизнеса. Важно создать и закрепить процесс принятия на рассмотрение, оценки, отбора и утверждения таких инициатив на различных стадиях внедрения. Ценность и целесообразность инициатив должны переоцениваться неоднократно.
- **Какие источники данных использовать?** Внутренние источники обычно проще использовать, но они содержат весьма ограниченные данные. Внешние источники могут быть весьма полезными, но неподконтрольными в плане оперативного управления (находиться под управлением других организаций или вовсе никем не контролироваться, как, например, многие соцсети). Это обширное поле, и конкуренция на нем серьезная, а потому бывает трудно разобраться и определиться с выбором из множества предлагаемых источников нужных элементов или наборов данных. Единственное, что можно порекомендовать: старайтесь приобретать наборы данных, которые достаточно хорошо совместимы с уже накопленными, чтобы минимизировать расходы на интеграцию и освоение.
- **Своевременность и полнота данных.** Одни элементы данных могут регистрироваться или поступать из внешних источников в потоковом режиме, другие — в виде моментальных снимков состояний через заданные интервалы времени, третьи — и вовсе поступать в интегрированной или обобщенной форме. Оперативные данные в идеале должны поступать с минимальным запаздыванием, но только не в ущерб машинному обучению в тех случаях, когда оно предусмотрено. В целом, алгоритмы обработки динамических потоковых данных принципиально отличаются от алгоритмов обработки статичных наборов данных. Постарайтесь придерживаться следующего правила: степень интеграции данных на стадии приема входных сигналов должна соответствовать минимальным потребностям пользовательских процессов ниже по потоку.
- **Согласованность с другими структурами данных.** Могут потребоваться изменения в структуре или контенте других данных с целью их согласования с наборами больших данных.
- **Учет влияния на существующие модели данных.** Планируйте распространение полученных в результате анализа и обобщения больших данных знаний на модели данных, используемые в управлении отношениями с клиентами, планировании продуктов, маркетинге и т. п.

На основе этой стратегии разрабатывается дорожная карта использования потенциала больших данных.

2.2 Выбор источников данных

Как и в случае любого другого проекта перспективных разработок, выбор источников больших данных, необходимых для развития науки о данных, должен диктоваться проблемами, которые стремится решить организация. Главное отличие в случае больших данных состоит в крайне

широком спектре потенциальных источников. Снимаются всякие ограничения по формату, что позволяет добавлять колоссальные объемы внешних данных к и без того расширенному спектру нарабатываемых внутри организации. Однако способность инкорпорировать внешние данные во внутренние решения привносит и множественные риски. Качество и достоверность данных и надежность их источников подлежат как первичной проверке, так и последующему подтверждению согласно запланированному графику. Среды больших данных позволяют быстро принимать колоссальные потоки информации, но для того, чтобы накопленные данные можно было использовать и хоть как-то ими распоряжаться, всё равно требуются учет и контроль хотя бы базовых фактов, относящихся к исходным данным, включая:

- происхождение;
- формат;
- смысл элементов данных;
- связи с другими данными;
- частоту обновления.

По мере появления обновленных и дополненных данных (например, демографической статистики, данных о спросе и продажах, спутниковых метеонаблюдений, новых наборов результатов масштабных научных экспериментов и т. п.) входные данные подлежат проверке и оценке на предмет их ценности, надежности и достоверности. Периодического пересмотра требуют и доступные источники данных, и процессы создания этих источников, а также планы поиска и подключения новых источников. При проработке источников больших данных основное внимание надлежит уделять следующим компонентам.

- **Основные данные.** Определите фундаментальные показатели, которые вас интересуют (например, продажи по каналам сбыта, если речь идет о торговле).
- **Детализация.** В идеале данные должны собираться на максимально доступном уровне детализации. Обобщить их по различным параметрам и признакам вы всегда успеете, сделав это согласно требуемому назначению.
- **Согласованность.** По возможности выбирайте источники данных таким образом, чтобы в них последовательно и согласованно отображались одни и те же показатели и применялись одни и те же ограничения. Это упростит визуализацию.
- **Проверка надежности источников.** Старайтесь убеждаться в достоверности и регулярности обновлений данных. Используйте только авторитетные источники с хорошей репутацией.
- **Выявление и подключение новых источников.** С одной стороны, важно своевременно выявлять ставшие доступными новые источники интересующих вас данных; с другой стороны, нельзя подключаться к ним без предварительной проверки их надежности. Не исключена и возможность нежелательных результатов вследствие подключения новых источников: например, искажения в отчетах или визуальных представлениях данных.

Риски, связанные с внешними источниками данных, обусловлены, в частности, необходимостью следить за соблюдением правил защиты конфиденциальных данных. Способность к быстрому усвоению и масштабной интеграции данных из множества разнородных источников делает возможным объединение исходных закрытых данных из различных защищенных, казалось бы, источников методом рекомбинации. Аналогичным образом и аналитический отчет может невольно выдать — через вводное описание, сводные данные или моделируемое состояние — не только группу населения, к которой он относится, но и конкретных лиц. Риск подобных нежелательных побочных эффектов особенно высок в тех случаях, когда результаты массово накопленной статистики применяются к узкой локальной выборке населения или граждан — и публикуются именно в таком виде. Например, демографические данные на национальном и региональном уровне обезличены; однако, если имеется возможность фильтрации по почтовым индексам, а тем более адресам, становятся вполне вычислимы и реальные лица, которые этими данными описываются[1].

Критерии выбора или фильтрации данных также сопряжены с риском. В любом случае выборочное включение данных в интегрированную модель требует объективного обоснования и управления во избежание привнесения субъективных искажений. Кроме того, не исключены и негативные последствия для визуальных представлений усеченных данных. С осторожностью следует применять и такие методы, как отбраковка данных, выходящих за пределы предельно допустимых отклонений, и искусственное ограничение области допустимых значений, и отсев редких элементов. В целом, практика улучшения фокусировки входных за счет удаления откровенно выбивающихся из общего ряда результатов широко распространена, но оправданной она может считаться лишь в тех случаях, когда имеет под собой объективные основания и применяется последовательно и единообразно[2] (см. главу 2).

2.3 Определение источников и загрузка данных

После выявления источников требуемых данных нужно получить к ним доступ (иногда речь может идти и о покупке или платной подписке) и загрузить исходные наборы данных, а также наладить бесперебойную загрузку обновлений в среду больших данных. В процессе этого не забывайте регистрировать все необходимые метаданные об источнике (происхождение, размер, датировку и прочие доступные сведения о контенте). Многие системы обработки вводных данных генерируют, как минимум, часть метаданных автоматически. После поступления данных в озеро их можно оценивать на пригодность к использованию для анализа различными методами. Поскольку построение моделей в рамках науки о данных — процесс по определению итерационный, поэтапно происходит и освоение данных. Шаг за шагом выявляйте пробелы в имеющихся массивах, ищите и подключайте ресурсы, необходимые для их заполнения. Для определения доступных

[1] В этом плане интересной представляется мысль Мартина Фаулера (Martin Fowler) о «прореживании данных» (*нем.* Datensparsamkeit), то есть об избавлении от бюрократической привычки собирать «полные данные». См.: http://bit.ly/1f9Nq8K

[2] Подробнее о систематических ошибках наблюдения или регистрации данных и их пагубном влиянии на интерпретацию результатов см.: http://bit.ly/2sANQRW, http://bit.ly/2oz2o5H и http://bit.ly/1rjAmHX

источников недостающих данных используйте средства профилирования, визуализации, добычи и иные методы науки о данных, позволяющие определять алгоритмы сбора и ввода данных сообразно гипотетическим моделям.

Перед интеграцией обязательно проверяйте качество данных. Объем контролируемых параметров качества может варьироваться от минимальной проверки процента заполнения полей до пропускания вводных данных через сложную последовательность аналитических проверок, позволяющих профилировать и классифицировать вводные данные и даже выявлять логические связи между их элементами. Подобная оценка позволяет оценить, действительно ли анализируемая выборка служит основой для проработки, и, если это так, перейти к проработке порядка хранения и доступа к данным (будь то мультипроцессорная параллельная архитектура, федеративная или распределенная схема и т. п.). Подобная проработка ведется непременно с участием экспертов в предметных областях (включая самих статистиков, то есть специалистов по большим данным) и инженеров-разработчиков платформенных решений.

По результатам экспертизы вырабатываются ценные заключения относительно возможности интеграции результатов обработки больших данных с другими наборами данных, такими как основные данные или архивные копии данных в хранилище. Эти же результаты могут использоваться для настройки параметров моделей обучения машинных алгоритмов, распознавания и проверки данных и т. п.

2.4 Выработка гипотез и выбор методов

Наука о данных сводится, по сути, к формулировке набора вопросов, ответы на которые помогают получать осмысленную картину или выявлять статистически значимые закономерности в имеющихся данных. Технически такая задача решается построением статистических моделей для выявления корреляций между элементами и наборами данных, а также временны́х тенденций их изменения. Ответы на любой вопрос априори неоднозначны, поскольку зависят от выбора исходных данных для обработки в рамках модели. Например, для прогнозирования будущей стоимости портфеля финансовых активов нужно задавать показатели доходности входящих в него ценных бумаг. При этом в моделях зачастую фигурирует множество подобных переменных, и рекомендуемой практикой считается поиск детерминированных результатов — иными словами, ожидаемые значения интересующей нас переменной рассчитываются при фиксировании остальных в качестве параметров на уровне с наибольшей вероятностью ожидаемых значений. Однако по мере накопления данных и развития модели значения таких параметров также должны переоцениваться и корректироваться методами машинного обучения. Работоспособность модели и результаты моделирования зависят от выбранного метода прогнозного или статистического анализа. Перед внедрением метод следует проверить на предмет допустимости и правдоподобия получаемых результатов на различных наборах вводных, включая самые маловероятные.

Качество результатов моделирования зависит, во-первых, от качества вводных данных, а во-вторых — от качества самой модели. Хорошие модели часто сами выявляют имеющиеся

в обработанном массиве корреляции, учет которых помогает уточнить модель и результаты. Например, метод кластеризации по k-средним позволяет для начала определить число кластеров, на которые следует разбить анализируемый массив данных с целью оптимизации дальнейшего анализа (см. главу 13).

2.5 Предварительная интеграция / Согласование данных для анализа

Подготовка данных к анализу требует понимания содержания данных, выявления дублирующих друг друга или вторичных данных в различных источниках и приведения общих для различных наборов элементов данных к согласованному представлению, чтобы их можно было корректно использовать.

Во многих случаях объединение источников данных — скорее искусство, чем наука. Например, предположим, что данные из первого источника поступают раз в сутки, а из второго — в виде ежемесячной сводки. Следовательно, чтобы те и другие можно было согласованно анализировать, в рамках представлений науки о данных ежесуточные данные подлежат накопительному обобщению в ежемесячные сводки, синхронизированные по срокам со сводками из второго источника.

Один из распространенных методов — модель интеграции данных через общий внешний ключ. Другой — сканирование и объединение данных через индексацию в СУБД; используется он, как правило, в рамках методов и алгоритмов сравнительного анализа наборов данных с целью выявления сходств и определения ссылочных связей. Часто данные подвергают предварительной обработке на предмет определения экспертным путем наиболее подходящего для них метода анализа. Для определения групп выходных данных полезно использовать кластеризацию. Есть и другие методы выявления коэффициентов корреляции между различными элементами моделируемых данных, помогающие правильно определять структуру отображаемых на выходе результатов. Все подобные приемы должны применяться на начальных этапах с целью понять, как будут выглядеть публикуемые данные в случае выбора в пользу той или иной методологии моделирования и анализа.

Большинство решений требуют интеграции основных и справочных данных для интерпретации результатов анализа (см. главу 10).

2.6 Исследование данных с помощью моделей

2.6.1 Заполнение и настройка предиктивной модели

Чтобы отконфигурировать любую предиктивную модель, нужно для начала ввести все предусмотренные моделью начальные или исторические данные: например, о клиентах, рынке, продуктах или иных факторах, учитываемых в модели, за исключением пускового фактора. Все расчеты по формулам с использованием предварительно загруженных начальных данных обычно производят заранее, чтобы обеспечить максимально оперативное реагирование на запуск события. Например, по истории онлайновых покупок для каждого клиента предварительно моделируется

рекомендуемая корзина. При прогнозировании поведения розничных рынков исторические данные о ценах, динамике их изменения и сезонных колебаниях дополняются информацией о покупательной способности, социальном и демографическом составе местного населения и метеорологическими данными.

2.6.2 Обучение модели

Первые прогоны данных через модель используются для уточнения параметров, — в этом и состоит смысл обучения модели. Впоследствии обучение повторяется всякий раз, когда нужно проверить обоснованность каких-либо новых гипотез или адекватность предлагаемых уточнений модели. Результатом обучения становится перенастройка модели. Пользоваться этим приемом нужно в меру. Злоупотребление обучением, особенно с использованием ограниченных наборов учебных данных, чревато переобучением модели на далекие от оптимальных результаты.

До получения устойчивых положительных результатов модель не передается в производственную среду. Нужно выявить все искажения и систематические погрешности в данных модели и скомпенсировать их изменением параметров или перетренировкой; последующая тонкая настройка может производиться уже в производственной среде, в том числе и с использованием алгоритмов самообучения, позволяющих модели постепенно корректировать параметры исходной настройки по мере замещения первоначальных вводных данных фактическими данными, получаемыми по результатам взаимодействий с целевыми группами населения. Для оптимизации наборов настраиваемых параметров можно использовать формулу расчета вероятностей взаимозависимых событий по теореме Байеса, причинно-следственную инверсию или вывод правил методом индукции. Наконец, можно разработать и составную модель данных для согласованного (так называемого «ансамблевого») обучения прогнозирующей модели, построенной по принципу сочетания сильных сторон, позаимствованных из нескольких более простых моделей.

Выявление отклонений или аномальных элементов данных (объектов, не укладывающихся в общую картину по тем или иным характеристикам) — критически важный компонент оценки модели. В случае волатильных наборов данных используйте дисперсный критерий расчета допустимых стандартных отклонений от среднего и доверительного интервала. Оба метода можно спокойно применять и к результатам профилирования данных. В таком случае вполне может оказаться, что выпадающие из общего ряда результаты как раз и являются искомыми или целевыми, если ставится задача поиска альтернативных вариантов, то есть прямо противоположная выявлению отклонений от сложившихся тенденций.

Для прогностического анализа используйте потоковые вводные данные для продолжения заполнения прогнозной модели и расчета текущего значения триггерного показателя, применяемого для запуска сигнализации о наступлении события или реакции на него. Обработка потока данных в режиме, близком к реальному времени, может потребовать особого внимания к таким инженерно-техническим аспектам систем, как обеспечение сверхнизкого времени запаздывания или сверхскоростной обработки сигнала. В некоторых моделях способность фиксировать изменение прогнозируемого значения за доли секунды — важнейшее условие пригодности решения,

а бывают и ситуации, требующие использования инновационных технологий, позволяющих регистрировать изменения чуть ли не со скоростью света.

Модели могут использовать всевозможные статистические функции и приемы, доступные из библиотек с открытыми исходными кодами, самой популярной из которых, вероятно, является «проект R». В бесплатной прикладной программной среде статистических вычислений проекта R все стандартные статистические функции реализованы через обращения к службам[1]. Дополнительные настраиваемые функции можно разрабатывать с помощью языка программирования и открывать совместный доступ к ним всем нуждающимся в них программным средствам, платформам и организациям.

По завершении создания, испытания, доработки и оценки прогностической модели организации остается решить, стоит ли разрабатывать и внедрять основанное на ней прикладное прогнозно-аналитическое решение. Средства операционной аналитики, работающие в режиме, близком к реальному времени, увы, часто бывают крайне ресурсоемкими и требуют столь существенных вложений в архитектуру, инфраструктуру и инженерное проектирование, что далеко не всегда окупаются.

2.6.3 Оценка модели

Итак, модель разработана и реализована, данные загружены на платформу и готовы к анализу, — тут-то и начинается, собственно, их изучение при помощи науки о данных. Поскольку на стадии проектирования модель прошла проверку и отладку на учебных наборах, с этого момента от нее разумно ожидать практических результатов, позволяющих уточнять бизнес-требования и судить о целесообразности и направлении дальнейших усилий по управлению изучаемым набором данных или же, напротив, о его непригодности. Вполне вероятен и вариант развития событий, ведущий к возникновению новых гипотез, проверка которых потребует дополнительных наборов данных.

Ученые исследуют данные с помощью сложных запросов и всевозможных алгоритмов обработки с целью выявить в изучаемых наборах тенденции и закономерности. Частенько приходится перепробовать множество различных математических и статистических функций, прежде чем удается усмотреть в данных хоть какие-то структурные признаки (например, кластеры или всплески и спады, чередующиеся с достаточно выраженной периодичностью, и т. п.). На этом этапе исследователи данных часто выстраивают гипотезы на основе результатов серий последовательных пакетных обработок, пока методом итерационных приближений не удается выявить достаточно закономерностей и корреляций для построения стройной модели связей между элементами данных.

Прикладная наука о данных имеет этическую составляющую, которую нужно учитывать и при моделировании данных. Модели могут давать непредсказуемые результаты, а могут и привносить в картину мира искажения, обусловленные необъективностью или даже предвзятостью

[1] Подробную информацию см. на веб-сайте *проекта R*: http://bit.ly/19WExR5

создателей этих моделей. В обязательном порядке следует требовать должной подготовки по вопросам этики от специалистов по разработке искусственного интеллекта. В идеале в программу обучения студентов по таким специализациям, как искусственный интеллект, информатика, вычислительная математика, кибернетика, аналитические и статистические методы обработки данных, — то есть всей совокупности предметов, прямо или косвенно относящихся к науке о данных, — должны входить курсы этики, информационной безопасности и защиты данных. Тем не менее одного только теоретического знания законов, принципов и правил этики недостаточно. Такие познания помогают практикам лишь сознавать свои этические обязательства перед всеми заинтересованными сторонами, а вот для их соблюдения нужно еще и подкрепить знание теории техническими возможностями для воплощения благих намерений в жизнь (Executive Office, 2016) (см. главу 2).

2.6.4 Создание визуальных представлений данных

Визуализация моделируемых данных должна соответствовать специфическим потребностям, напрямую связанным с назначением модели. Каждое наглядное представление должно давать ответ на какой-либо актуальный вопрос или позволять делать глубокие выводы. Следует прежде всего определить назначение и параметры визуального представления: отображаемые временны́е точки состояний; что требуется подчеркнуть — тенденции или исключения, связи между подвижными частями, географические различия или что-то еще.

Соответствующим назначению образом выберите тип визуального представления. Убедитесь в соответствии выбранного средства визуализации целевой аудитории; при необходимости повысьте или упростите уровень сложности и глубину детализации структуры графического представления сообразно восприимчивости или образованности аудитории. Не всякий пользователь осилит сложный интерактивный график. Сопроводите визуальное представление пояснительными текстами.

Визуализации должны представлять собой, по сути, связный рассказ в картинках о наглядно описываемых данных. В процессе составления такого «повествования о данных» могут всплывать и новые вопросы относительно контекста данных, которые стоят того, чтобы их изучить. Именно жанр рассказа с иллюстрациями позволяет сделать использование визуализаций максимально эффективным.

2.7 Внедрение и мониторинг

Модель, соответствующую бизнес-потребностям и недорогую с точки зрения технической реализации, можно внедрять в производственную среду с целью текущего мониторинга отслеживаемых с ее помощью показателей. После развертывания такие модели требуют доработки и эксплуатационного сопровождения. В моделировании существует несколько отработанных стандартных подходов к технической реализации процессов. Модели могут служить основой как для процессов пакетной обработки данных, так и для обмена сообщениями между службами интеграции данных в режиме реального времени. Также их можно встраивать в аналитические приложения,

обеспечивающие вводными данными системы управления решениями, средства ретроспективного анализа данных или приборные панели мониторинга и управления рабочими показателями.

2.7.1 Представление результатов анализа

Доходчивое отображение найденных в данных закономерностей (как правило, средствами визуализации) — последний шаг научного исследования данных. Находки должны представляться в связке с действенными рекомендациями, чтобы организация могла оценить отдачу от вложений в исследования методами науки о данных.

Для изучения выявленных новых связей полезно использовать, опять же, средства визуализации данных. По мере использования модели могут всплывать изменения в данных и связях между ними, тем самым раскрывая всё новую информацию о данных.

2.7.2 Итерации с добавлением источников

Презентация результатов и выводов часто приводит к инициированию нового цикла исследований. Наука о данных по определению строится по итерационному принципу; соответственно, и разработка больших данных — процесс итерационный: уроки, извлеченные из анализа предыдущего набора данных, часто ставят вопрос о необходимости привлечения альтернативных или дополнительных источников с целью окончательного подтверждения предварительных заключений или доработки и углубления существующей модели или моделей.

3. ИНСТРУМЕНТЫ

Технологический прогресс (достаточно вспомнить закон Мура[1] плюс экспоненциальный рост числа персональных мобильных устройств и техники с веб-интерфейсами), по сути, и привел к созданию индустрии больших данных и науки о данных. Для понимания того, что в этой отрасли происходит, нужно разобраться прежде всего с движущими факторами и направлениями ее развития. В настоящем разделе рассказано об основных инструментах и технологиях, сделавших возможным изучение больших данных.

Массово-параллельная обработка (Massive Parallel Processing, MPP[2]) стала одним из первых инструментов обработки больших данных, позволившим за кратчайшее время обрабатывать и анализировать колоссальные объемы информации. Сегодня мы только тем и занимаемся, что ищем иголки в стогах сена, и в будущем эта тенденция будет только усиливаться.

---

[1] Закон Мура — сформулированное основателем Intel Гордоном Муром (*англ.* Gordon Earle Moore, p. 1929) эмпирическое наблюдение об удвоении числа транзисторов в интегральных микросхемах процессоров и, как следствие, скорости обработки данных и производительности ЭВМ каждые два года. Эта закономерность вполне соблюдалась с момента ее формулировки в 1965 году и вплоть до середины 2000-х годов, когда был достигнут, по сути, физический предел возможностей повышения производительности одноядерных процессоров. — *Примеч. пер.*

[2] *сокр.* от *англ.* Massively Parallel Processing. — *Примеч. пер.*

Другие достижения, изменившие наши взгляды на данные и информацию, включают:

- продвинутые аналитические средства, встроенные в базы данных;
- аналитику неструктурированных данных (Hadoop, MapReduce и т. п.);
- интеграцию результатов анализа в операционные системы;
- универсальные средства визуализации данных в различных средах и на различных устройствах;
- семантическое связывание структурированной и неструктурированной информации;
- новые источники данных, ставшие доступными благодаря интернету вещей;
- продвинутую функциональность визуализации;
- новые методы и технологии обогащения данных;
- технологии и наборы инструментов для совместной работы.

Существующие архитектуры централизованных хранилищ данных с витринами и локальными хранилищами операционных данных (ODS) также всё чаще дополняются функциональностью, позволяющей нести дополнительную рабочую нагрузку по обработке больших данных. Технологии без реляционных связей (NoSQL) позволяют запрашивать, обрабатывать и сохранять слабоструктурированные и вовсе неструктурированные данные.

Раньше доступ к неструктурированным данным обычно происходил посредством обработки пакетных запросов по расписанию, что снижало оперативность согласования данных в локальных хранилищах с источниками. Теперь некоторые СУБД класса NoSQL включают технологии, позволяющие обходить эту проблему и существенно ускорять получение данных из источников. Масштабируемые распределенные базы данных так и вовсе обеспечивают автоматическое сегментирование (распределение потоков данных по серверам) с целью параллельного исполнения обработки запросов. Конечно, как и в любой другой базе данных, определение структуры данных и сопоставление структурированным элементам неструктурированных данных из анализируемого набора остается процессом, который приходится выполнять по большей части вручную.

Функциональность немедленных запросов к данным, отчетов и анализа может на вполне удовлетворительном уровне реализовываться с помощью технологий обращения к большим данным в оперативной или виртуальной памяти, которые позволяют конечным пользователям конструировать SQL-подобные запросы к неструктурированным данным. В некоторых инструментальных средах предусмотрены еще и адаптеры SQL/NoSQL, позволяющие отправлять стандартные для реляционных моделей запросы к неструктурированным данным и получать вполне совместимые с SQL-представлениями результаты (понятно, что с ограничениями и скрытыми подвохами). Немаловажно, что такие адаптеры нередко позволяют распространять привычные средства анализа данных на неструктурированные массивы.

Предлагаемые наборы технических средств определения критериев принятия решений, реализации процессов и формирования пакетов предложений профессиональных услуг также способствуют упрощению и ускорению процесса выбора исходного набора инструментов. Как

и в случае оснастки, необходимой для хранилища данных / бизнес-аналитики, тут критически важно учесть все доступные варианты и сравнить их плюсы и минусы: строить/создавать собственные решения? покупать/арендовать готовые продукты и услуги (в частности, SaaS)? Как отмечалось в главе 11, следует взвешенно соизмерять затраты и выгоды от использования средств, доступных в облаке, по сравнению с издержками и выигрышами от проектирования с нуля собственного или приобретения и развертывания коммерческого программного обеспечения. При этом должны учитываться и затраты на обновления, продление подписки или возможные замены неподходящих приложений. Согласование всех этих вопросов с действующими соглашениями об уровнях обслуживания (SLA) или операционной поддержки (OLA) также будет не лишним, поскольку позволит хоть как-то спрогнозировать издержки реализации и согласовать между собой привлекательные ставки платы за обслуживание с суммами штрафов за нарушения.

3.1 Технологии и архитектуры MPP без разделения ресурсов

Технологии баз данных с массово-параллельной обработкой (MPP) без разделения ресурсов (shared-nothing) сделались стандартом вычислительных технологий, используемых для анализа и изучения наборов больших данных. В базах данных с MPP потоки обрабатываемых данных сегментируются (логически распределяются) по множеству серверов (вычислительных узлов), каждый из которых располагает достаточным объемом выделенной памяти для локальной обработки адресованного ему потока данных. Согласование же обработки осуществляется, как правило, с помощью головного сервера хост-системы, контролирующего все процессы на задействованных в распределенной сети обработки локальных серверах. Никаких совместно используемых вычислительных ресурсов, дисковых пространств или оперативной памяти при таких архитектурных решениях не предусмотрено, отсюда и уточнение — «полностью раздельные» вычисления.

Появление MPP-архитектуры стало логичной реакцией на неспособность традиционных вычислительных схем (с индексацией, распределенными наборами данных и т. д.) обеспечивать достаточно высокую скорость обработки запросов, обращенных к огромным массивам табличных данных. Самые мощные классические вычислительные платформы (включая суперкомпьютеры Cray и им подобные) будут часами, если не сутками, обсчитывать сложную модель, примененную к таблице данных, содержащей триллион строк.

А теперь представьте себе батарею из сотен серийных серверных компьютеров, работающих параллельно под управлением головной хост-машины. Каждый получает запрос на обработку своей доли вычислений. Скажем, если та же таблица с триллионом строк распределяется для обработки между тысячей параллельно подключенных серверов, то скорость обработки запроса, обращенного к триллиону записей, повышается на три порядка, поскольку каждому из 1000 компьютеров нужно обработать «всего лишь» миллиард строк. Причем MPP-архитектура хороша еще и линейной масштабируемостью, что делает ее крайне привлекательной для исследователей и пользователей больших данных: перестало хватать вычислительных

мощностей — к платформе без каких-либо архитектурных изменений пристраиваются дополнительные серверы.

Технология MPP также позволила реализовать встроенные в СУБД на уровне процессоров исполняемые функции статистического анализа (кластеризацию методом k-средних, регрессионный анализ и т. п.). Распределение рабочей нагрузки на процессорный уровень значительно ускоряет обработку аналитических запросов и тем самым стимулирует всё новые инновационные методы изучения больших данных.

Система, обеспечивающая автоматическое распределение данных и параллельную рабочую нагрузку на все доступные (локально) серверные процессоры, — оптимальное решение для анализа больших данных.

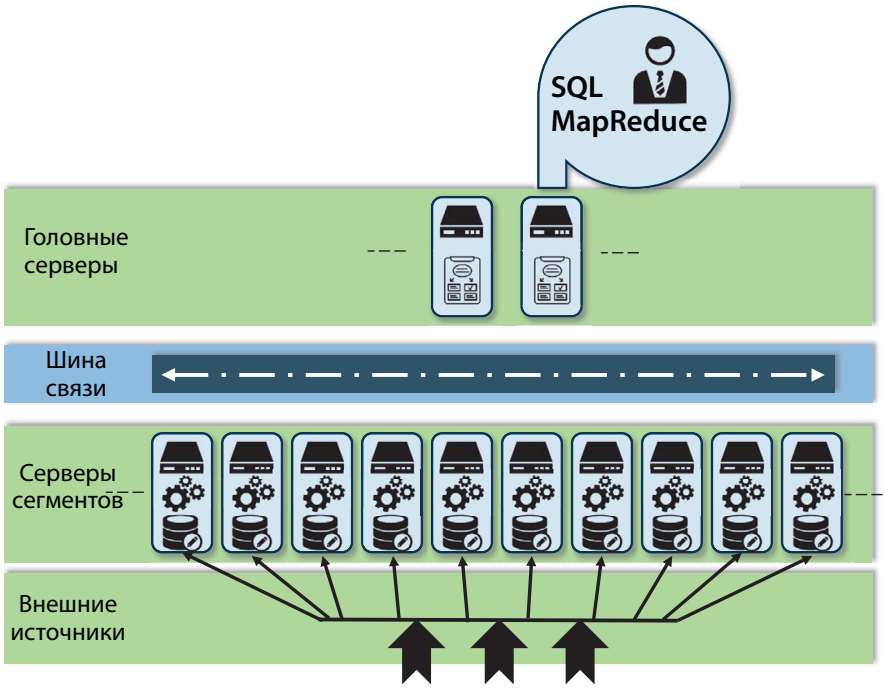


Рисунок 102. Колоночная архитектура[1]

Объемы данных продолжают стремительно расти. Компании могут реагировать на это, наращивая вычислительные мощности по мере надобности простым добавлением новых вычислительных узлов, поскольку архитектура MPP предельно упрощает параллельное подключение десятков, сотен или тысяч ядер, выстраивающихся в ЭВМ. При этом в полностью раздельной по ресурсам МПП-архитектуре с поддержкой линейного масштабирования каждое ядро используется с максимальным КПД, и это дополнительно повышает производительность обработки огромных массивов данных.

[1] *Источник*: «Greenplum Database 4.0: Critical Mass Innovation», White Paper, August 2010.

3.2 Базы данных на основе распределенных файловых систем

Технологические решения на основе распределенных файловых систем, подобные HDFS (Hadoop Distributed File System), служат недорогим способом хранения больших объемов разнородных данных. В HDFS можно сохранять файлы любого размера, формата и типа — структурированные, частично структурированные и не структурированные вовсе. Как и в MPP-архитектуре, файлы данных распределяются между серверами. Решение идеально подходит для надежного хранения данных (поскольку файлы реплицируются), а вот с доступом к ним с помощью структурированных запросов (наподобие SQL) и, как следствие, с онлайновым анализом данных, хранящихся в распределенных файловых системах, возникнут серьезные проблемы.

Благодаря относительно низкой стоимости Hadoop стала популярной перевалочной базой, выбираемой многими организациями. А из Hadoop данные затем можно по мере надобности переносить в поддерживающие обработку аналитических запросов среды базы данных, например в MPP. Впрочем, некоторые организации, не особо озабоченные оперативностью, обрабатывают сложные запросы в рамках проектов науки о данных и прямо в Hadoop; правда, на получение результата в этом случае уходят часы и сутки, а не минуты, как в MPP.

В распределенных файловых системах используется специфическая терминология модели MapReduce[1]. Три основных этапа аналитической обработки больших данных на этом языке называются так:

- **Map — отображение**: идентификация и получение данных для анализа;
- **Shuffle — перетасовка**: выборка и компоновка в соответствии с выбранной схемой анализа;
- **Reduce — свёртка**: вычистка дублей или агрегирование данных с целью радикального уменьшения объема данных в полученном результате и сохранения в нем только нужных элементов.

Эти этапы могут в различных сочетаниях, последовательно или параллельно, включаться во многие аналитические инструменты, что обеспечивает возможность весьма сложных манипуляций с данными.

3.3 Алгоритмы «в базе данных»

Алгоритм «в базе данных» основан на принципе полностью независимой обработки каждым процессором в архитектуре MPP своего собственного аналитического алгоритма, что открывает возможность нового подхода к анализу больших данных по принципу раздельной реализации различных математических или статистических функций на уровне вычислительных узлов. Открытые библиотеки встраиваемых в масштабируемые БД алгоритмов машинного обучения, решения статистических и аналитических задач как в ядре, так и во внешней памяти разработаны для различных архитектур, включая MPP самых современных СУБД, что обеспечивает

[1] MapReduce (~ «отображение-свёртка») — модель и язык распределенных параллельных вычислений на больших данных, предлагаемые компанией Google. — *Примеч. пер.*

максимальное приближение вычислений к данным. А чем ближе вычислительные мощности к данным, тем меньше непродуктивные затраты времени и больше возможностей для расчетов по сложным алгоритмам (таким, как кластеризация по k-средним, линейная или логистическая регрессия, U-критерий Манна — Уитни, расчет сопряженных градиентов, анализ когорт и т. д).

3.4 Облачные хранилища больших данных

Ряд поставщиков предлагают облачные решения для хранения и интеграции больших данных, иногда с поддержкой аналитических возможностей. Руководствуясь стандартами, определяемыми такими провайдерами, клиенты загружают свои данные в облачную среду, после чего поставщик решения может дополнительно дорабатывать данные, распоряжаясь ими либо как открытыми наборами, либо на условиях, определяемых подключенными к облачному хранилищу организациями. В итоге любой клиент получает возможность изучать и анализировать весь массив больших данных, накопленный в облаке. Пример применения: агрегирование розничных предложений по предметным областям в сочетании с географическими профилями спроса и продаж в обмен на бонусные мили авиакомпаний — участников схемы, предлагаемые всем покупателям, соглашающимся на использование их данных подобным образом.

3.5 Языки статистических вычислений и графических представлений

Упоминавшийся уже в разделе 2.6.2 проект R предлагает всем желающим язык написания сценариев и бесплатную среду для статистических вычислений и графического представления их результатов. Язык R позволяет реализовывать широкий спектр методов статистического анализа данных, включая линейное и нелинейное моделирование, классические статистические испытания, анализ временны́х рядов, классификацию и кластеризацию данных в неизученных массивах. Поскольку это язык сценарного анализа, модели, разработанные на *R*, можно затем реализовывать в самых разнообразных средах и на различных платформах, что открывает широкие возможности для совместной работы и интеграционных усилий поверх географических и организационных границ. Плюс к тому среда *R* поддерживает графопостроение на уровне, пригодном для публикации без доработок, а также математические символы и формулы, доступные конечным пользователям.

3.6 Средства визуализации данных

Традиционные средства визуализации данных включают два компонента — численное и графическое представления. Продвинутые средства визуализации и раскрытия данных используют оптимизированную для обработки в оперативной памяти архитектуру поддержки интерактивного взаимодействия пользователя с данными. Закономерности и связи в больших наборах данных в численном представлении бывают трудноуловимыми, а вот при выборе сложного графического режима визуализации динамики загрузки данных даже с тысячами точек любые неравномерности сразу бросаются в глаза и вызывают желание их проанализировать.

Инфографика (как теперь принято называть эффектные стилизованные наглядные графические представления данных) также может быть сделана интерактивной для большей доходчивости.

В маркетинге и рекламе это поняли первыми и стали использовать анимированную и иными способами приукрашенную графику для повышения привлекательности образа всего, что ею описывается. Следом за рекламщиками журналисты, блогеры, учителя всех уровней, школ и предметов осознали всю степень полезности инфографики для анализа трендов, эффектности презентаций и распространения сообщений. Современные методы компьютеризованного наглядного представления информации — лепестковые диаграммы («радарные карты»), графики сопоставления многих измерений, облака тегов, тепловые карты и многие другие типы карт данных — теперь поддерживаются многими наборами инструментов. При таком богатстве визуализаций пользователям гораздо проще оперативно улавливать изменения в динамике данных, переходить к просмотру связанных элементов, разбираться в причинно-следственных связях и выявлять настораживающие признаки до того, как начнутся реальные неприятности. Подобные средства визуализации обладают, как минимум, следующими преимуществами по сравнению с традиционными:

- наличие самых сложных разновидностей аналитической визуализации, таких как спектры с мультиплетами, осциллограммы, тепловые карты, гистограммы, каскадные диаграммы и т. д. и т. п.;
- встроенная поддержка обеспечения соблюдения стандартов и рекомендаций в области наглядных представлений;
- интерактивные возможности углубленного исследования визуально выявленных закономерностей и тенденций.

4. МЕТОДЫ

4.1 Аналитическое моделирование

Имеется несколько источников с открытыми кодами, а также ряд провайдеров услуг облачных хранилищ, предлагающих программное обеспечение для разработки моделей данных, в том числе с поддержкой визуальной разработки, веб-сканирования и оптимизации линейного программирования. Для создания, распространения и обработки моделей, созданных другими приложениями, лучше использовать средства, поддерживающие язык разметки прогнозных моделей PMML (predictive model markup language), основанный на XML.

Доступ в режиме реального времени решает большинство проблем с запаздыванием, неизбежных при пакетной обработке. Apache Mahout — проект с открытым кодом, который занимается разработкой библиотеки алгоритмов машинного обучения, всерьез нацелившийся на решение задачи автоматизации исследований *больших данных* посредством разведки рекомендованных источников, классификации документов и кластеризации элементов. Это направление развития *науки о данных* идет в обход традиционных для решений класса MapReduce методов доступа к данным и, в частности, вовсе не предусматривает пакетной обработки. Вместо этого в полной мере используются преимущества прямого доступа через API-интерфейсы в слой хранения

распределенных данных HDFS, что позволяет поддерживать широкий спектр разнообразных функций, начиная от обработки SQL-запросов и потоковой трансляции контента и заканчивая машинным обучением и библиотеками шаблонов визуализации.

Аналитические модели могут варьироваться по глубине и направленности.

- **Описательное моделирование** нацелено на получение сводных данных или компактных представлений структурированных данных. Такой подход далеко не всегда позволяет проверять гипотезы, выявлять причинно-следственные связи или предсказывать исходы, но зачастую он и не преследует подобных целей. Зато он позволяет использовать алгоритмы определения или уточнения связей или параметрических зависимостей между множеством переменных и тем самым делает их пригодными для описания.
- **Разъясняющее моделирование** заключается в применении к данным статистических моделей с целью проверки гипотезы о причинно-следственных связях в рамках теоретических построений. Технические приемы в данном случае мало отличаются от используемых при добыче данных и в аналитическом прогнозировании, а вот цель преследуется иная: не предсказать исходы или смоделировать прогнозы, а просто проверить соответствие результатов, получаемых с помощью гипотетической модели, реальным данным.

Предиктивная аналитика, или аналитическое прогнозирование, ставит целью обучение на примерах в процессе отладки (тренировки) модели. Эффективность итеративного обучения проверяется по способности модели к достоверным прогнозам на основе независимого набора контрольных данных. Полученный результат используется для выбора следующего урока и одновременно оценки качества модели. Кроме того, работоспособность модели оценивается еще и по такому критерию, как способность к обобщению при обработке нового набора данных, включая выявление ошибочных обобщений.

Избегайте избыточной подгонки или чрезмерной обученности модели. Такой риск присутствует при обучении на непрезентативных выборках, при излишней сложности модели относительно данных, а также при особой чувствительности модели, приводящей к описанию шумовых помех вместо глубинных связей. Дополнительно используйте скользящую перекрестную проверку на последних k-частях данных, чтобы своевременно уловить момент, когда дальнейшее обучение не требуется, поскольку не приводит к улучшению способности модели к обобщению.

Ошибки обучения неуклонно снижаются по мере усложнения модели и теоретически могут быть сведены к нулю. Следовательно, они не могут использоваться в качестве показателя ошибки тестирования. Случайным образом разделите массив данных на три части — учебную, проверочную и контрольную выборки. На учебном наборе натаскивайте модель, контрольный набор используйте для прогнозирования ошибки выбора, а проверочный набор — для оценки ошибки обобщения в окончательной модели.

Многократное использование одного и того же тестового набора может повлечь заниженную оценку ошибки тестирования по сравнению с ее истинной величиной. В идеале нужна

скользящая перекрестная проверка на k-частях. Разбейте набор данных на k случайных равновеликих выборок. Проведите курс обучения модели на k–1 выборке из k, значения прогнозируемых переменных в которых, естественно, сильно коррелированы. Наконец, протестируйте модель на последней, k-й выборке, а затем определите ошибку обобщения по всем k-частям. Для получения численной оценки пригодности модели для использования в анализируемом контексте можно применить к полученным результатам различные статистические критерии.

4.2 Моделирование больших данных

Моделирование больших данных — задача технически крайне сложная, но решать ее необходимо, если организация действительно нацелена на описание имеющихся в ее распоряжении данных с целью их постановки под контроль. Традиционные принципы архитектуры данных предприятия в равной мере применимы и к большим данным: они требуют интеграции, спецификации и управления.

Главный стимул к разработке физической модели хранилища данных — обеспечить их накопление и возможность быстрой обработки запросов. На большие данные этот стимул не распространяется. Но это не повод отказываться от моделирования или отдавать его на откуп первому попавшемуся стороннему разработчику. Ведь ценность моделирования данных заключается еще и в том, что в его процессе люди учатся понимать данные и разбираться в их содержании и смысле. Так что применяйте проверенные методы моделирования данных, но только отдавая себе отчет в множественности и разнообразии источников. Разработайте модель предметной области — хотя бы обобщенную, — чтобы ее можно было использовать для определения объектов и отношений между ними в привязке к контексту и для создания дорожной карты, в точности так же, как это делается применительно к любым другим видам данных. Труднее всего дается именно составление понятной и полезной общей картины применения этих гигантских массивов данных, да еще и ценой разумных затрат.

Выработайте понимание связей между различными данными и наборами. В случае данных различного уровня детализации внимательно следите за тем, чтобы какие-либо элементы или значения данных не были учтены два и более раз на разных уровнях. К примеру, категорически не рекомендуется сочетать наборы атомарных и сводных данных.

5. РЕКОМЕНДАЦИИ ПО ВНЕДРЕНИЮ

Многие общие принципы управления хранилищами данных автоматически переносятся и на управление большими данными, включая: обеспечение надлежащей проверки надежности источников и достоверности данных; наличие метаданных в объеме, достаточном для понимания и возможности использования данных; управление качеством данных; изыскание способов интеграции данных из различных источников; обеспечение информационной безопасности и защиты данных (см. главы 6–8). Основные отличия при реализации среды больших данных обусловлены

тем, что приходится, по сути, решать систему уравнений со многими неизвестными: как и для чего будут использоваться данные? Какие данные будут считаться особо ценными? Каковы будут сроки хранения данных?

Скорость и объемы поступлений больших данных способны зародить мысль о том, что их можно пустить на самотек, тем более что времени на реализацию механизмов управления такими потоками, по сути, не остается. Это крайне опасное заблуждение. Чем больше массивы накопленных данных, тем важнее управлять их своевременной обработкой и инвентаризацией. В противном случае озеро данных быстро превращается в гнилое болото.

Освоение больших данных далеко не всегда требует от организации получения законных прав на обладание ими или принятия на себя обязательств по их непременному изучению. Возможны варианты периодической аренды платформы больших данных на срок, требующийся для исследования и предварительной оценки заинтересовавших вас данных. По результатам таких изысканий вы быстро определите области потенциального интереса. Да, и проводить подобную предварительную оценку нужно всегда и в любых ситуациях, прежде чем скачивать какие бы то ни было массивы данных в ИТ-среду организации, будь то озеро, хранилище данных или даже буферный накопитель.

5.1 Согласование со стратегией организации

Любая программа в области сбора и изучения больших данных должна выстраиваться сообразно стратегическим целям, планам и задачам организации. Утвержденная стратегия больших данных сразу же должна приниматься во внимание и включаться в стратегические планы управления доступом пользователей, защиты данных, управления метаданными, включая генеалогию, и управления качеством данных.

В стратегии должны быть задокументированы цели, подход и руководящие принципы. Для извлечения максимума из больших данных организации как таковой нужно наработать определенные навыки и способности. Используйте стандартные приемы управления возможностями и потенциалом для согласования бизнес-инициатив и ИТ-проектов в рамках дорожной карты программы. Обязательными документами являются стратегии управления:

- жизненный цикл информации;
- метаданные;
- качество данных;
- сбор данных;
- доступ к данным и защита данных;
- руководство данными;
- конфиденциальность данных;
- обучение и восприятие;
- текущая работа.

5.2 Оценка готовности / Оценка рисков

Как и любой проект развития, инициатива в области больших данных или науки о данных должна соответствовать реальным потребностям бизнеса. Прежде чем ее выдвигать, оцените готовность организации к адекватному восприятию подобного проекта по следующим критическим параметрам успеха.

- **Польза для бизнеса.** Насколько хорошо инициативы в области больших данных / науки о данных соответствуют потребностям и вписываются в канву деятельности компании? Для успеха они должны сулить перспективы качественного скачка в плане усиления бизнес-функций или развития бизнес-процессов.
- **Готовность бизнеса.** Готовы ли бизнес-партнеры к долгосрочной, поэтапной поставке продукта? Согласны ли создать центры повышения квалификации для устойчивой поддержки новых версий? Не слишком ли расплывчаты в целевом сообществе общие представления, или не слишком ли скудны практические навыки, чтобы эту пропасть можно было перепрыгнуть одним махом?
- **Экономическая целесообразность.** Проводилась ли консервативная оценка материальных и нематериальных выгод от реализации проекта? Были ли учтены на стадии экономического обоснования варианты покупки/аренды готовых решений вместо построения собственных с нуля?
- **Прототипирование.** Нельзя ли оперативно построить прототип предлагаемого решения для какой-то подгруппы целевой пользовательской аудитории, чтобы наглядно продемонстрировать ценность модели? Масштабные реализации методом «Большого взрыва» — идея эффектная, но многим она представляется чересчур рискованной, поскольку на кону стоят слишком большие деньги. Поэтому маломасштабная, но надежная реализация сбора в меру больших, но очень полезных и прибыльных данных — хороший способ побороть настороженность.

Однако самые трудные решения, вероятно, придутся на стадию согласования выделения средств на приобретение данных, разработку платформы и обеспечение программы прочими необходимыми ресурсами.

- Существует множество источников цифровых данных, и все их не подключишь, не купишь и не скачаешь. Какие данные реально нужны? Как это обосновать? Какие данные нужно приобрести в постоянное пользование, а какие достаточно взять во временную аренду или получить по подписке?
- На рынке имеется множество программных средств и методологий. Какие из них лучше всего подойдут для общих нужд? В каком сочетании?
- Как вовремя привлечь специалистов, обладающих всеми необходимыми на данном этапе навыками? Как удержать таланты на стадии реализации проекта? Какие альтернативы имеются в плане аутсорсинга, сетевого и облачного сотрудничества?
- Воспитание собственных талантов — дело стоящее, но не в тех случаях, когда сроки сдачи работы поджимают.

5.3 Организационные и культурные изменения

Бизнесмены — это по определению люди, полностью занятые делом, а потому нужно как-то заставить их осознать пользу, которую способна принести их делу продвинутая статистическая аналитика. Непонимание вполне успешно устраняет грамотно поставленная программа информационно-разъяснительной работы с целевой аудиторией. Центр компетенций, к примеру, может предлагать тренинги, распространять стартовые наборы, разрабатывать образцовые стандарты практики и рекомендации по их применению, полезные советы по изысканию источников данных и хитрые приемы их подключения, да и просто делать массу всего полезного с точки зрения доказательства полезности решений или предъявления бизнес-пользователям наглядных свидетельств богатства возможностей, которые открывает постепенный переход на модель самообслуживания. В дополнение к управлению знаниями такой центр способен выполнять и функции информационно-коммуникационного центра, обеспечивающего непрерывность взаимодействия между разработчиками, конструкторами, аналитиками и сообществами потребителей данных.

Как и в случае DW/BI, реализация среды больших данных требует согласованных усилий специалистов различных профилей. В частности, в проекте должны принимать участие следующие лица.

- **Архитектор платформы больших данных**: подбор и конфигурирование аппаратного обеспечения, операционных систем, файловых систем, служб и т. п.
- **Архитектор загрузки данных**: анализ данных, системы записей, моделирование, карты преобразования данных и т. п. Также может отвечать за сопоставление источников кластерам Hadoop с целью обработки запросов и анализа.
- **Специалист по метаданным**: интерфейсы, архитектура и контент *метаданных*.
- **Ведущий аналитик**: выбор или разработка аналитических средств для конечных пользователей, реализация новейших методологических рекомендаций в связанных наборах инструментов, оптимизация доступа конечных пользователей к результатам обработки данных.
- **Специалист в области науки о данных**: снабжение всех вышеперечисленных специалистов необходимыми сведениями о теории, методологии и практике статистического анализа, а также содействие в разработке необходимых средств прикладных вычислений и технических приложений.

6. РУКОВОДСТВО В ОБЛАСТИ БОЛЬШИХ ДАННЫХ И НАУКИ О ДАННЫХ

Большие данные, как и любые другие, требуют обеспечения надлежащего руководства. Изыскание и анализ источников, поглощение и усвоение, обогащение и публикация — все эти процессы требуют, помимо технического контроля, еще и механизмов контроля со стороны бизнеса, в частности с целью решения вопросов следующего характера.

- **Изыскание источников.** Что и когда искать? Как выбрать наилучший источник данных для конкретного исследования?
- **Совместное использование.** Какие соглашения и договоры о совместном доступе к данным, распространении результатов, обмене данными и т. п. нужно заключить и на каких условиях, включая внутриорганизационные договоренности и контракты со сторонними поставщиками, клиентами и партнерами?
- **Метаданные.** Как трактуется смысл и значение различных данных там, откуда они поступают? Как обеспечить правильную интерпретацию результатов получателями?
- **Обогащение.** Нуждаются ли данные в обогащении? Какими методами? Какую пользу принесет обогащение данных?
- **Доступ.** Какие из результатов публиковать? Кому и когда открывать к ним доступ? Как регулировать порядок доступа?

Для грамотного решения вопросов, касающихся оборота данных и обращения с данными, требуется целостное представление о данных, имеющихся в распоряжении предприятия.

6.1 Управление каналами визуализации

Важнейший фактор успеха реализации программы статистических исследований — правильный выбор средств визуализации, которые должны максимально соответствовать потребностям пользовательского сообщества. В зависимости от размера и характера организации возможно использование самых разнообразных средств визуального отображения данных в различных процессах. Важно всякий раз убеждаться, что используемые средства визуализации не слишком переусложнены и соответствуют уровню понимания целевой группы пользователей. Высокообразованные продвинутые пользователи будут со временем становиться всё требовательнее в своих запросах к сложным визуальным представлениям. Скоординированный подход к проектированию архитектуры данных предприятия, управлению портфелем информационных ресурсов и техническому сопровождению систем — необходимое условие надлежащего контроля каналов визуализации как внутри портфеля, так и внешних. Не забывайте, что любая смена поставщиков данных, провайдеров контента или критериев выбора отображаемой информации с большой вероятностью приводит к изменению набора и структуры элементов, доступных для визуального просмотра ниже по потоку, и может потребовать перенастройки средств визуализации с целью восстановления их эффективной работы.

6.2 Наука о данных и стандарты визуализации

Передовой практикой сегодня считается создание экспертного сообщества для определения и публикации стандартов визуализации, руководств по их применению и спецификаций артефактов, производимых при использовании различных методов выдачи визуального контента. Особую важность соблюдение стандартов визуализации имеет в тех случаях, когда контент адресован клиентам или строго регламентирован. Стандарты визуализации могут регулировать:

- выбор программных средств в зависимости от аналитической парадигмы, сообщества пользователей или предметной области;
- порядок и сроки запросов новых данных и/или обновлений;
- технологические процессы обработки наборов данных различных типов;
- процедурные правила и нормы нейтрального и объективного представления экспертных заключений во избежание привнесения искажений в результаты или их предвзятой интерпретации; соблюдение всех методологических требований, предъявляемых к статистическим исследованиям, включая:
  - объективность критериев формирования выборок, включения и исключения точек/элементов данных;
  - формулировку гипотез, проверяемых моделями;
  - статистическую достоверность и значимость результатов;
  - обоснованность и корректность интерпретации результатов;
  - применимость и уместность использованных методов.

6.3 Безопасность данных

Наличие надежного процесса обеспечения информационной безопасности и защиты данных само по себе является ценнейшим ресурсом в активе организации. Для больших данных, как и для любых других, должны устанавливаться правила обращения, защиты и контроля доступа, дополненные средствами мониторинга их соблюдения. Особое внимание должно уделяться воспрепятствованию злоупотреблениям персональными данными и обеспечению их защищенности на протяжении всего жизненного цикла.

Проработайте уровни доступа к данным авторизованных сотрудников и пользователей. Уровни доступа к данным, получаемым по подписке, должны соответствовать соглашениям с провайдерами. Настройте службы данных раздельно по профилям сообществ пользователей, с тем чтобы можно было ограничивать выдачу конфиденциальных и иных данных лишь сообществами, имеющими право обрабатывать эти данные с целью освоения; в выдачах остальным категориям данные должны быть скрыты. Часто организации определяют и прямые запретительные правила (например, блокируют возможность запросов данных по фамилиям, адресам или номерам телефонов). Для защиты строго конфиденциальных или персональных идентификационных данных (номеров ID-документов, кредитных карт и т. п.) используются шифрование или обфускация. При необходимости может быть выбран метод шифрования, при котором сохраняются соотношения значений, но не сами значения, что позволяет пользователям выявлять статистические закономерности без доступа к фактическим данным.

Рекомбинацией называют возможность реконструкции или восстановления исходных персональных идентификационных или конфиденциальных данных. Такой риск нужно учитывать при обеспечении ИБ и защиты не только «обычных», но и больших данных. В частности, результаты анализа могут нарушать неприкосновенность личной информации, даже если до начала анализа было невозможно определить, к кому относится каждый отдельно взятый элемент данных.

Во избежание подобных и иных недоразумений, приводящих к нарушениям в сфере информационной безопасности и защиты данных, критически важно получать четкое понимание результатов обработки еще на уровне управления метаданными. А для этого требуется знание назначения и задач использования или анализа данных, а также распределение ролей исполнителей этих задач. Отдельным доверенным лицам может быть санкционирован доступ к незашифрованным данным подобного рода в режиме чтения, но лишь по крайней служебной необходимости, далеко не всем и уж точно не с целью углубленного анализа (см. главы 2 и 7).

6.4 Метаданные

В рамках инициативы по сбору и исследованию больших данных организация формирует общий набор данных, созданный с использованием различных подходов и стандартов. Интеграция столь разнородных данных — задача крайне трудная. Без метаданных, описывающих каждый набор, на успешное использование всей совокупности данных рассчитывать не приходится. Управление метаданными должно вестись тщательнейшим образом, начиная со стадии освоения данных. Сообществу пользователей нужно предоставить инструменты, позволяющие создавать и вести главный список наборов данных, в котором каждому набору должны соответствовать метаданные, характеризующие структуру, содержание и качество данных, включая первоисточник и происхождение данных; определения и назначения объектов и элементов данных. Технические метаданные можно собирать с помощью разнообразных инструментальных средств работы с большими данными, включая слои хранения, инструменты интеграции, управления основными данными, а иногда и получать прямо из файловых систем — источников данных. Следует также сопоставить и оценить плюсы и минусы обработки входящих данных с целью определения метаданных в потоковом режиме или в статике, а также определить, не требуются ли какие-то дополнительные вычисляемые элементы данных, необходимые для поддержки возможности отслеживания происхождения данных до первоисточника.

6.5 Качество данных

Под качеством данных понимается мера их соответствия ожиданиям: чем меньше отклонение, тем выше степень соответствия данных ожиданиям и, как следствие, качество данных. В высокотехнологичных средах стандарты качества, по идее, должны определяться достаточно просто (хотя на практике приходится наблюдать немало организаций, где они усложнены, и еще больше организаций, вовсе не занимающихся определением стандартов качества данных). Находятся и скептики, ставящие под сомнение и целесообразность, и саму возможность управления качеством больших данных. Здравый смысл, однако, подсказывает, что управлять качеством больших данных можно и нужно. Достоверная аналитика немыслима на основе недостоверных данных. В проектах, предусматривающих сбор и анализ больших данных, судить о качестве вводных навскидку действительно невозможно, но именно поэтому и требуются особые усилия по оценке качества источников, иначе никакой уверенности в том, что результаты анализа соответствуют действительности, не будет. Для этого можно провести первичную экспертизу набора данных из

планируемого к подключению источника, необходимую для получения полного понимания качественной картины и соответствующего определения измеримых показателей и критериев качества, по которым и будут оцениваться последующие экземпляры набора данных из этого источника. В процессе первичной экспертной оценки качества данных также определяются и ценные метаданные, которые потребуются для любых работ по интеграции этих данных.

Самые зрелые в плане использования больших данных организации сканируют потенциальные источники вводных данных с помощью специальных средств инструментальной диагностики качества данных, которые позволяют понять, какая именно информация реально содержится в источнике. Самые передовые из таких инструментариев проверки качества поддерживают функциональность, которая дает организациям возможность проверять гипотезы и получать исчерпывающую информацию о том, что в действительности представляют собой предлагаемые источником данные. Примерами такой функциональности могут служить:

- **раскрытие** фактической структуры первоисточников и мест хранения данных, представленных в наборе;
- **классификация** представленных данных по стандартизованным типам и схемам;
- **профилирование**: насколько полны и как именно структурированы данные;
- **сопоставление и соотнесение** значений с данными из других источников/наборов.

Как и в случае DW/BI, в исследовании больших данных всегда есть место искушению отложить экспертизу качества до лучших времен. Однако без нее может оказаться затруднительным определение как содержания и смысла накапливающихся больших данных, так и определение связей и зависимостей между наборами данных. Интеграция так или иначе понадобится, а вероятность того, что данные из различных входящих потоков будут иметь идентичную структуру и состав элементов, близка к нулю. А это означает, что, например, коды и прочие связующие элементы данных в наборах от разных провайдеров почти наверняка совпадать не будут. Без выявления подобных рассогласованностей в рамках первичной экспертизы источников они так и останутся незамеченными, пока у аналитиков не возникнет потребности в интеграции или обобщении данных от различных поставщиков, — вот тогда-то и обнаружится их несовместимость и, как следствие, непригодность для решения поставленных задач.

6.6 Метрики

Измеримые показатели качества жизненно необходимы в любом процессе управления. Они позволяют не только выражать в цифрах качественные показатели проделанной работы, но и определять предельно допустимые отклонения наблюдаемых параметров от желаемых.

6.6.1 Технические метрики использования

Многие средства управления большими данными имеют продуманную функциональность административной аналитической отчетности, которая позволяет учитывать спрос на данные

различного типа непосредственно по результатам обработки пользовательских запросов к контенту. Метрики технического использования позволяют анализировать «горячие точки» (чаще других запрашиваемые данные) с целью оптимизации распределения данных и поддержания высокой производительности. Темпы роста спроса могут использоваться также для планирования развития вычислительных мощностей.

6.6.2 Метрики загрузки и сканирования

Метрики загрузки и сканирования отражают темпы освоения данных и интенсивность взаимодействия с пользовательским сообществом. С приобретением каждого нового источника данных метрики загрузки должны ожидаемо демонстрировать всплеск, а по завершении освоения данных из него — выравниваться. Оперативные данные, поступающие в потоковом режиме, могут обрабатываться в порядке очередности сетевыми службами, а могут накапливаться и обрабатываться партиями по расписанию; во втором случае ожидаемым эффектом будут цикличные всплески загрузки.

Слой или слои приложений, вероятно, служат оптимальным источником показателей использования данных, которые можно считывать из журналов исполняемых процессов. Мониторинг потребления или доступа можно вести и посредством регистрации статистики обращений к метаданным, тем более что такой подход позволяет анализировать еще и структуру и частотность запросов.

Метрики сканирования следует использовать в тех случаях, когда предусмотрена обработка внешних запросов, поступающих из-за пределов среды аналитической обработки данных. Средства администрирования должны обеспечивать учет таких взаимодействий в числе показателей здоровья служб управления большими данными.

6.6.3 Показатели эффективности и истории успешных внедрений

Для демонстрации ценности программы больших данных / науки о данных следует измерять показатели материальной отдачи от вложений в разработку решений и управления изменениями процессов. Подобные метрики могут включать количественные оценки дополнительной прибыли, стоимостного выражения полученных преимуществ, экономии за счет предотвращения или минимизации издержек, а также показатели сроков от инициации компонентов программы до их реализации и получения осязаемых результатов. Распространенные метрики включают:

- число и точность разработанных моделей и схем;
- дополнительные поступления и выгоды от реализации выявленных возможностей;
- экономию за счет снижения издержек и устранения выявленных угроз.

Иногда результаты аналитических изысканий похожи на увлекательные истории о том, как организации удалось переориентироваться и обрести второе дыхание, ухватившись за открытые новые возможности. Поэтому важнейшим показателем эффективности анализа больших данных может стать число новых проектов и инициатив, взятых на вооружение отделом маркетинга или утвержденных высшим руководством.

7. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Abate, Robert, Peter Aiken and Joseph Burke. *Integrating Enterprise Applications Utilizing A Services Based Architecture*. John Wiley and Sons, 1997. Print.

Arthur, Lisa. *Big Data Marketing: Engage Your Customers More Effectively and Drive Value*. Wiley, 2013. Print.

Barlow, Mike. *Real-Time Big Data Analytics: Emerging Architecture*. O'Reilly Media, 2013. Kindle.

Davenport, Thomas H. «Beyond the Black Box in analytics and Cognitive». *DataInformed* (website), 27 February, 2017, http://bit.ly/2sq8uG0. Web.

Davenport, Thomas H. *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Harvard Business Review Press, 2014. Print.

EMC Education Services, ed. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley, 2015. Print.

Executive Office of the President, National Science and Technology Council Committee on Technology. *Preparing for the Future of Artificial Intelligence*. October 2016, http://bit.ly/2j3XA4k

Inmon, W. H., and Dan Linstedt. *Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault*. 1st Edition. Morgan Kaufmann, 2014.

Jacobs, Adam. «Pathologies of Big Data». *AMCQUEU*, Volume 7, Issue 6. July 6, 2009, http://bit.ly/1vOqd80. Web.

Janssens, Jeroen. *Data Science at the Command Line: Facing the Future with Time-Tested Tools*. O'Reilly Media, 2014. Print.

Kitchin, Rob. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications Ltd, 2014. Print.

Krishnan, Krish. *Data Warehousing in the Age of Big Data*. Morgan Kaufmann, 2013. Print. The Morgan Kaufmann Series on Business Intelligence.

Lake, Peter and Robert Drake. *Information Systems Management in the Big Data Era*. Springer, 2015. Print. Advanced Information and Knowledge Processing.

Lake, Peter. *A Guide to Handling Data Using Hadoop: An exploration of Hadoop, Hive, Pig, Sqoop and Flume*. Peter Lake, 2015. Kindle. Advanced Information and Knowledge Processing.

Laney, Doug. «3D Data Management: Controlling Data Volume, Velocity, and Variety». *The Meta Group* [Gartner]. 6 February 2001, http://gtnr.it/1bKflKH

Loshin, David. *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools*, Techniques, NoSQL, and Graph. Morgan Kaufmann, 2013. Print.

Lublinsky, Boris, Kevin T. Smith, Alexey Yakubovich. *Professional Hadoop Solutions*. Wrox, 2013. Print.

Luisi, James. *Pragmatic Enterprise Architecture: Strategies to Transform Information Systems in the Era of Big Data*. Morgan Kaufmann, 2014. Print.

Marz, Nathan and James Warren. *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications, 2014. Print.

McCandless, David. *Information is Beautiful*. Collins, 2012.

Provost, Foster and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, 2013. Print.

Salminen, Joni and Valtteri Kaartemo, eds. *Big Data: Definitions, Business Logics, and Best Practices to Apply in Your Business*. Amazon Digital Services, Inc., 2014. Kindle. Books for Managers Book 2.

Sathi, Arvind. *Big Data Analytics: Disruptive Technologies for Changing the Game*. MC Press, 2013. Print.

Sawant, Nitin and Himanshu Shah. *Big Data Application Architecture Q&A: A Problem — Solution Approach*. Apress, 2013. Print. Expert's Voice in Big Data.

Slovic, Scott, Paul Slovic, eds. *Numbers and Nerves: Information, Emotion, and Meaning in a World of Data*. Oregon State University Press, 2015. Print.

Starbird, Michael. *Meaning from Data: Statistics Made Clear* (The Great Courses, Parts 1 and 2). The Teaching Company, 2006. Print.

Tufte, Edward R. *The Visual Display of Quantitative Information*. 2nd ed. Graphics Pr., 2001. Print.

van der Lans, Rick. Data Virtualization for Business Intelligence Systems: Revolutionizing Data Integration for Data Warehouses. Morgan Kaufmann, 2012. Print. The Morgan Kaufmann Series on Business Intelligence.

van Rijmenam, Mark. *Think Bigger: Developing a Successful Big Data Strategy for Your Business*. AMACOM, 2014. Print.

Оценка зрелости управления данными

1. ВВЕДЕНИЕ

Оценка зрелости возможностей (Capability Maturity Assessment, CMA) — подход к совершенствованию процессов, основанный на использовании специальной рамочной структуры — модели зрелости возможностей (Capability Maturity Model, CMM)[1], описывающей развитие процесса от бессистемной (ad hoc) организации до оптимального состояния. Концепция CMA начала формироваться в рамках работ по определению критериев оценки потенциальных поставщиков программного обеспечения, которые проводились Министерством обороны США. В середине 1980-х годов модель зрелости возможностей для программного обеспечения (Capability Maturity Model for Software) была опубликована Институтом программной инженерии (Software Engineering Institute) Университета Карнеги — Меллона. Хотя изначально CMM использовалась только применительно к разработке ПО, в дальнейшем аналогичные модели были созданы и для других областей деятельности, в частности для области управления данными.

Модели зрелости описываются в терминах продвижения по уровням зрелости, которым соответствуют определенные характеристики процесса. Научившись понимать эти характеристики, организация может начать повышать уровень зрелости процесса и внедрить план совершенствования его возможностей. Она может также измерять степень продвижения и сравнивать себя с конкурентами или партнерами на основе уровней зрелости модели. С переходом на очередной уровень процесс становится более стабильным, предсказуемым и надежным. Переход осуществляется, если обеспечиваются соответствующие данному уровню характеристики процесса. Прогресс носит строго ступенчатый характер: порядок следования уровней неизменен, и ни через одну ступень перескочить невозможно

Обычно выделяют следующие уровни[2].

[1] Под возможностью процесса (process capability) понимается характеристика его способности к достижению текущих или планируемых бизнес-целей. — *Примеч. науч. ред*

[2] См.: Select Business Solutions, «What is the Capability Maturity Model?» (http://bit.ly/IFMJI8, ссылка проверена 02/07/19).

- **Уровень 0.** Отсутствие возможностей.
- **Уровень 1.** Начальный (или бессистемный — ad hoc): успех зависит от компетенции отдельных сотрудников.
- **Уровень 2.** Повторяемый: присутствует минимальная дисциплина выполнения процессов.
- **Уровень 3.** Установленный: введены и используются стандарты.
- **Уровень 4.** Управляемый: обеспечена возможность измерения характеристик процессов и осуществляется их контроль.
- **Уровень 5.** Оптимизированный: обеспечена возможность измерения степени достижения целей процессов.

На каждом уровне описываются критерии оценки характеристик процессов. Например, модель зрелости может включать критерии, относящиеся к выполнению процессов, включая уровень автоматизации. Она может фокусироваться на политиках и механизмах контроля, а также на деталях процессов.

Такая оценка позволяет определить, что работает хорошо, что — недостаточно хорошо и где организация имеет пробелы (gaps). Основываясь на полученных данных, организация может разработать дорожную карту, нацеленную на:

- совершенствование по наиболее важным (обеспечивающим наибольшую выгоду) направлениям, относящимся к процессам, методам, ресурсам и средствам автоматизации;
- обеспечение возможностей, которые соответствуют бизнес-стратегии;
- поддержку процессов руководства (governance), которые необходимы для периодической оценки прогресса организации, основанной на характеристиках, заложенных в модель.

Модель оценки зрелости управления данными (Data Management Maturity Assessment, DMMA) может быть использована как для оценки процессов управления данными в целом, так и применительно к отдельным областям знаний по управлению данными или даже к отдельным процессам. Независимо от выбранного фокуса DMMA помогает ликвидировать разрыв между взглядами бизнеса и блока ИТ на текущее состояние и эффективность практик управления данными. Модель предоставляет общий язык для описания прогресса по отдельным областям знаний по управлению данными и предлагает поэтапный сценарий совершенствования, который может быть привязан к стратегическим приоритетам организации. Таким образом, подход DMMA можно использовать как для формулировки целей и задач организации, так и для измерения степени их достижения, а также для сравнения одной организации с другими организациями и с отраслевыми эталонными показателями.

Перед началом любого процесса DMMA организация должна оценить текущее состояние (базовый уровень — baseline) своих возможностей, ресурсов, целей и приоритетов. Некоторый уровень организационной зрелости требуется, чтобы провести первичную оценку, а также чтобы эффективно отреагировать на ее результаты, определив цели, утвердив дорожную карту и наладив мониторинг прогресса.

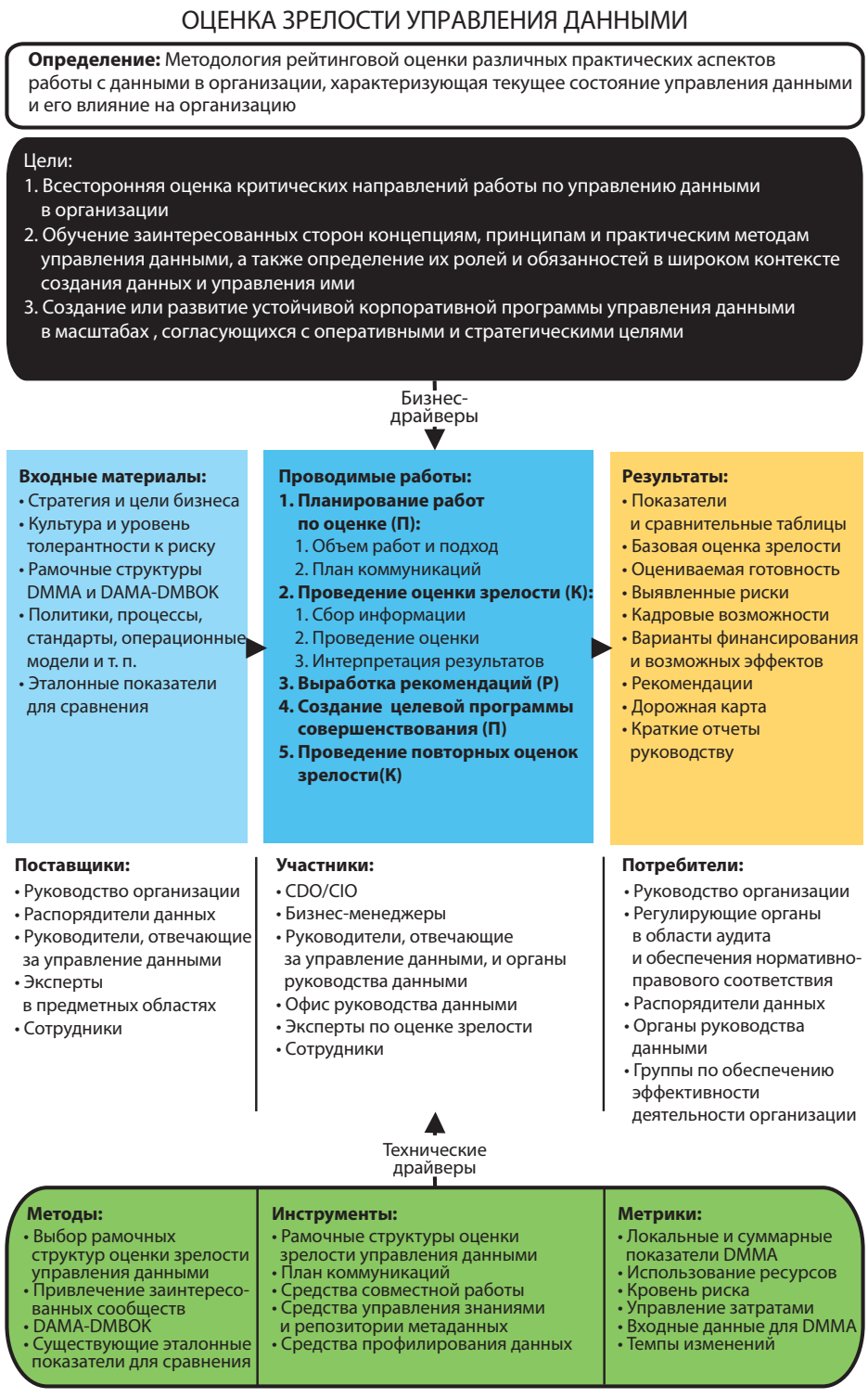


Рисунок 103. Контекстная диаграмма: оценка зрелости управления данными

1.1 Бизнес-драйверы

Организация проводит оценку зрелости возможностей по ряду причин:

- **Нормативно-правовые требования.** Законы и подзаконные акты, ведомственные инструкции, отраслевые регламенты или постановления надзорных органов часто в явном виде предписывают соблюдение определенного минимального уровня зрелости управления данными организации.
- **Реализация функции руководства данными.** Функция руководства данными предполагает проведение оценки зрелости в целях осуществления деятельности по планированию и обеспечению соответствия требованиям.
- **Готовность организации к совершенствованию процессов.** Осознание того факта, что нужно что-то менять в сложившейся практике для выхода на новый качественный уровень, естественным образом приводит к возникновению потребности в комплексной оценке текущего состояния. Например, руководство понимает, что в современных условиях не обойтись без управления основными данными, и инициирует оценку готовности организации к внедрению процессов и инструментов MDM.
- **Организационные изменения.** Любые реорганизации, особенно слияния, бывают сопряжены с серьезными трудностями в сфере управления данными. Вводные для решения встающих задач проще всего получить, придерживаясь модели DMMA.
- **Новые технологии.** Технологический прогресс открывает всё новые возможности и способы управления данными и их применения, и организация не хочет их упускать.
- **Проблемы с управлением данными.** При возникновении систематических проблем в области управления данными или обеспечения качества данных комплексная экспертиза текущего состояния — обязательный первый шаг по пути поиска и устранения их причин.

1.2 Цели и принципы

Главная цель оценки возможностей процессов управления данными организации — оценить текущее состояние критически важных участков работ и выявить недостатки, нуждающиеся в устранении. В процессе оценочной экспертизы организация получает оценки по различным шкалам измерения зрелости; низкие оценки и есть прямое указание на слабые стороны. Ориентируясь на результаты проведенной экспертизы, организации проще выявлять, ранжировать и реализовывать возможности для усовершенствований.

Помимо достижения первичных целей, DMMA также может положительно влиять на организационную культуру, помогая:

- разъяснять всем сторонам концепции, принципы и методы управления данными;
- уточнять роли и обязанности различных лиц по отношению к данным организации;
- подчеркивать важность отношения к данным как к ценному активу;

- добиваться всеобщего понимания необходимости работ по управлению данными;
- налаживать сотрудничество на благо эффективного распоряжения данными.

По результатам проведенной экспертизы организация может спланировать необходимые усовершенствования своей программы управления данными в привязке к рабочим и стратегическим планам. Обычно управление данными, будучи пущено на самотек, приводит к несогласованности и разрозненности программ управления данными, формирующихся в различных уголках организации. Редко где первичная инициатива носит централизованный характер и приводит сразу же к формированию общеорганизационного ви́дения данных. Методология DMMA помогает организации вооружиться всем необходимым для выработки связной картины и — на ее основе — общеорганизационной стратегии. Кроме того, DMMA помогает организации уяснять приоритеты, выкристаллизовывать объективные задачи и разрабатывать комплексный план усовершенствований.

1.3 Основные понятия и концепции

1.3.1 Оценки уровня зрелости и их характеристики

В стандартных CMM обычно насчитывается пять-шесть уровней зрелости с характеристиками в диапазоне от «нулевого» или отсутствующего управления до высокоэффективного или полностью оптимизированного. Рисунок 104 иллюстрирует пример классификации уровней.

Ниже приведено описание самой общей модели с описаниями уровней зрелости управления данными. При детальной экспертизе оценки обязательно должны ставиться раздельно по областям знаний по управлению данными (DM) и категориям и подкатегориям направлений DM в каждой области (стратегия, политика, правила, стандарты, определение ролей и т. д. и т. п.).

- **Уровень 0. Отсутствие возможностей**: нулевая организация управления данными; полная неуправляемость данных на практике или пущенные на самотек чисто формальные процессы сбора данных для отчетности. В реальном мире организации с нулевым уровнем DM — крайняя редкость. В CMM этот уровень обычно определяется исключительно в качестве базовой точки.
- **Уровень 1. Начальный** (или бессистемный — ad hoc): с использованием ограниченного набора инструментально-технических средств при отсутствующем или слабом централизованном распоряжении данными. Грамотно обращаться с данными умеют в лучшем случае несколько специалистов, от которых всецело зависят результаты. Роли и обязанности определяются исключительно в рамках параллельных вертикалей организационно-функциональных подразделений. В каждом подразделении есть свой хозяин данных, которые собирает/получает/генерирует данные и передает их наверх в автономном от остальных вертикалей режиме. Механизмы управления, если даже и существуют, применяются непоследовательно и несогласованно. Платформенные решения по DM не используются или применяются в ограниченных

пределах. Проблемы с качеством данных носят систематический и повсеместный характер, и никто не озабочен их выявлением, не говоря уже о решении. Техническая поддержка инфраструктуры DM реализована на уровне бизнес-подразделений.

Дополнительные критерии оценки на этом уровне могут включать наличие хоть каких-то контрольных механизмов — например, журналов ошибок или регистрации проблем с качеством данных.

Рисунок 104. Пример модели оценки зрелости управления данными

- **Уровень 2. Повторяемый**: появляются согласованные наборы инструментов управления данными, назначаются ответственные за исполнение различных процессов. На втором уровне организация начинает использовать и централизованные средства распоряжения данными и административного надзора. Активно определяются ролевые функции, процессы обработки данных перестают всецело зависеть лично от экспертов. Проявляются общеорганизационная обеспокоенность качеством данных и понимание последствий проблем с данными. Зарождается понимание необходимости управления основными и справочными данными.

Критерием оценки может дополнительно служить наличие формальных определений ролей в должностных инструкциях, технологической документации процессов и инструментов DM.

- **Уровень 3. Установленный**: появляются функции систематического управления данными. Именно на третьем уровне мы становимся свидетелями внедрения и институционального закрепления масштабируемых процессов DM и представления об эффективном

высокоуровневом распоряжении данными как о залоге успешного развития организации. Отличительные характеристики этого уровня зрелости включают репликацию данных в масштабах организации с обязательными механизмами контроля согласованности и, в целом, повышение качества всей совокупности данных, а также скоординированное определение политики и правил DM, а также централизованный контроль их соблюдения. Формализация определений процессов приводит к устойчивому сокращению доли ручной обработки за счет внедрения средств автоматизации, а это, наряду с централизацией проектирования, в свою очередь делает результаты всех процессов более предсказуемыми.

В число критериев оценки могут быть включены наличие политик, правил или регламентов управления данными, использование масштабируемых процессов и решений, согласованность моделей данных и средств управления системами.

- **Уровень 4. Управляемый**: институциональные знания, накопленные в процессе прохождения уровней 1–3, позволяют организации уверенно прогнозировать результаты планируемых новых проектов, решать поставленные задачи и на фоне этого приступить к управлению рисками, обусловленными данными. В сферу DM включаются текущие метрики производительности систем и оперативного контроля качества данных. К характерным признакам четвертого уровня зрелости относятся стандартизация средств DM на всех уровнях — от рабочих мест до инфраструктуры предприятия; полностью сформировавшаяся система централизованного планирования и осуществления административно-надзорных функций распоряжения данными. Достижение четвертого уровня характеризуется ощутимым подъемом общего уровня качества данных и развитием таких общеорганизационных функционалов, как комплексный аудит всех данных во всех процессах.

Критерии оценки могут также включать показатели успешности проектов, технико-эксплуатационные характеристики систем и текущие показатели качества данных.

- **Уровень 5. Оптимизированный**: после оптимизации всех практических аспектов DM результаты процессов становятся максимально предсказуемыми благодаря предельной автоматизации и эффективному управлению технологическими изменениями. Достигнув высшего уровня зрелости, организации перефокусируют внимание на непрерывное совершенствование всего комплекса систем и процессов DM. На пятом уровне становится возможным прозрачный просмотр данных в полном срезе процессов, что позволяет минимизировать избыточные и повторяющиеся данные и процессы. Метрики качества данных и процессов, как и механизмы управления ими, понятны и доступны всем сотрудникам.

Оцениваться могут также документы по управлению изменениями, методологии измерения показателей и совершенствования процессов.

1.3.2 Оценочные критерии

Каждому уровню зрелости DM соответствуют наборы оценочных критериев процессов, характерных для соответствующего уровня и специфичным образом распределенных по направлениям оценки. Например, если оценивается зрелость функции моделирования данных, на уровне 1 могут ставиться вопросы о существовании каких-либо моделей как таковых и понимании назначения моделирования данных и систем, которые должны охватываться этими моделями; на уровне 2 — вопросы о качестве выбранного подхода к определению моделей данных в архитектурной среде предприятия; на уровне 3 — о степени реализации выбранного подхода; на уровне 4 — об эффективности контроля соблюдении стандартов моделирования; на уровне 5 — о наличии процессов непрерывного совершенствования моделей и практик моделирования (см. главу 5).

На любом уровне и по любому направлению текущее состояние оценивается по единой шкале, например: 0 — не начато; 1 — планируется; 2 — внедряется; 3 — функционирует; 4 — полностью эффективно. По этим показателям отслеживается прогресс на текущем уровне и близость к состоянию готовности к переходу на следующий качественный уровень. Полученные оценки (в баллах) по различным параметрам могут обобщаться в сводные показатели или отображаться визуально с целью лучшего понимания разрыва между текущим и желаемым состояниями.

При оценке с использованием модели, совместимой по классификации с областями знания об управлении данными DAMA-DMBOK, критерии можно и нужно формулировать в привязке к категориям, представленным на контекстной диаграмме.

- **Направление.** Насколько проект или процесс соответствует требуемому составу работ по заявленному направлению? Вписывается ли в общий контекст? Определены ли критерии действенности и эффективности исполнения? Насколько хорошо определен и исполняется план работ? Соответствуют ли результаты на выходе лучшим образцам?
- **Технические решения и инструментальные средства.** Достаточно ли автоматизированы процессы на данном направлении? Согласованы ли наборы инструментов с используемыми в других областях? Проводится ли обучение работе с инструментами всех сотрудников, которым полагается их использовать в силу их функциональной роли или должностных обязанностей? Всегда ли доступны все нужные для работы инструментальные средства? Оптимально ли отконфигурированы под достижение наилучших результатов? В какой мере долгосрочное ИТ-планирование учитывает потребности, которые возникнут по достижении в будущем желаемого состояния?
- **Стандарты.** В какой степени выполнение работ согласовано с общепринятым набором стандартов? Хорошо ли задокументированы стандарты? Поддерживается ли соблюдение стандартов административными средствами обеспечения их соблюдения и управления изменениями?
- **Кадровые ресурсы.** Укомплектован ли штат организации всеми квалифицированными специалистами, требующимися для реализации проекта? Какие навыки, подготовка, знания и опыт требуются для выполнения работ? Хорошо ли определены все роли и обязанности?

Рисунок 105 иллюстрирует возможное визуальное представление результатов экспертной оценки зрелости управления данными по методологии DMMA. Внешний контур задает необходимые для обеспечения конкурентоспособности организации оценки зрелости по всем функциональным областям (руководство данными, архитектура и т. д.), а внутренний отражает фактическое положение дел по этим направлениям, выявленное по результатам экспертизы. Области с наибольшим разрывом между желаемым и текущим состояниями являются источником наибольших рисков для организации. Ознакомление с таким отчетом весьма полезно для определения приоритетов, а периодические повторные экспертизы могут использоваться для мониторинга достигнутого прогресса.

Рисунок 105. Пример визуального представления результатов оценки зрелости управления данными

1.3.3 Существующие рамочные структуры оценки зрелости управления данными

Любая методика оценки зрелости управления данными предусматривает сегментацию на отдельные объекты управления. Акцент внимания может смещаться в сторону того или иного направления, а состав и содержание параметров оценки — варьироваться в зависимости от ориентации на общее или узкое/отраслевое применение. Однако большинство методик так или иначе рассматривают вопросы, вполне укладывающиеся в предметные области рамочной структуры DAMA-DMBOK. Ниже описаны лишь избранные примеры, иллюстрирующие широту спектра

применения CMM к управлению данными. Многие поставщики комплексных решений к тому же разрабатывают и собственные модели оценки, поэтому организациям рекомендуется произвести сравнительный анализ предлагаемых на рынке методологий, прежде чем выбирать поставщика или приступать к конструированию собственной рамочной структуры оценки зрелости управления данными[1].

1.3.3.1 МОДЕЛЬ CMMI-DMM

Институт моделирования зрелости возможностей — CMMI (Capability Maturity Model Institute) — разработал и предлагает собственную модель оценки зрелости управления данными (DMM) по критериям, отнесенным к следующим областям:

- стратегия управления данными;
- руководство данными;
- обеспечение качества данных;
- платформа и архитектура;
- оперативное управление данными;
- поддерживающие процессы.

В рамках каждой области эта модель определяет процессы и подпроцессы, подлежащие оценке. Например, раздел *обеспечение качества данных* включает параметры оценки стратегии обеспечения качества данных, процедур проверки, профилирования и очистки данных. Модель позволяет учитывать связи между областями управления данными, что дает возможность согласованно учитывать интересы различных сторон и участников[2].

1.3.3.2 МОДЕЛЬ EDM COUNCIL-DCAM

Совет по управлению корпоративными данными (Enterprise Data Management Council) — международная некоммерческая организация по отстаиванию отраслевых интересов в сфере финансовых услуг со штаб-квартирой в США, занимающаяся стандартизацией данных на рынках финансовых услуг. Разрабатываемая EDM Council® по принципу совместного вклада и выработки консенсуса модель оценки возможностей по управлению данными (Data Management Capability Assessment Model, DCAM) описывает 37 категорий и 115 подкатегорий функциональности и устойчивости программ управления данными. Подсчет баллов ведется с учетом показателей

[1] Дополнительную информацию о существующих моделях оценки функциональной зрелости управления данными (DM CMM) см.: Alan McSweeney, *Review of Data Management Maturity Models*, SlideShare.net (http://bit.ly/2spTCY9); Jeff Gorball, *Introduction to Data Management Maturity Models*, SlideShare.net (2016-08-01). Правда, в первой из двух указанных презентаций (McSweeney) модель DAMA-DMBOK также ошибочно классифицирована как одна из моделей оценки зрелости, хотя таковой не является ни концептуально, ни структурно, а лишь содержит рекомендации относительно применимости методологий DM CMM в различных ситуациях.

[2] http://bit.ly/1Vev9xx

вовлечения всех заинтересованных сторон, формализации процесса, наличия подтверждающих достижения и способности документов, и т. д. и т. п.[1]

1.3.3.3 МОДЕЛЬ СОВЕТА ПО УПРАВЛЕНИЮ ДАННЫМИ IBM

Модель зрелости Совета по управлению данными IBM стала результатом согласования предложений, представленных 55 организациями — членами совета. Участники проекта сформулировали наборы наблюдаемых и желаемых схем поведения организаций, озабоченных оценкой имеющихся и разработкой перспективных программ распоряжения данными. Основное назначение модели IBM — помощь организациям в обеспечении согласованности моделей и качества данных путем последовательного применения проверенных технологий бизнес-управления в комплексе с передовыми методами совместной работы. Модель строится вокруг четырех ключевых категорий.

- **Результаты**: минимизация риска, соблюдение установленных требований, создание добавленной стоимости.
- **Необходимые условия (предпосылки)**: организованность, сознательность, политика, надзор.
- **Ключевые дисциплины**: управление качеством и жизненным циклом данных, обеспечение ИБ и защиты данных, и т. п.
- **Вспомогательные/технические дисциплины**: архитектура и классификация данных, управление метаданными, аудит, регистрация и отчетность.

Модель IBM предлагается как в формате комплексной модели оценки зрелости, так и в формате опросных листов, позволяющих оценить уровни зрелости управления данными организации по различным направлениям[2].

1.3.3.4 МОДЕЛЬ ЗРЕЛОСТИ УПРАВЛЕНИЯ ДАННЫМИ СТЭНФОРДСКОГО УНИВЕРСИТЕТА

Стэнфордская модель зрелости управления данными разработана для сугубо внутренних нужд Стэнфордского университета и не претендует на роль отраслевого стандарта, однако заслуживает упоминания, поскольку может служить образцом строгого академического подхода к определению стандартов оценки качества измерений. Первоочередное внимание в этой модели уделяется контролю, а не управлению, но базисные показатели оценки выглядят очень внушительно. Модель проводит четкое разграничение между фундаментальными (осознание, формализация, метаданные) и проектными (обслуживание, обеспечение качества, основные данные) компонентами. В рамках каждого компонента формулируются движущие стимулы для участников, правила и функциональные возможности. И лишь после этого четко определяются критерии оценки зрелости по каждому из этих компонентов. Кроме того, предлагаются и количественные показатели оценки, соответствующие каждому критерию[3].

[1] http://bit.ly/2sqaSga

[2] https://ibm.co/2sRfBIn (ссылка проверена 04/07/19).

[3] См.: http://stanford.io/2sBR5bZ и http://stanford.io/2rVPyM2 (ссылки проверены 04/07/19).

1.3.3.5 МОДЕЛЬ GARTNER'S ENTERPRISE INFORMATION MANAGEMENT MATURITY MODEL

Консалтинговая компания Gartner опубликовала модель зрелости корпоративного управления информацией (Enterprise Information Management, EIM), которая устанавливает критерии оценки ви́дения, стратегии, метрик, управления, ролей и обязанностей, жизненного цикла и инфраструктуры.

2. ПРОВОДИМЫЕ РАБОТЫ

Оценки зрелости управления данных должны проводиться на плановой основе. При планировании не забудьте отвести достаточно времени на подготовку материалов и оценку результатов: это поспособствует получению качественных и осмысленных результатов, которые можно будет принять за основу при последующей выработке практических мер. Сами экспертизы должны проводиться в сжатые сроки, установленные планом-графиком работ. Помните, что цель оценки зрелости — получить снимок текущего состояния управления данными со всеми его сильными и слабыми сторонами, а не заниматься устранением выявляющихся по ходу оценки проблем.

Оценивание производится по результатам углубленных собеседований с участниками бизнес-процессов, специалистами по управлению данными и ИТ-системам. Цель — достигнуть консенсуса относительно текущего функционального состояния каждого из компонентов DM, подкрепленного объективными, задокументированными свидетельствами, например выявленными артефактами (У вас имеется резервная копия этой БД? — да/нет) и/или показаниями (Эта система регистрации разрабатывалась с учетом возможности ее многоцелевого использования? — да/нет).

Экспертизы допускают масштабирование в зависимости от потребностей организации. Однако в случае урезания части вопросов или изменения их формулировок вы рискуете снизить уровень тщательности проверки или исказить смысл, вкладывавшийся в вопросы в рамках исходной модели. В любом случае, редактируя и приспосабливая модель оценки зрелости DM под нужды своей организации, следите за сохранением ее целостности и непротиворечивости.

2.1 Планирование работ по оценке

Нужно определить общий подход, выбрать модель, согласовать план мероприятий со всеми заинтересованными сторонами — и оставаться с ними на связи до завершения экспертизы, подпитывая тем самым их заинтересованность. Тем более что проведение самой оценки зрелости — процесс, по сути, интерактивный, поскольку включает сбор и оценку вводных данных, сообщение результатов и согласование рекомендаций и планов действий.

2.1.1 Определение задач

Любая организация, решившая озаботиться оценкой зрелости управления данными, самим этим решением демонстрирует ненулевой уровень готовности к совершенствованию практики DM.

В большинстве случаев у такой организации и бизнес-стимулы на момент проведения экспертизы определены вполне четко, но всё же их следует дополнительно уточнить, сформулировав в виде объективных задач, на которые и будет ориентироваться оценка. Главное, чтобы задачи оценки зрелости DM четко понимались высшим руководством и менеджерами направлений бизнеса, поскольку от них потребуется помощь в согласовании и выстраивании экспертизы согласно стратегическим направлениям развития организации.

В задачи оценки входит также выработка критериев выбора модели, приоритетных для проведения экспертизы участков работы, и определение лиц, непосредственно отвечающих за предоставление вводных данных.

2.1.2 Выбор рамочной структуры

Как отмечалось в разделе 1.3.3, существующие концептуальные модели различаются преимущественно расстановкой приоритетов и фокусировкой на различных аспектах управления данными. Внимательно проанализируйте эти модели в контексте предполагаемого текущего состояния и задач оценки на предмет выбора наиболее информативной, осмысленной и всесторонне полезной для вашей организации. Области фокусировки модели экспертной оценки можно будет и подкорректировать исходя из специфики вашей организации.

От выбора модели будет во многом зависеть и порядок практического проведения оценочных работ. Важно, чтобы команда экспертов имела навыки работы с выбранной моделью и знала методологию, лежащую в ее основе.

2.1.3 Определение объема и содержания работ

Большинство методологий оценки DMM предусматривают экспертизу в масштабах всего предприятия. Однако такой подход бывает труднореализуемым или избыточным. Для первичной оценки достаточно определить посильный объем работ: например, запланировать экспертизу единственной области бизнеса или реализуемой организацией программы. Выбранные области должны представлять осмысленный срез организации в целом, содержащий подмножество данных и состав участников, позволяющий рассчитывать на благотворное влияние оптимизации DM в выбранной области на общеорганизационную ситуацию в сфере управления информационными ресурсами. В рамках поэтапного внедрения можно предусмотреть воспроизведение модели экспертной оценки зрелости DM, отработанной на первичном участке работ, в других частях организации. При выборе между локальным и общеорганизационным подходами или поиске оптимального компромиссного варианта учитывайте следующие соображения.

- **Локальные оценки** позволяют глубже исследовать детали и к тому же обычно занимают меньше времени просто по причине меньших объемов исследуемых данных и процессов. Для проведения пробной точечной экспертизы выберите зарегулированную функцию, например финансовую отчетность, если речь идет о публичной компании. Многие вводные, роли, инструменты и потребители могут оставаться за рамками оцениваемых функций, что весьма

осложняет планирование и проведение оценки. Хорошо спланированный комплекс локальных оценок часто позволяет получать средневзвешенные показатели, достаточно четко описывающие ситуацию в целом по предприятию, поскольку многие данные так или иначе находятся в совместном пользовании.

- **Оценка в масштабах организации** охватывает широкий и не всегда связный круг данных и процессов, имеющихся в разных частях организации. Программа DMMA предприятия может вырасти из локального проекта оценки зрелости управления данными, а может быть реализована в рамках отдельной масштабной инициативы. Например, организация может запланировать и осуществить переход к оценке работы с данными различных функций (исследование и разработка, производство, финансы и т. п.) по единому для всех набору критериев. В рамках общеорганизационной многоуровневой модели проще учесть все вводные, роли, инструменты и потребителей данных.

2.1.4 Определение подхода к обеспечению взаимодействия

Проводя оценку DMM, организация должна строго придерживаться методологических рекомендаций выбранной модели. Для сбора информации могут проводиться рабочие семинары, собеседования, опросы, выемка и исследование документов, и т. п. Постарайтесь выбирать методы, хорошо сочетающиеся с организационной культурой, отнимающие минимум времени у участников и позволяющие быстро завершать этап оценки и приступать к определению плана действий, пока участники процесса еще помнят детали процессов.

Во всех без исключения случаях требуется формализованная обратная связь, позволяющая участникам выставлять рейтинговые оценки текущей ситуации по всем выбранным критериям. Во многих случаях экспертиза также будет включать физическую инспекцию на местах со сбором, изучением и оценкой документов, свидетельств, прочих артефактов и доказательств.

При затягивании сроков завершения экспертизы заинтересованные стороны могут разочароваться в программе управления данными, утратить всякий энтузиазм и даже начать чинить препятствия и противиться позитивным изменениям. Разумный совет: не перегружайте целевую аудиторию деталями и исчерпывающим анализом, а делайте упор на основные моменты и подчеркивайте объективность и обоснованность суждений, подкрепленных компетенцией и опытом ведущих экспертов. Все методологии оценки DMM включают измеримые критерии и соответствующие им методики усовершенствований. По совокупности всё это позволяет получать как целостную картину текущего состояния программы управления данными, так и представление о зрелости ее отдельных компонентов.

2.1.5 Планирование коммуникаций

Взаимопонимание — неотъемлемый атрибут успеха экспертной оценки и мер, выработанных по ее результатам. Следует доносить смысл проделываемой работы и результаты заключений до всех участников и заинтересованных сторон. Ведь выводы экспертизы могут напрямую сказаться на работе многих людей, пусть даже и опосредованно (например, через изменение методологий

и реорганизацию), — поэтому важно четко информировать сотрудников о целях и задачах, назначении и порядке процессов оценки, конкретных ожиданиях от различных групп и отдельных сотрудников. Нужно донести до понимания участников, во-первых, смысл модели оценки, а во-вторых — практическое назначение ее результатов.

Прежде чем приступать к экспертной оценке, обязательно разъясните всем заинтересованным сторонам, что именно будет требоваться от них в процессе изучения ситуации. Информация должна включать описание следующих моментов:

- назначение экспертизы зрелости управления данными;
- порядок проведения DMMA;
- степень участия, вклад и роли сотрудников (по ситуации);
- график мероприятий.

В ходе любого мероприятия (например, проведения фокусной группы) придерживайтесь заранее выработанной четкой повестки, которая должна включать и ответы на возникающие у аудитории вопросы, и заготовки самих ответов. Постоянно напоминайте участникам о целях и задачах. Всякий раз благодарите их за внесенный вклад и описывайте следующие этапы.

Оценивайте вероятность успеха запланированного подхода применительно к целевому направлению бизнеса, исходя из выявляющихся в процессе собеседований и фокусных групп факторов, включая неприятие / готовность к сотрудничеству, нежелание раскрывать истинное положение дел из опасения дисциплинарных последствий для себя или юридических осложнений у организации из-за выявленных нарушений (в том случае, если оценка проводится извне); не исключено замалчивание из боязни прослыть нелояльным к своей организации у начальства или кадровиков.

Наконец, план коммуникаций должен обязательно включать график предоставления отчетов с результатами экспертной оценки и рекомендаций для работников всех уровней, включая сводный отчет и краткие сводки для высших руководителей.

2.2 Проведение оценки зрелости

2.2.1 Сбор информации

Следующий этап — сбор вводных данных, предусмотренных моделью экспертной оценки. Как минимум, должны быть собраны все данные, необходимые для формального расчета рейтинговых оценок по всем критериям модели. Кроме того, они могут быть дополнены информацией, полученной по результатам интервью и фокусных групп, анализа систем и проектной документации, исследований источников и маршрутов движения данных, цепочек электронных писем, процедурных руководств, стандартов, политик и правил, файлохранилищ, административно-деловых процедур и порядков согласования решений, всевозможных рабочих продуктов, репозиториев метаданных, архитектуры данных и эталонной архитектуры интеграции данных, шаблонов и форм.

2.2.2 Проведение оценки

Общеорганизационные рейтинги обычно рассчитываются и интерпретируются по многофазной схеме. Участники непременно будут иметь разные мнения относительно положения вещей и оценки ситуации по многим изучаемым темам. Потребуются дискуссии и рациональные обоснования тех или иных оценок, прежде чем удастся прийти к согласию относительно каждого рейтингового показателя. Согласованное мнение участников служит основанием лишь для исходных оценок, которые затем дорабатываются и уточняются командой экспертов в ходе изучения артефактов. Конечная цель — выработка консенсуса относительно текущего состояния DM по всем направлениям. Полученная картина должна подкрепляться доказательной базой (то есть документальными свидетельствами соответствия практики ее описаниям). Если у заинтересованных сторон не сложится единого представления относительно текущего состояния, то ни о каком консенсусе относительно оптимальных путей совершенствования работы организации и речи быть не может.

Процесс уточнения и согласования оценок текущих параметров обычно осуществляется по следующему алгоритму:

- сравнение результатов с критериями выбранного рейтингового метода с выставлением предварительной оценки каждому продукту или виду работы;
- документирование оснований для выставления каждой оценки;
- разбор и анализ предварительных результатов совместно с участниками с целью выработки консенсуса относительно окончательной рейтинговой оценки по каждой области; при необходимости — подсчет общего рейтинга как средневзвешенного показателя всех оценок с весами, пропорциональными значимости критерия;
- документирование интерпретации рейтинга с использованием формулировок критериев модели и комментариями экспертов-оценщиков;
- разработка визуальных представлений, иллюстрирующих результаты экспертной оценки.

2.3 Интерпретация результатов

Интерпретация результатов заключается в выявлении возможностей для совершенствования в привязке к организационной стратегии и в выработке практических рекомендаций по использованию этих возможностей. Иными словами, интерпретация заключается в определении следующих шагов на пути к целевому состоянию. То есть по завершении оценки текущего состояния организации нужно прежде всего переосмыслить, какого желаемого состояния она надеется достичь в сфере управления данными. Затраты времени и усилий на достижение цели будет варьироваться в зависимости от исходной точки, амбициозности цели, культуры организации и стимулов к изменениям.

Представляя результаты проведенной экспертизы, начните с объяснения смысла рассчитанных рейтингов и их значения для организации. Для доходчивости лучше разъяснять рейтинги на примерах, контекстуально близких организации, с учетом ее культурной специфики, бизнес-целей и стимулов — например, повышения степени удовлетворенности клиентов или объемов

продаж. Проиллюстрируйте также связь между функциональными возможностями организации на текущий момент, бизнес-процессами и стратегиями, которые могут ими поддерживаться, и преимуществами, которые получит организация в результате расширения возможностей по достижении целевого состояния.

2.3.1 Подготовка отчета с результатами оценки

По результатам оценки зрелости управления данными составляется отчет, включающий следующие разделы с описаниями и экспертными заключениями.

- Бизнес-драйверы, послужившие стимулами к проведению экспертизы.
- Общие результаты обследования.
- Рейтинги по направлениям с указанием степени тяжести недоработок.
- Рекомендуемые подходы по устранению недоработок.
- Сильные стороны организации на момент наблюдения.
- Прогрессирующие риски.
- Доступные варианты инвестиций и их результаты.
- Административный контроль и измеримые показатели прогресса.
- Анализ имеющихся ресурсов и потенциальных будущих потребностей.
- Артефакты, которые могут разово или многократно использоваться организацией.

Отчет с результатами оценки становится источником исходных данных для развития программы управления данными — либо в целом, либо в части предметных областей управления данными, которые оценивались. Отталкиваясь от него, организация сможет развивать, двигать вперед или совершенствовать стратегию управления данными. Стратегия должна включать инициативы по ускорению продвижения к бизнес-целям за счет совершенствования высокоуровневого управления процессами и стандартами.

2.3.2 Подготовка кратких отчетов руководству

Экспертная группа должна периодически подготавливать и представлять руководству краткие сводки с основными результатами заключений о выявленных сильных и слабых сторонах, недостатках и возможностях, а также со сжатыми рекомендациями, на которые руководство сможет ориентироваться, принимая решения о выборе приоритетных инициатив и постановке целей и сроков их реализации. Составлять такие докладные нужно таким образом, чтобы они лаконично и доходчиво разъясняли каждой целевой группе руководителей негативные последствия сохранения недоработок и выгоды от их устранения.

Амбициозные начальники часто замахиваются сразу на нечто более грандиозное, нежели то, что содержится в рекомендациях экспертов, — иначе говоря, нацеливаются на скачки через ступени лестницы модели зрелости DM. Нацеленность руководства на наивысший уровень зрелости должна находить отражение в анализе последствий, включаемом в рекомендации, чтобы донести

до понимания руководителей оборотные стороны и издержки подобного чрезмерного ускорения хода событий и заставить задуматься о балансировке рисков и выгод.

2.4 Создание целевой программы совершенствования управления данными

Экспертная оценка зрелости управления данными по одной из методологий DMM — не самоцель. Она призвана служить инструментом прямого воздействия на стратегию развития ИТ-систем, вырабатываемую высшим руководством, равно как и на программу и стратегию работы с данными. Рекомендации DMMA должны расцениваться как обязательные для практической реализации. Но для этого и сами рекомендации должны составляться с учетом ресурсных возможностей и функциональных потребностей организации. Если такой баланс будет достигнут, оценка зрелости DM может сделаться мощным средством определения приоритетов и распределения ресурсов организации совместными усилиями специалистов по ИТ, DM и бизнес-руководства.

2.4.1 Определение состава работ и создание дорожной карты

Рейтинговые оценки DMMA высвечивают компоненты DM, требующие особого внимания со стороны руководства и специалистов. Поначалу каждый рейтинг, вероятно, будет использоваться как самостоятельный показатель способности организации успешно решать узкий круг задач, к которому относится соответствующая оценка. Однако в скором времени различные рейтинговые оценки оперативно увязываются в согласованные текущие показатели процессов, в частности требующих усовершенствования (например: «Целевое значение показателя N должно быть не ниже N = n, потому что без этого мы не сможем обеспечить требуемый уровень Z = z в смежном процессе). Если показатели модели экспертной оценки включаются в текущие измерения, критерии зрелости управления данными не только становятся инструментами самонаведения организации на высшие уровни зрелости, но и не позволяют забывать о самой необходимости постоянных усовершенствований.

Результаты первичной экспертизы зрелости управления данными должны быть полными и детализированными в той мере, в которой это необходимо для информационной поддержки многолетней программы совершенствования, включая инициативы как по развитию функциональных возможностей и вычислительных мощностей, так и по внедрению передового опыта. Поскольку реальные изменения обычно привносятся в жизнь организаций через проекты, нужно обеспечить рычаги влияния на новые проекты, с тем чтобы они изначально планировались с учетом рекомендаций по управлению данными. Для этого в типовую схему или шаблон дорожной карты новых проектов следует включить:

- строго упорядоченные последовательности действий, направленных на совершенствование конкретных функций управления данными;
- планы-графики реализации мер по совершенствованию управления данными;
- ожидаемые результаты, выражающиеся в повышении рейтингов DMMA по завершении реализации проектов;

- мероприятия по внедрению и поэтапному совершенствованию механизмов надзора, включая планы-графики реализации.

Такие дорожные карты будут задавать промежуточные цели и темпы изменений в рамках приоритетных потоков работ, а соответствующие им согласованные показатели — помогать отслеживать прогресс.

2.5 Проведение повторных оценок зрелости

Повторные оценки проводятся с установленной периодичностью. Содержание повторных экспертиз включает:

- оценку базовых показателей по методологии первичной экспертизы;
- определение параметров, нуждающихся в уточнении или переопределении в целях повторной оценки, включая состав и объем организационных составляющих;
- повторную полную экспертизу по выбранной модели DMM, если это требуется согласно установленному плану-графику;
- отслеживание текущих тенденций относительно исходных (базовых) и промежуточных показателей;
- выработку дополнительных и обновленных рекомендаций по результатам переоценки текущего состояния.

Повторные экспертизы бывают полезны с точки зрения переключения фокуса внимания сотрудников на вновь выявленные проблемы или наиболее актуальные направления. Измеримые показатели прогресса — лучшее средство подпитки энтузиазма в масштабах организации. Дополнительный плюс периодического проведения повторных комплексных экспертиз — своевременное выявление изменений нормативно-правового режима, внутренней или внешней политики, а также инноваций, требующих коренного пересмотра подхода к распоряжению и стратегий управления данными.

3. ИНСТРУМЕНТЫ

- **Методика оценки зрелости управления данными**: выбранная рамочная модель DMM, собственно, и является главным инструментом экспертов.
- **План коммуникаций**, включая модель привлечения к участию всех фигурантов и интересантов, типы информации для сбора и распространения, график публикации или передачи руководству результатов оценки и экспертных заключений.
- **Средства обеспечения совместной работы** позволяют оперативно доводить до сведения всех участников полученные оценки, результаты и заключения экспертизы. Кроме того, переписка по e-mail, заполненные формы и шаблоны, документы, созданные в рамках стандартных

процедур совместного проектирования, обмена рабочими данными, расследования инцидентов, ревизий, согласований и т. п., могут послужить источником важных сведений о реальных практиках управления данными.

- **Средства управления знаниями и репозитории метаданных**: стандарты, политики, правила, методики, повестки и протоколы рабочих совещаний, постановления, приказы и т. п. по вопросам, касающимся руководства и управления данными, наряду с техническими артефактами и документацией должны (в идеале) храниться в хорошо каталогизированных и управляемых библиотеках и архивах. В рамках некоторых моделей CMM/DMM наличие/отсутствие подобных хранилищ и надлежащего порядка в этих хранилищах подлежит оценке как отдельный показатель зрелости организации. Хранилища метаданных могут оказаться сегментированными на ряд никак не связанных между собой логических структур, о чем участники порой даже и не догадываются. Например, некоторые аналитические приложения всецело полагаются на собственные репозитории метаданных при формировании представлений и отчетов; при этом в документации к ним эти артефакты описываются без упоминания слова «метаданные» (например, «библиотека шаблонов»); как следствие, в организации появляются неучтенные и рассогласованные между собой источники метаданных.

4. МЕТОДЫ

Обычно методы проведения процедур DMMA определяются выбранной рамочной структурой. Поэтому ниже представлены лишь описания самых общих методов, понимание смысла которых поможет вам выбрать подходящую рамочную структуру, где реализованы подходящие для вашей организации методы.

4.1 Выбор рамочной структуры DMM

При выборе рамочной структуры DMM следует руководствоваться следующими критериями.

- **Векторы перспективного развития методом приращений.** Специфика приоритетов у каждой организации своя, однако рамочная модель DMM должна содержать логический алгоритм поэтапного движения вперед в каждой описываемой ею функциональной области.
- **Воспроизводимость.** Рамочная структура DMM может и должна обеспечивать однозначную интерпретацию результатов оценки и воспроизводимость результатов, иначе теряет смысл всякое сравнение показателей организации с отраслевыми показателями и делается невозможной оценка динамики изменения показателей.
- **Всесторонность.** Рамочная структура DMM должна охватывать максимально широкий спектр аспектов и компонентов DM, включая учет бизнес-факторов и обеспечение заинтересованности бизнес-пользователей, а не одни лишь ИТ-процессы.

- **Гибкость, масштабируемость и модульная архитектура.** Структура должна предусматривать возможность как надстройки или включения расширений и дополнений с целью учета, например, отраслевых стандартов или дополнительных методик из смежных дисциплин, так и выборочного использования отдельных компонентов или модулей в зависимости от фактических потребностей организации.
- **Доступность.** Практики DM должны описываться общедоступным языком с минимальным использованием узкоспециализированных технических терминов, чтобы формулировки простыми словами передавали функциональную суть каждой операции, компонента или направления работы.
- **Методическая поддержка.** Рамочная структура должна дополняться полным комплектом материалов, необходимых для ее изучения, освоения и практического применения, включая оптимизацию.
- **Рекомендательный характер.** Хорошая рамочная структура DMM описывает, что нужно сделать, а не прописывает готовые рецепты и инструкции, как это делается.
- **Сопровождение независимой организацией.** Рамочная структура должна, во-первых, иметь собственника, а во-вторых, сопровождаться коммерчески нейтральной организацией во избежание выборочного отражения рекомендуемых практик вместо полного и объективного.
- **Тематическая организация.** Рамочная структура должна распределять работы по DM по надлежащим контекстным областям, чтобы каждое направление можно было оценивать и по отдельности, однако учитывает и зависимости между различными областями DM.
- **Технологический нейтралитет.** Рамочная структура должна фокусироваться на практиках и методах, а не продвигать и навязывать конкретные инструменты.
- **Универсальная или отраслевая.** При наличии проработанных отраслевых подходов организации следует взвешенно подойти к выбору рамочной структуры DMM: что важнее — педантичный учет специфики отраслевых стандартов DM или ориентация на лучшие образцы DM и структурное согласование моделей поверх межотраслевых границ?
- **Уровень абстракции/детализации.** Практики и критерии оценки должны описываться достаточно детально, чтобы их можно было соотнести практически со всеми специфическими аспектами структуры и работы организации, но не настолько, чтобы многие детали модели оказывались неприменимыми в контексте организации.

4.2 Возможность использования рамочной структуры DAMA-DMBOK

Рекомендации DAMA-DMBOK могут использоваться для подготовки к проведению оценки зрелости управления данными или для определения критериев выбора рамочной структуры DMM и методики DMMA. Исполнители должны сами определить прямые связи между функциональными сегментами (областями знаний) DMBOK и соответствующими им задачами (направлениями работ), с одной стороны, и классификацией выбранной структуры DMM — с другой. Области знаний, направления работ и результаты (продукты) в терминологии DMBOK можно сконфигурировать в соответствии с конкретной рамочной структурой DMM по таким параметрам, как

области исследования, необходимые технические работы по проведению измерений в каждой из областей, релевантность измерений, доступные сроки их проведения и т. п. При таком ускоренном подходе (по сути, по принципу проверочного листа) можно весьма оперативно определить области, требующие углубленного анализа, содержащие явные пробелы и недоработки или очаги острых проблем, требующие срочного лечения.

Однако рамочная структура DMBOK дает и еще одно дополнительное преимущество при ее использовании в качестве инструмента планирования экспертизы зрелости DM. Множество профессиональных сообществ знающих специалистов используют DMBOK в качестве справочного руководства и настольной книги в самых разных отраслях человеческой деятельности, в результате чего вокруг DMBOK формируется мощное сообщество по обмену практическим опытом ее применения.

5. РЕКОМЕНДАЦИИ ПО ВНЕДРЕНИЮ DMMA

5.1 Оценка готовности / Оценка рисков

Перед проведением оценки зрелости полезно выявить потенциальные риски и выработать стратегии по их смягчению. В таблице 33 обобщены основные факторы и источники риска, а также представлены возможные подходы к устранению или минимизации этих рисков.

Таблица 33. Типичные риски, связанные с проведением DMMA, и меры по их смягчению

Риск	Меры по устранению/снижению риска
Незаинтересованность руководства и сотрудников организации	Используйте неформальное общение для разъяснения концепции оценки зрелости DM и выгод от ее применения. Сформулируйте выигрыши от проведения экспертизы, прежде чем к ней приступать. Распространяйте статьи и рассказы об успехах. Заручитесь поддержкой влиятельного куратора из числа высших руководителей, чтобы формально экспертиза проводилась под началом этого лица и результаты докладывались ему же
Отсутствие опыта проведения DMMA Дефицит времени или штатных специалистов Отсутствие навыков или стандартов планирования информационно-разъяснительной работы	Привлекайте ресурсы и специалистов со стороны. Включайте передачу знаний и опыта сторонним экспертам в обязательную программу или состав работ по таким договорам
Незнание «языка данных», неумение на нем изъясняться и/или быстрый перевод разговоров о данных в плоскость обсуждения ИТ-систем и приложений	Обсуждайте DMMA в привязке к знакомым аудитории бизнес-проблемам или понятным и конкретным сценариям. Включите ликбез в план разъяснительной работы. Со временем DMMA сама научит всех участников пониманию языка данных — вне зависимости от их образования, профиля и технического опыта. Вводная ориентировка по основным понятиям DM и DMMA должна предшествовать началу проведения оценки
Неполный или устаревший арсенал аналитических средств	Делайте отметки «датируется» или просто снижайте оценки. Например, вычитайте балл из рейтинга за отсутствующее или не обновлявшееся больше года приложение

Риск	Меры по устранению/снижению риска
Узость рамок рассмотрения	Снизьте глубину фокусировки исследования до простейшей оценки DMMA и быстро пройдитесь по всем смежным областям с целью определения базовых рейтингов. Сделайте первый раунд DMMA пилотным, а затем примените извлеченные уроки в более широких масштабах. Представляйте оцениваемые вопросы строго в контексте областей знаний модели DAMA-DMBOK — и наглядно демонстрируйте, какие предметы остались за кадром и почему они обязательно должны учитываться в будущих оценках
Отсутствие доступа к нужным сотрудникам или системам	Понизьте уровень горизонтального среза DMMA, сфокусировав внимание только на доступных областях знания и сотрудниках
Непредвиденные обстоятельства и сюрпризы наподобие изменения регламентирующих требований	Добавьте гибкости в рабочие процессы и предусмотрите возможность перенастройки и смены фокусировки

5.2 Организационные и культурные изменения

Инициирование или кардинальное усиление общеорганизационной программы управления данными требует изменения процессов, методов и инструментов. Масштабные технологические перемены, в свою очередь, должны сопровождаться соответствующими изменениями в культуре организации. Организационно-культурная трансформация начинается с признания того факта, что не всё благополучно в устоявшемся и привычном порядке вещей — и он требует изменений. Контрольно-измерительные функции обычно как раз и позволяют сигнализировать о проблемах и подстегивают организацию к осмысленным изменениям. Экспертиза зрелости управления данными (DMMA) позволяет организации понять свое истинное место в пространстве, описываемом градуированной системой координат зрелости DM, и проложить оптимальный маршрут к совершенствованию. При этом она же и подсказывает организации направление каждого следующего шага, выступая в роли навигационного прибора, помогающего ориентироваться в изменениях. Результаты DMMA должны включаться в повестку общеорганизационного обсуждения более широкого круга вопросов стратегического развития. При должной поддержке со стороны органа руководства данными результаты DMMA позволяют синтезировать из различных точек зрения и взглядов на перспективы разделяемое всеми ви́дение желанного будущего — и ускорять прогресс организации на пути к его воплощению в жизнь (см. главу 17).

6. РУКОВОДСТВО УПРАВЛЕНИЕМ ЗРЕЛОСТЬЮ

Обычно процесс оценки зрелости управления данными включается в состав мероприятий по руководству данными организации, каждое из которых имеет собственный жизненный цикл. Жизненный цикл DMMA включает первичное планирование и начальную оценку, выработку рекомендаций, согласование и утверждение плана действий с последующими периодическими

переоценками и корректирующими мерами. При этом и сам жизненный цикл DMMA нуждается в руководстве.

6.1 Надзор за процессом DMMA

Надзорные функции в отношении процесса DMMA относятся к сфере компетенции команды по руководству данными. Если формальная функция руководства данными в организации не представлена, надзор по умолчанию сохраняется за управляющим комитетом или звеном управления, инициировавшим DMMA. У проекта должен быть куратор на высшем уровне исполнительного руководства (в идеале на эту роль лучше всего подходит CDO), который будет обеспечивать прямое согласование работ по совершенствованию управления данными с бизнес-задачами.

Объем и глубина надзора зависят от масштабности DMMA. Каждая функция или сторона, участвующая в процессе, должна иметь право голоса в принятии решений, касающихся исполнения, методов, результатов и оперативных планов, принимаемых по результатам общей оценки зрелости DM. Каждая затрагиваемая область управления данными и каждая функция организации имеет право на собственную независимую точку зрения, но именно через рамочную модель DMM они и должны находить общий язык и приходить к консенсусу.

6.2 Метрики

Будучи ключевым компонентом любой стратегии совершенствования, измеримые параметры служат еще и базовым средством обеспечения взаимопонимания. Первичные оценки параметров модели DMMA служат рейтингами текущего состояния управления данными. Периодические переоценки позволяют демонстрировать тенденции к улучшению. Каждая организация должна разрабатывать собственные метрики в привязке к маршрутной карте перехода в желаемое состояние. Примеры возможных измеримых показателей приведены ниже.

- **Показатели DMMA**: моментальный снимок текущего функционального состояния DM в организации. Каждую оценку можно сопровождать описанием. Можно использовать средневзвешенные оценки по направлениям или предметным областям с указанием целевых или рекомендуемых значений, соответствующих желаемому состоянию.
- **Показатели использования ресурсов**: мощные средства отражения затратности управления данными, включая лобовые показатели трудозатрат. Пример подобной метрики: «В среднем каждый информационный ресурс в организации отнимает у сотрудников 10% рабочего времени — на ручной сбор и ввод данных».
- **Уровень риска** (либо, напротив, защищенности от риска) или способности адекватно реагировать на возникновение угрозы отражает функциональные возможности организации согласно текущим рейтингам DMMA. Например, если организация собирается открыть новую линию, требующую высокого уровня автоматизации, а в текущей операционной модели управление данными осуществляется преимущественно вручную (уровень 1), с открытием новой линии лучше повременить из-за высокого риска срыва поставок.

- **Управление затратами** отражает структуру распределения затрат на управление данными по подразделениям организации и влияние этих затрат на устойчивость, себестоимость и окупаемость управления данными. Эта группа показателей пересекается с метриками руководства данными и может включать:
  - устойчивость управления данными;
  - достижение целей и решение задач проектами и инициативами;
  - эффективность коммуникации;
  - эффективность информационно-разъяснительной работы;
  - эффективность учебно-методической работы и профессиональной подготовки;
  - скорость восприятия и привития изменений;
  - стоимостную отдачу от управления данными;
  - вклады различных компонентов в решение бизнес-задач;
  - показатели снижения рисков;
  - показатели повышения эффективности и производительности процессов.
- **Объемы и качество входных данных для DMMA** — важный комплекс измеримых показателей, демонстрирующих полноту отражения, уровень анализа и глубину детализации среза данных, необходимых для подсчета и интерпретации рейтингов по каждому оцениваемому направлению. Ключевые вводные могут включать: число, охват, доступность, количество систем, объемы данных, задействованные команды и т. п.
- **Темпы изменений** отражают скорость повышения возможностей организации относительно базового уровня, установленного первичной экспертизой DMMA. Для измерения показателей этой группы требуются регулярные переоценки.

7. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Afflerbach, Peter. *Essential Readings on Assessment*. International Reading Association, 2010. Print.

Baskarada, Sasa. *IQM-CMM: Information Quality Management Capability Maturity Model*. Vieweg+Teubner Verlag, 2009. Print. Ausgezeichnete Arbeiten zur Informationsqualität.

Boutros, Tristan and Tim Purdie. *The Process Improvement Handbook: A Blueprint for Managing Change and Increasing Organizational Performance*. McGraw-Hill Education, 2013. Print.

CMMI Institute (website), http://bit.ly/1Vev9xx

Crawford, J. Kent. *Project Management Maturity Model*. 3rd ed. Auerbach Publications, 2014. Print. PM Solutions Research.

Enterprise Data Management Council (website).

Freund, Jack and Jack Jones. *Measuring and Managing Information Risk: A FAIR Approach*. Butterworth-Heinemann, 2014. Print.

Ghavami, Peter, PhD. *Big Data Governance: Modern Data Management Principles for Hadoop, NoSQL and Big Data Analytics*. CreateSpace Independent Publishing Platform, 2015. Print.

Honeysett, Sarah. *Limited Capability — The Assessment Phase*. Amazon Digital Services LLC., 2013. Social Insecurity Book 3.

IBM Data Governance Council. https://ibm.co/2sUKIng

Jeff Gorball, *Introduction to Data Management Maturity Models*. SlideShare.net, 2016-08-01, http://bit.ly/2tsIOqR

Marchewka, Jack T. *Information Technology Project Management: Providing Measurable Organizational Value*. 5th ed. Wiley, 2016. Print.

McSweeney, Alan. *Review of Data Management Maturity Models*. SlideShare.net, 2013-10-23, http://bit.ly/2spTCY9

Persse, James R. *Implementing the Capability Maturity Model*. Wiley, 2001. Print.

Saaksvuori, Antti. *Product Management Maturity Assessment Framework*. Sirrus Publishing Ltd., 2015. Print.

Select Business Solutions. «What is the Capability Maturity Model?», http://bit.ly/IFMJI8 (Accessed 2016-11-10).

Stanford University. *Stanford Data Governance Maturity Model*, http://stanford.io/2ttOMrF

Van Haren Publishing. *IT Capability Maturity Framework IT-CMF*. Van Haren Pub, 2015. Print.

Организация управления данными и ролевые ожидания

1. ВВЕДЕНИЕ

Ландшафт информационной среды стремительно эволюционирует, и организациям нужно к этому приспосабливаться, обеспечивая оперативное развитие способов управления и руководства данными. Большинство организаций сегодня буквально утопают в данных всяческих видов и форматов, которые собираются в рамках всевозможных процессов. Лавинообразный рост объемов данных серьезно осложняет задачи управления ими. Тем временем потребители требуют быстрого и простого доступа к информации и к тому же хотят, чтобы все данные были представлены в понятной форме и позволяли оперативно получать ответы на все животрепещущие вопросы бизнеса. Для того чтобы эффективно функционировать в такой быстро эволюционирующей среде, организационные системы руководства данными и управления данными должны быть достаточно гибкими[1]. Им нужно иметь предельно ясные ответы на базовые вопросы относительно владения информационными ресурсами, обеспечения совместной работы, ответственности и принятия решений.

Настоящая глава посвящена описанию общих принципов, которые следует принимать во внимание при формировании организационной системы управления данными или руководства данными. Принципы относятся как к руководству данными, так и к управлению данными, поскольку руководство данными задает направления и предоставляет бизнес-контекст для работ, выполняемых организационной системой управления данными. Для любой из этих организационных систем невозможно предложить идеальной структуры. В то время как общие принципы распространяются на каждую из них, детали будут сильно зависеть от отраслевых факторов и корпоративной культуры самой организации.

[1] Понятия «организационная система руководства данными» (data governance organization) и «организационная системы управления данными» (data management organization) обсуждаются в главе 3. — *Примеч. науч. ред.*

2. ВЫРАБОТКА ПОНИМАНИЯ СУЩЕСТВУЮЩЕЙ ОРГАНИЗАЦИОННОЙ СИСТЕМЫ И КУЛЬТУРНЫХ НОРМ

Осведомленность, определенность в вопросах владения и подотчетность — вот ключи к тому, чтобы обеспечить активное участие людей в инициативах по управлению данными и выполнении соответствующих политик и процессов. Прежде чем браться за определение любой новой организационной системы или перестройку имеющейся с целью ее усовершенствования, важно понять текущее состояние отдельных составляющих такой системы, связанных с культурой, существующей операционной моделью и людьми (см. рис. 106).

Рисунок 106. Оценка текущего состояния с целью создания организационной системы

Например:

- **Роль данных в организации.** Какие ключевые процессы являются управляемыми на основе данных (data-driven)? Как определяются требования к данным? Насколько хорошо осознаётся роль данных в организационной стратегии?
- **Культурные нормы в отношении данных.** Какие потенциальные препятствия, обусловленные особенностями культуры, имеются на пути внедрения или совершенствования структур управления и руководства?
- **Сложившиеся практики управления и руководства данными.** Кто, где и с какими данными работает? Как и кем принимаются решения относительно данных?

- **Как организована и как выполняется работа?** Например, как соотносится деятельность в рамках проектов с операционной деятельностью? Какие комитеты и/или иные структуры осуществляют поддержку усилий по управлению данными?
- **Как организована система подотчетности?** Например, централизована организация или децентрализована? Имеет иерархическую систему управления или представляет собой «плоскую» структуру?
- **Уровни навыков.** Каков уровень знаний в области данных и управления данными у экспертов в предметных областях и других заинтересованных лиц — от рядовых сотрудников до высших руководителей?

Сформировав картину текущего состояния, оцените уровень удовлетворенности существующим положением дел, чтобы глубже понять потребности и приоритеты организации в отношении управления данными. Например:

- Располагает ли организация всей информацией, необходимой для своевременного принятия обоснованных бизнес-решений?
- Уверена ли организация в достоверности собственной финансовой отчетности?
- Способна ли отслеживать ключевые показатели эффективности?
- Работает ли организация согласно всем законам и нормативным документам, регулирующим управление данными?

Большинство организаций на момент появления у них стремления к совершенствованию практики управления или руководства данными находятся где-то в середине общепринятой шкалы оценки зрелости возможностей (то есть имеют уже не нулевой уровень, но до «пятерки» по шкале CMM, описанной в главе 15, не дотягивают). Чтобы создать релевантную организационную систему управления данными, важно понять и учесть существующую корпоративную культуру и принятые нормы. Если создаваемая организационная система не будет соответствовать действующей структуре принятия решений и практике работы комитетов, она не сможет функционировать в течение долгого времени. Следовательно, разумнее создавать и развивать подобные организации постепенно, нежели навязывать радикальные изменения.

Организационная система управления данными должна соответствовать иерархической структуре управления и ресурсам компании. Для правильного подбора людей требуется понимание как функциональной, так и политической роли управления данными в жизни организации. Целью должно являться кросс-функциональное управление с участием всех заинтересованных сторон. Для этого нужно следующее.

- Выявить сотрудников, выполняющих различные функции управления данными на текущем этапе, признать их роль и задействовать именно их; привлечение дополнительных ресурсов

допустимо только по мере роста реальных потребностей в специалистах по управлению и руководству данными.

- Изучить используемые в организации методы управления данными и определить возможные пути совершенствования процессов; оценить объемы изменений, которые могут потребоваться для оптимизации практик управления данными.
- Составить дорожную карту внесения разного рода изменений, необходимых с точки зрения организации для обеспечения наилучшего соответствия требованиям.

3. СТРУКТУРЫ ОРГАНИЗАЦИОННЫХ СИСТЕМ УПРАВЛЕНИЯ ДАННЫМИ

Критически важным шагом на пути создания организационной системы управления данными является определение наиболее подходящей операционной модели. Операционная модель служит рамочной структурой для определения ролей, обязанностей и процессов принятия решений. Она описывает порядок взаимодействия людей и функций.

Надежная операционная модель помогает наладить механизмы подотчетности, поскольку все необходимые функции в ней представлены. Она способствует развитию коммуникаций и обеспечивает поддержку процесса разрешения проблемных ситуаций. Формируя базис для организационной структуры, операционная модель, однако, не задает ее фиксированную схему. Речь идет не о составлении штатного расписания, а об описании взаимосвязей между составными частями организационной системы.

Далее в этом разделе будет представлен высокоуровневый обзор плюсов и минусов децентрализованной, сетевой, гибридной, федеративной и централизованной операционных моделей.

3.1 Децентрализованная операционная модель

В рамках децентрализованной модели ответственность за управление данными распределяется по различным направлениям бизнеса и деятельности в области ИТ (см. рис. 107). Любое сотрудничество возможно только через комитеты; единый ответственный отсутствует. Многие программы управления данными начинаются по инициативе снизу, направленной на то, чтобы хоть как-то упорядочить практику управления данными в масштабах организации, а потому и носят по определению децентрализованный характер.

К преимуществам такой модели можно отнести практически плоскую (с минимальным количеством уровней иерархии) структуру и хорошую согласованность между различными аспектами управления данными, направлениями бизнеса и ИТ. Наличие такой согласованности обычно подразумевает четкое понимание требований к данным. К тому же децентрализованные модели проще внедрять и совершенствовать.

К минусам относится слишком большое число участников, вовлеченных в деятельность руководящих органов и принятие решений. Коллегиальные решения труднее вырабатывать

и реализовывать, нежели безапелляционные приказы. Децентрализованные модели к тому же часто носят весьма неформальный характер, что не способствует закреплению результатов. Для успешности им требуется находить способы обеспечения последовательного соблюдения наработанных практик, а такие усилия трудно координировать в условиях децентрализации. Также нередко возникают и трудности с определением владельцев данных в рамках децентрализованной модели.

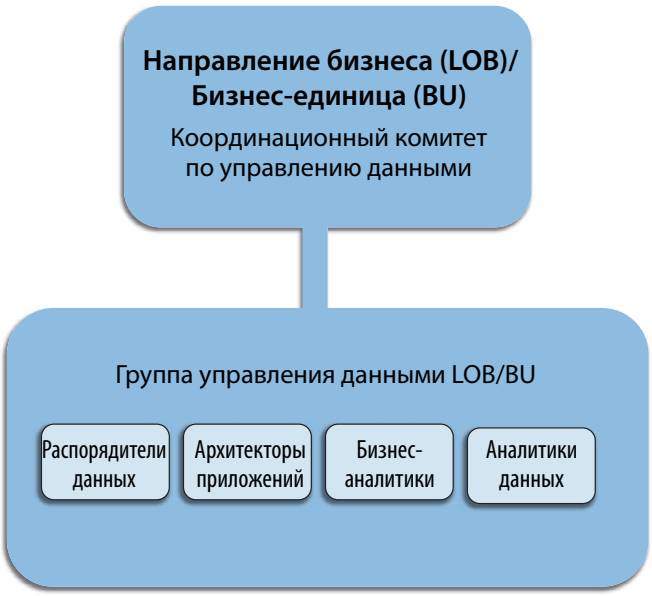


Рисунок 107. Децентрализованная операционная модель

3.2 Сетевая операционная модель

Децентрализованную модель можно сделать более строгой и формализованной посредством ее дополнения задокументированным распределением ролей и обязанностей, — как правило, описываемым через матрицу RACI (см. главу 7, раздел 5.4). Такую модель принято называть сетевой, поскольку она работает по принципу использования горизонтальных связей между людьми, участвующими в процессе в различных ролях, и может быть отражена в виде «сети» (см. рис. 108.)

Сетевая модель наследует большинство плюсов децентрализованной (плоская структура, простота согласования и настройки), а добавление распределения функциональной ответственности согласно матрице RACI позволяет еще и обеспечивать подотчетность исполнителей без выстраивания или перекройки организационных структур. Сохраняются и все недостатки децентрализации, усугубляющиеся еще и необходимостью учета и контроля выполнения своих функций участниками RACI-схемы.

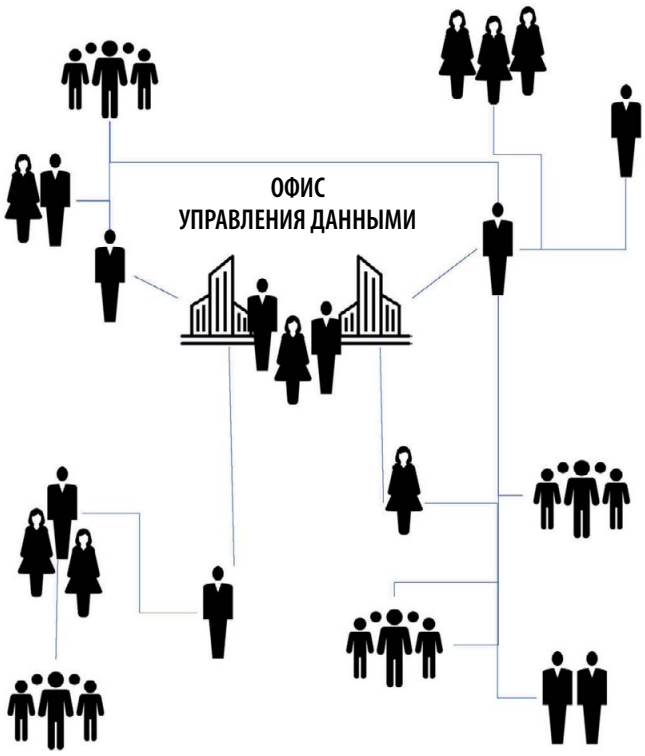


Рисунок 108. Сетевая операционная модель

3.3 Централизованная операционная модель

Самой формализованной и зрелой является централизованная операционная модель управления данными (см. рис. 109). В ней всё упорядочено и входит в сферу влияния организационной системы управления данными. Все участники процессов руководства и управления данными отчитываются непосредственно перед главным руководителем по вопросам управления данными, отвечающим за руководство и распоряжение данными, управление метаданными, управление качеством данных, управление справочными и основными данными, архитектуру данных, бизнес-аналитику и т. д.

Преимущество централизованной модели — наличие формального руководителя, отвечающего за управление и руководство данными. Наличие одного руководителя упрощает процессы принятия решений, распределения заданий и обязанностей, способствует определению четкой схемы подотчетности. Внутри организации управление данными может осуществляться отдельно в зависимости от их видов или предметных областей. Сдерживающим фактором для внедрения такой модели являются затратность и трудоемкость реализации. К тому же обычно она требует серьезной структурной перестройки всей организации. Кроме того, всегда имеется риск, что формальное выделение функции управления данными в самостоятельную вертикаль приведет к ее отрыву от основных бизнес-процессов, а со временем, как следствие, и к утрате их понимания.

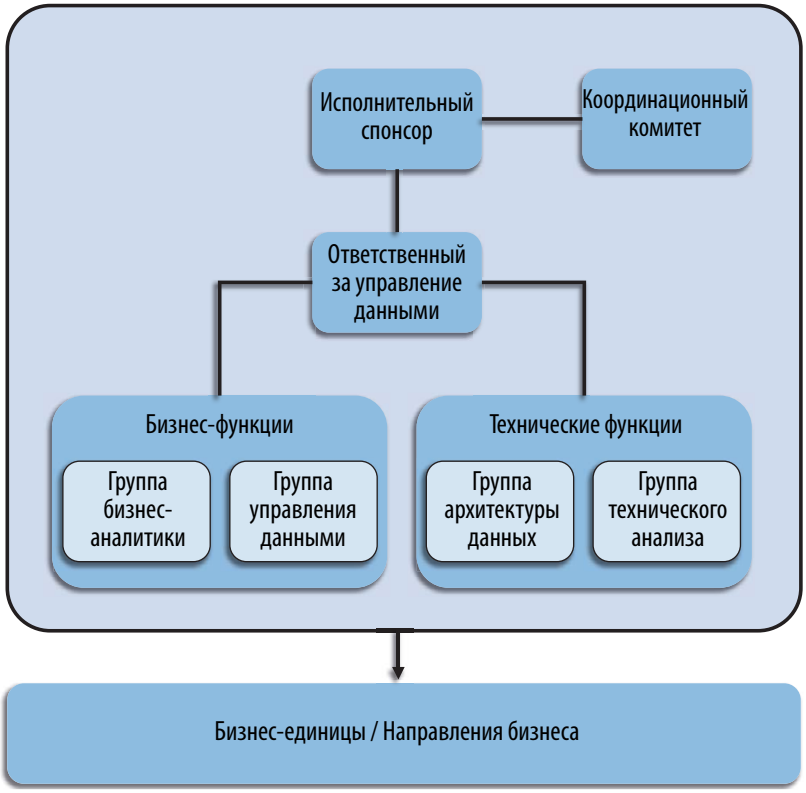


Рисунок 109. Централизованная операционная модель

Централизованная модель обычно требует выстраивания новой организационной системы. Встают вопросы: «Какое место должна занимать организационная система управления данными в общей структуре корпоративного управления? Кто должен ее возглавлять и кому он будет подотчетен?» — Ответом на эти вопросы всё чаще становится выведение организационной системы управления данными из-под начала директора по ИТ (CIO) для того, чтобы переориентировать ее на поддержку в первую очередь бизнес-функций, а не ИТ-аспектов управления данными. В результате организационные системы управления данными всё чаще входят в состав общих корпоративных сервисных или операционных команд или являются частью организационной системы, возглавляемой директором по данным (CDO) (см. раздел 6.1).

3.4 Гибридная операционная модель

Из самого ее названия понятно, что гибридная операционная модель — плод скрещивания централизованной и децентрализованной моделей с целью совмещения их преимуществ (см. рис. 110). В рамках гибридной модели головной Центр компетенций в области управления данными координирует работу децентрализованных групп управления данными бизнес-единиц, стратегические направления совершенствования обычно определяются Координационным комитетом, в котором представлены все ключевые направления бизнеса, а тактические вопросы

решаются на уровне рабочих групп, создаваемых внутри бизнес-единиц и руководствующихся методическими рекомендациями Центра компетенций.

Рисунок 110. Гибридная операционная модель

Гибридная модель оставляет часть ролевых функций децентрализованными. Например, проектировщики архитектуры данных организационно могут оставаться в составе группы управления корпоративной архитектурой, а функции обеспечения качества данных — в ведении руководства направлениями бизнеса, у каждого из которых будет собственная команда качества данных. Какие роли централизовать, а какие оставить децентрализованными — зависит от множества факторов, обусловливаемых преимущественно организационной культурой.

Главное преимущество гибридной модели состоит в том, что она позволяет задавать общее направление согласованного управления или руководства данными по всей организации сверху донизу. Во главе системы управления или руководства данными стоит исполнительный руководитель, отвечающий за скоординированность действий. При этом подотчетность бизнес-единиц в части управления данными носит весьма общий характер, что позволяет им выстраивать свою работу с данными в соответствии с бизнес-приоритетами и фокусироваться на решении оперативных задач. А помощь в этом им оказывает специализированный головной Центр компетенций в области управления данными.

К минусам можно отнести изначальную трудность создания такой организации, требующего, помимо прочего, расширения штатного расписания с целью укомплектования специалистами создаваемого с нуля Центра компетенций. Команды бизнес-единиц могут иметь принципиально различные взгляды на приоритетные направления, и их придется без конца согласовывать с целью обеспечения упорядоченного управления данными в масштабах предприятия. Кроме того,

не исключены и конфликты приоритетов и даже интересов между централизованными и децентрализованными организационными системами.

3.5 Федеративная операционная модель

Разновидностью гибридной операционной модели является федеративная организация управления данными, предусматривающая многоуровневую централизацию. Обычно такой подход используется лишь в крупнейших транснациональных корпорациях и глобальных организациях. Мысленно представьте себе корпоративную организационную систему управления данными, включающую множество различных параллельно реализованных в разных регионах и/или направлениях бизнеса гибридных моделей управления данными — это и будет федеративная операционная модель (см. рис. 111).

Рисунок 111. Федеративная операционная модель

Федеративная модель обеспечивает децентрализованное исполнение централизованно вырабатываемых стратегических планов. Следовательно, для крупных многопрофильных организаций она, возможно, является единственной работоспособной моделью организационной системы управления данными. Главе исполнительного органа управления данными всей организации подчинен Центр компетенций. И, конечно же, все направления бизнеса уполномочены осуществлять управление данными в соответствии со специфическими требованиями, диктуемыми их

потребностями и приоритетами. Федерализация позволяет организации избирательно выстраивать схемы управления на уровне сущностей данных в зависимости от задач, определяемых направлениями деятельности или региональными приоритетами.

Главное препятствие на пути реализации этой высокоэффективной модели — ее сложность. Многослойность плюс необходимость балансировки уровней автономии направлений бизнеса с нуждами всей организации способны сами по себе негативно сказаться на решении приоритетных корпоративных задач.

3.6 Выбор оптимальной для организации операционной модели

Выбор операционной модели — отправная точка совершенствования практики управления и руководства данными. Чтобы правильно сориентироваться, какую модель внедрять, для начала нужно разобраться с текущим состоянием организационной системы управления данными и ее вероятной эволюцией в обозримой перспективе. Поскольку операционная модель будет служить структурной основой для определения, согласования, утверждения и исполнения правил и процессов, критически важно выявить наиболее подходящий для организации вариант.

Оцените степени централизации/децентрализации текущей организационной системы управления данными, ее иерархичности или горизонтальности, а в случае гибридной структуры — выявите компоненты, относящиеся к различным типам операционных моделей. Опишите, насколько независимы структурные подразделения, направления, региональные представительства или филиалы в плане распоряжения/управления данными. Функционируют ли они в автономном режиме или полагаются на указания из центра и детально отчитываются о своей операционной деятельности? Насколько различаются стоящие перед ними цели и предъявляемые к ним требования? Самое важное: постарайтесь определить, как там принимаются решения (например, консенсусом, голосованием или единоличным приказом) и как именно реализуются принятые решения.

Ответы на все подобные вопросы как раз и позволят определить отправную точку с позиции текущего места организации на шкале централизации/децентрализации управления данными.

3.7 Альтернативные варианты организационной системы и соображения проектирования

Большинство организаций начинают с децентрализованной модели и лишь через какое-то время осознают необходимость перехода к более формальной организационной системе управления данными (Data Management Organization, DMO). Обратив внимание на тот факт, что упорядочение управления данными позитивно влияет на их качество, организация может приступить к формализации ответственности в сфере управления данными — например, с использованием матрицы RACI — и вскоре перейти на сетевую модель DMO. Еще через какое-то время синергетический эффект взаимодействия распределенных ролей станет очевидным, как и желательность экономии за счет укрупнения, и часть ролей будет вытянута из сети на формализованный общеорганизационный уровень, а на местах будут сформированы рабочие группы по управлению данными. Путем таких метаморфоз DMO самоорганизуется по схеме гибридной или федеративной модели.

Некоторые организации не могут себе позволить роскоши дожидаться естественного вызревания DMO. Скорость может потребоваться, допустим, вследствие рыночных потрясений или принятия каких-то новых законов или регламентов. В таких случаях важно сразу же уделить первоочередное внимание профилактике неприятия и отторжения быстрых организационных перемен сотрудниками, иначе устойчивого успеха попытка перехода на новую модель DMO не принесет (см. главу 17).

Вне зависимости от выбора модели важно помнить, что залогом ее восприятия культурной средой и устойчивого привития в организации являются простота и практичность. Если операционная модель хорошо укладывается в рамки корпоративной культуры, то все необходимые механизмы управления и надлежащего руководства данными можно встроить в текущие операции и согласовать со стратегией развития. При проработке проекта операционной модели руководствуйтесь следующими правилами.

- Начните с оценки текущей ситуации с целью определения отправной точки.
- Планируйте операционную модель в привязке к организационной структуре.
- Обязательно учитывайте следующие факторы:
  - сложность + зрелость организации;
  - сложность + зрелость предметной области;
  - масштабируемость.
- Заручитесь поддержкой на высшем уровне руководства: это обязательно для долгосрочной устойчивости модели.
- Важнейшие решения должны приниматься высокоуровневым форумом или органом (оргкомитетом, руководящим, консультативным или координационным советом, правлением и т. п.). Добейтесь закрепления этого положения в уставе или ином регламентирующем документе организации.
- При необходимости используйте пилотные проекты, поэтапное, итерационное, каскадное или волновое внедрение.
- Фокусируйтесь на областях данных, имеющих особо высокое значение.
- Используйте уже имеющиеся в организации наработки.
- Никогда не стремитесь объять необъятное и выработать единый, универсальный и всех устраивающий подход, — это нереалистично.

4. КРИТИЧЕСКИЕ ФАКТОРЫ УСПЕХА

Можно выделить десять факторов, играющих ключевую роль в успехе и обеспечении эффективности организационной системы управления данными вне зависимости от ее модели и фактической структуры. К таковым относятся:

1) наличие куратора в высшем руководстве;
2) четкое представление о желаемом состоянии;
3) активное управление изменениями;
4) отсутствие разногласий в руководстве DMO;
5) информирование и обратная связь;
6) привлечение к участию всех заинтересованных сторон;
7) инструктаж и подготовка кадров;
8) мониторинг восприятия новых практик;
9) непреложность соблюдения руководящих принципов;
10) эволюционный характер изменений (никаких революций и больших скачков).

4.1 Куратор в высшем руководстве

Правильный выбор высокопоставленного куратора программы реорганизации управления данными позволяет рассчитывать на координацию усилий по проведению всех необходимых в переходный период изменений действенным и эффективным образом, позволяющим получить на выходе новую, информационно-ориентированную организацию, способную функционировать долго и устойчиво. Важно, чтобы поручитель(ница) понимал(а) смысл инициативы и верил(а) в ее необходимость и реалистичность успеха, а также обладал(а) достаточным весом и возможностями для обеспечения поддержки изменений другими членами высшего руководства.

4.2 Четкость ви́дения

Ясное представление об идеальной организационной системе управления данными в сочетании с планом ее материализации — важнейшее условие успеха. Задача лидеров — обеспечить понимание и усвоение всеми участниками и сторонами, заинтересованными в управлении данными, в том числе и вне организации, смысла, содержания и значения организационной системы управления данными, в том числе и для их собственной работы и встречного влияния их самих на DMO.

4.3 Упреждающее планирование изменений

Управление изменениями при формировании структуры организационной системы управления данными требует не только стратегического, но и оперативно-тактического планирования, обеспечивающего своевременный характер управления изменениями с целью обеспечения их устойчивости. Применение методологии управления организационными изменениями к процессам становления организационной системы управления данными помогает решать проблемы, обусловленные человеческим фактором, и повышает вероятность долгосрочной устойчивости DMO по завершении перехода к желаемой модели (см. главу 17).

4.4 Согласование позиций руководства

Единодушие руководства относительно необходимости всесторонней поддержки развития корпоративной программы управления данными и выработка единой, согласованной позиции по

вопросам определения оптимальных путей реализации и критериев успеха DMO — еще одно условие перехода на общеорганизационную модель. Согласованные позиции должны быть выработаны как в отношении преследуемых целей, так и в отношении желаемых результатов и критериев ценности управления данными, и даже в отношении намерений и назначения различных компонентов.

Если в руководстве имеются разногласия или недопонимание хоть по каким-то вопросам DMO, рано или поздно оттуда пойдут рассогласованные и просто противоречащие друг другу распоряжения в адрес исполнителей — и кончится всё подрывом доверия к DMO вплоть до саботажа дальнейших изменений. Следовательно, критически важно регулярно сверять позиции руководителей всех подразделений и уровней с целью выявления и оперативного устранения нестыковок.

4.5 Прямая и обратная связь

Информационно-разъяснительная работа должна стартовать одновременно с программой DMO и вестись регулярно с непременными каналами обратной связи. Организаторы обязаны обеспечивать полное и четкое понимание всеми заинтересованными сторонами смысла, содержания и значения управления данными для компании; всем и каждому должно быть разъяснено, что именно изменяется, как эти изменения скажутся на их работе, от каких привычек им придется отказаться и какие приобрести. Чтобы лучше управлять данными, людям нужно знать, во-первых, что именно они делают не так в настоящем и, во-вторых, как это исправить, изменить или исполнять по-другому, чтобы соответствовать будущим требованиям. Преподносите инициативу по реорганизации управления данными креативно, как красивую историю со встроенными в нее ключевыми суггестивными посланиями, способствующими переосмыслению процессов DMO.

Содержание месседжей должно последовательно подчеркивать важность управления данными и способствовать формированию у аудитории целостного представления об идеальной DMO. Кроме того, послания следует адаптировать сообразно менталитету, специфике восприятия и уровню образования целевых групп. Разъяснения должны быть неназойливыми, но по мере надобности повторяться вплоть до полного усвоения. Эффективность посланий и рост уровней понимания нуждаются в мониторинге. В случае низкой эффективности сообщений нужно их изменять, а не интенсифицировать пропаганду.

4.6 Обеспечение заинтересованности и участия

Людям, затрагиваемым инициативой по DMO как на индивидуальном, так и на групповом уровне, свойственны различные и не всегда предсказуемые реакции на новую программу и отводимые им в рамках этой программы роли. От способности организации заинтересовать и увлечь всех участников управления данными, найти правильный подход и общий язык, отнестись с пониманием к их нуждам, найти мотиваторы и стимулы зависят шансы на успех начинания.

Анализ заинтересованных лиц и сторон помогает организации точнее очертить круг затрагиваемых лиц и выявить группы риска, которые могут болезненно отреагировать на изменения. Сопоставив эту информацию со степенями значимости, уровнями влияния и заинтересованности различных лиц и групп в реализации программы DMO, а также выявив потенциальные очаги

сопротивления, организация сможет оптимальным образом определить подходы к различным фигурантам процесса изменений, обеспечивающие максимум содействия и нейтрализующие риск активного противодействия (см. раздел 5.3).

4.7 Ориентировка, инструктаж и подготовка

Без дополнительного обучения и повышения квалификации сотрудников цели реорганизации управления данными недостижимы. При этом разным целевым группам, очевидно, потребуются различные типы и уровни программ переподготовки.

Лидерам понадобится расширение кругозора и вводные ориентировки по широкому спектру вопросов управления данными в контексте их ценности для компании. Распорядителям, владельцам и попечителям данных (то есть всем тем, кто будет выступать на переднем крае реализации практических изменений) потребуется глубокое понимание всех технических аспектов инициативы по DMO. Углубленная подготовка, сфокусированная на практических аспектах изменений в их областях работы, призвана помочь им впоследствии более эффективно исполнять свои функциональные роли. А это подразумевает овладение навыками работы с новыми правилами, процессами, приемами, процедурами и даже инструментальными средствами.

4.8 Мониторинг восприятия и освоения новых методов

Важно выстроить систему измеримых показателей, всесторонне описывающих прогресс внедрения, восприятия и освоения новых принципов и правил управления данными, а также спланировать работу по объективному контролю выполнения мероприятий, предусмотренных дорожной картой внедрения DMO, и обеспечения ее последующей устойчивой работы. Регулярный мониторинг должен позволять получать объективную информацию по следующим вопросам:

- уровни восприятия/освоения новых методов;
- процентные или абсолютные показатели улучшения ситуации или роста контрольных цифр («дельты») по сравнению с предыдущим состоянием;
- различные аспекты благотворного влияния управления данными, то есть измеримые показатели улучшения работы различных систем и решений, которые могут быть объяснены исключительно реализацией программы DMO;
- усовершенствованные процессы, реализованные проекты;
- улучшение показателей своевременного выявления и устранения рисков;
- инновационный аспект управления данными, то есть вклад DMO в фундаментальные изменения технологических и бизнес-процессов;
- достоверность аналитики.

К числу показателей благотворного влияния управления данными могут быть отнесены любые аспекты совершенствования информационно-зависимых процессов — от сроков сдачи и точности отчетности по итогам месяца до управления рисками и мониторинга эффективности исполнения

проектных работ. Мониторинг инновационного аспекта управления данными можно сфокусировать на показателях совершенствования процессов принятия решений и бизнес-анализа за счет повышения качества и достоверности данных.

4.9 Соблюдение руководящих принципов

Руководящим принципом называют четко сформулированное официальное заявление позиции организации по какому-либо важному вопросу, в полной мере характеризующее общепринятые ценности и стратегическое ви́дение миссии организации и включаемое в число незыблемых основ комплексного принятия решений. Совокупность руководящих принципов диктует детализированные правила и ограничения, критерии и исключения, этические нормы и рабочие процедуры, которых будет придерживаться организация в повседневной работе и которые остаются незыблемыми в долгосрочной перспективе. Вне зависимости от выбора в пользу централизованной, децентрализованной или какой-либо разновидности гибридной операционной модели DMO согласовать и утвердить руководящие принципы следует с самого начала, чтобы все участники имели возможность синхронизировать и выверять по ним все свои действия и решения. Руководящие принципы задают систему координат, в которой принимаются решения. Следовательно, их утверждение — важнейший первый шаг из начальной точки отсчета на пути к созданию программы управления данными, способной эффективно регулировать и изменять поведение людей, подразделений и всей организации.

4.10 Эволюции — да! Революции — нет!

Во всех без исключения аспектах управления данными отказ от революционных преобразований в пользу поэтапной эволюции способствует минимизации риска, неизбежно сопутствующего любым масштабным переменам. Поэтому важно выработать и закрепить в организации привычку к постепенному и поэтапному внедрению изменений по мере вызревания условий для их осуществления в рамках находящейся в процессе становления программы DMO. Пошаговые приращения возможностей и усовершенствование методов управления данными, согласованные с первоочередными приоритетами и задачами бизнеса, позволят обеспечить и привитие новых правил и процессов в организационной культуре, и устойчивые изменения в привычках и моделях поведения людей. Пошаговые изменения к тому же гораздо проще обосновывать, чтобы заручиться заинтересованной поддержкой руководства и разъяснять детали ключевым участникам.

5. ПОСТРОЕНИЕ ОРГАНИЗАЦИОННОЙ СИСТЕМЫ УПРАВЛЕНИЯ ДАННЫМИ

5.1 Выявление действующих участников управления данными

Приступая к реализации выбранной операционной модели, опирайтесь на уже имеющиеся группы сотрудников, задействованных в управлении данными. Этим вы поспособствуете, во-первых, минимизации организационных перетрясок, а во-вторых — фокусировке на рабочих вопросах управления данными, а не на кадровых или политических решениях.

Начните с ревизии существующих работ по управлению данными, чтобы установить, кто имеет отношение к созданию или получению данных, кто ими распоряжается, кто и как проверяет их качество. В этом плане полезно бывает даже просто внимательно изучить детали организационной структуры и штатного расписания — там вполне могут отыскаться артефакты вроде «отдел информационного обеспечения» или «менеджер по сбору данных». Также обследуйте всю организацию, опросите сотрудников всех подразделений, чтобы выяснить, кто реально работает с какими бы то ни было данными и в чем заключаются ролевые функции и служебные обязанности этих лиц. Названия их должностей могут быть самыми разнообразными и внешне ни о чем не говорить. Считайте, что вы занимаетесь раскрытием разветвленной сети тайных агентов-информаторов. Составив список таких «осведомителей», определите, кто какую роль играет и каких ролей в этом списке недостает. Какие дополнительные функции должны быть реализованы в соответствии с предполагаемой стратегией DMO? Какими наборами навыков должны обладить лица, призванные восполнить выявленные пробелы? Во многих случаях комплексы навыков работы с данными, отсутствующие на одном участке организации, имеются на другом — и многие пробелы заполняются посредством простой передачи навыков или обмена опытом. Не забывайте, что люди со значительным стажем работы в организации часто обладают незаменимыми практическими знаниями и опытом, которые остается лишь выявить, пересмотреть и грамотно вписать в контекст усилий по DMO.

Завершив инвентаризацию кадровых ресурсов и планирование перераспределения ролей, оцените, соответствуют ли текущие размеры оплаты труда участников ожидаемому от них вкладу в управление данными, и при необходимости инициируйте устранение выявленного дисбаланса. Понятно, что без взаимодействия с отделом кадров или департаментом управления персоналом трудно рассчитывать на согласование и утверждение нового штатного расписания с уточнением названий и описаний должностей, функций, рабочих показателей, размеров и порядка оплаты и премирования, и т. д. и т. п. Так вот: именно на стадии согласования новой кадровой структуры критически важно обеспечить назначение подходящих людей на все ключевые должности и роли на всех уровнях организации, чтобы впоследствии не возникало сомнений ни в их компетентности и надежности, ни в их готовности принимать решения в полном соответствии с выбранной линией DMO.

5.2 Определение состава участников Координационного комитета

Вне зависимости от выбора операционной модели DMO, некоторую работу должен проделать Координационный комитет, под началом которого действуют рабочие группы на местах. Крайне важно заполучить в состав этого органа нужных людей и ценить их время, расходуя его с пользой. Всегда держите членов этой команды в курсе происходящего и фокусируйте их внимание на том, каким именно образом предлагаемые усовершенствования в сфере управления данными помогут им в решении текущих бизнес-задач и достижении стратегических целей.

Многим организациям претит сама мысль об учреждении еще одного профильного комитета в дополнение к немалому числу прочих заседающих не первый год. В такой ситуации бывает

проще вынести насущные вопросы руководства и управления данными на рассмотрение существующих комитетов и протолкнуть решения о реорганизации управления данными через них. Но такой путь требует соблюдения мер предосторожности. Главный риск продвижения DMO через любой существующий комитет смежного профиля состоит в том, что инициатива, будучи одобренной на словах, на деле не будет удостоена должного внимания — особенно на начальных этапах — именно по причине ее побочности по отношению к главному кругу решаемых задач. Процессы комплектования координационного комитета и тактических рабочих групп в любом случае требуют комплексного анализа интересов сторон и выявления потенциальных кураторов (спонсоров, гарантов, поручителей) DMO.

5.3 Выявление и анализ заинтересованных сторон

Под заинтересованными сторонами понимаются любые физические или юридические лица, объединения или группы лиц, способные повлиять на программу управления данными или зависящие от нее. Заинтересованные стороны могут находиться как внутри, так и за пределами формальной структуры DMO. Внешние по отношению к DMO заинтересованные стороны обычно включают экспертов по предметным областям, высокопоставленных руководителей, различные группы сотрудников, комитеты, клиентов, потребителей, надзорные и регулирующие органы, брокеров, посредников, агентов, поставщиков и т. д. и т. п. Внутренние заинтересованные стороны — это все, кто тем или иным образом задействован в DMO, включая не только специалистов по ИТ и собственно управлению данными, но и всевозможные операционные подразделения — и надзорные, и юридические, и кадровые, и финансовые — и вообще любые функции и службы, работа которых зависит от надежности и достоверности вводных данных. При этом степень влияния на организацию внешних факторов обычно бывает столь существенна, что именно с учета нужд, интересов и потребностей внешних по отношению к организации сторон чаще всего и начинается планирование общеорганизационного управления данными.

Анализ заинтересованных сторон помогает организации определить наилучший подход к вовлечению действующих участников процесса управления данными в операционную модель DMO с максимальной отдачей от их ролевого вклада. Полученная по результатам анализа картина также помогает оптимизировать распределение времени и иных ограниченных ресурсов. Чем раньше будет проведен анализ заинтересованных сторон, тем лучше для дела, поскольку он помогает организации предсказывать реакции фигурантов на изменения и корректировать планы с целью минимизации негативных последствий. Анализ заинтересованных сторон позволяет получать ответы на вопросы из следующего ряда.

- На ком скажется программа управления данными?
- Как изменятся их роли и обязанности?
- Какие реакции на изменения могут последовать с их стороны?
- С какими проблемами столкнутся люди? Чего они могут опасаться?

Результатом анализа станет список заинтересованных сторон с указанием их целей и приоритетов, а также расшифровкой причин, по которым эти цели и приоритеты для них важны. Исходя из результатов анализа, обдумайте меры, которые необходимо предпринять в отношении различных фигурантов списка. Особое внимание уделяйте планам привлечения на свою сторону критически важных участников, которым по силам как внести решающий вклад в успех DMO, так и пустить инициативу под откос в масштабах организации, в частности изначально дискредитировать ее приоритеты. Обязательные для рассмотрения и учета факторы:

- Кто контролирует распределение важнейших ресурсов?
- Кто способен заблокировать инициативы по DMO — прямо или косвенно?
- Кто обладает серьезным влиянием на других ключевых фигурантов и способен склонить их к решениям за или против программы DMO?
- В достаточной ли мере заинтересованные стороны поддерживают грядущие изменения?

Рисунок 112 иллюстрирует простейший графический подход к приоритизации заинтересованных сторон по степени их собственной влиятельности и по степени влияния планируемой программы управления данными на их интересы.

Рисунок 112. Карта интересов участников

5.4 Привлечение заинтересованных сторон

Определив и заручившись поддержкой надежного куратора в руководстве или составив шортлист потенциальных претендентов на эту роль, важно четко сформулировать, зачем и почему

каждой из заинтересованных сторон нужно непременно участвовать в инициативе. Вполне возможно, что они вовсе не сочтут ее перспективной. Поэтому инициатору или инициативной группе проекта следует четко артикулировать причины, по которым успешная программа управления данными необходима заинтересованным сторонам, а их содействие — программе. Это подразумевает понимание личных и профессиональных целей, преследуемых заинтересованными сторонами, и способность увязывать результаты процессов управления данными с этими целями, причем самым прямым образом, чтобы связь сделалась им предельно ясна и понятна. Без понимания прямой зависимости собственных успехов от эффективной DMO они, возможно, и согласятся оказать программе содействие, но лишь разово, а на долгосрочную помощь и поддержку от них рассчитывать не приходится.

6. ВЗАИМОДЕЙСТВИЕ DMO С ДРУГИМИ ОРГАНАМИ УПРАВЛЕНИЯ

Определив операционную модель и состав участников, пора выдвигать людей на новые ответственные позиции, согласованные с руководством. Переход DMO к функционированию подразумевает образование комитетов и налаживание взаимодействия со значимыми заинтересованными сторонами. При централизованной модели большая часть работ по управлению данными будет поставлена под прямой контроль единой организационной системы. При децентрализованной или сетевой модели, однако, DMO придется налаживать совместную работу с другими группами влияния на практику управления данными — и влияния серьезного. Обычно к таковым относятся:

- организационная система под началом директора по данным (CDO);
- органы руководства данными;
- команды качества данных;
- корпоративный архитектор.

6.1 Директор по данным

Формально признавая важное значение данных как ценного корпоративного ресурса, большинство компаний пока что не дозрели до введения должности директора по данным (CDO) или аналогичной ей на уровне совета директоров с целью наведения надежного моста через пропасть, отделяющую ИТ от бизнеса, и возведения управления данными в масштабах организации в ранг официально исповедуемой на высшем уровне культовой стратегии. Но тенденция к стремительному росту числа подобных компаний очевидна, особенно в регулируемых отраслях; по оценке агентства Gartner, к 2017 году в половине поднадзорных организаций будет иметься штатная должность CDO (Gartner, 2015).

Требования к CDO и функции этого должностного лица сильно варьируются в зависимости от корпоративной культуры, организационной структуры и нужд бизнеса, но многие CDO

участвуют в стратегическом планировании бизнеса, выступают в роли советников высшего руководства, гарантов качества данных, ну и практически всегда и везде исполняют представительские функции по всем вопросам управления данными.

В 2014 году на портале сообщества Dataversity было опубликовано исследование общераспространенных полномочий CDO[1]. К ним относятся:

- определение организационной стратегии в области данных;
- согласование требований к данным с имеющимися ИТ- и бизнес-ресурсами;
- определение и утверждение стандартов, правил и процедур руководства данными;
- консультационная поддержка (и, при необходимости, обслуживание) инициатив бизнеса, зависящих от данных, в таких областях, как, например, бизнес-аналитика, большие данные, обеспечение качества данных и технологии сбора, обработки и передачи данных;
- донесение до понимания всех значимых сторон внутри и вне организации важности соблюдения принципов качественного управления данными и информацией;
- обеспечение надзора за использованием данных в бизнес-аналитических целях.

Полученные Dataversity результаты также ярко высветили смещение фокусов внимания в самых разных отраслях.

Вне зависимости от отрасли всё более широкое распространение получает практика переподчинения организационной системы управления данными CDO. В случае же менее централизованной операционной модели за CDO сохраняется ответственность за определение стратегии управления данными, а *исполняется* эта стратегия уже силами блока ИТ, а также операционных и иных бизнес-подразделений. В некоторых случаях DMO, изначально созданные при участии CDO лишь в роли главного консультанта по вопросам стратегического планирования, со временем передают под его оперативное руководство и многие другие аспекты управления и руководства данными и даже бизнес-аналитики, поскольку такая централизация оказывается эффективнее или экономичнее благодаря упразднению излишних параллельных организационных структур.

6.2 Руководство данными

Руководство данными является организационной рамочной структурой для выработки стратегии, целей и задач, политики и правил эффективного управления данными. К нему также относятся процессы, регламенты, организационные системы и технологии, необходимые для обеспечения доступности, годности, целостности, согласованности, достоверности и защищенности данных. Поскольку программа руководства данными состоит из хитросплетения рабочих стратегий, стандартов, политик и коммуникаций, так или иначе относящихся к данным, она работает в тесной синергетической связи с управлением данными, определяя рамки выстраивания процессов управления в соответствии с приоритетами бизнеса и заинтересованных сторон.

[1] http://bit.ly/2sTf3Cy

При централизованной модели Офис руководства данными либо подчинен организационной системе управления данными, либо наоборот. В тех случаях, когда программа управления данными ориентирована прежде всего на выработку политики и правил управления данными как ценным активом, логично осуществлять ее под общим руководством Офиса руководства данными, сделав организационную систему управления данными подчиненной или подотчетной (или выстраивать DMO в формате матричного подчинения Офиса руководства данными по направлениям работы). Такой подход сплошь и рядом наблюдается в сильно зарегулированных средах, где главное — соблюдение правил и строгая отчетность.

Но и в самых децентрализованных вариантах моделей управления данными необходимо тесное партнерство между DMO и Офисом руководства данными с целью четкого разграничения и согласования функций разработки политик, правил и инструкций в сфере управления данными (руководство данными) и их реализации (DMO). Джон Лэдли предельно лаконично разъясняет разницу и неразрывную связь между двумя этими областями: руководство данными нужно для того, чтобы «делать правильные вещи» (Doing the right things), а управление данными — для того, чтобы «делать вещи правильно» (Doing things right) (Ladley, 2012). Это две стороны одной медали и два основных компонента, необходимых для получения ценных данных. Таким образом, можно считать руководство данными методом упорядочения управления данными или, образно выражаясь, функцией штабного планирования боевых операций.

Главное же, нужно четко понимать эту синергию и согласовывать роли, обязанности и сферы ответственности с целью одновременного выполнения указаний и инструкций по руководству данными и обеспечения эффективности оперативного управления данными. К участию в рабочей группе по руководству данными можно и нужно привлекать представителей организационной системы управления данными, а DMO, в свою очередь, может работать в качестве уполномоченного исполнительного звена и осуществлять надзорные функции от имени и по поручению представителей руководства данными и под «прикрытием с воздуха» их административным ресурсом.

6.3 Управление качеством данных

Управление качеством данных — ключевой функциональный аспект практики управления данными в любой организации. Многие DMO как раз и создаются изначально в рамках усилий по обеспечению качества вследствие понимания необходимости согласованных усилий по повышению качества данных в общеорганизационных масштабах. Можно, конечно, заниматься качеством данных и на уровне отдельно взятого направления работы или компьютерного приложения, не привлекая никого из смежных областей и не вдаваясь в кросс-функциональные сложности. Однако по мере вызревания практик управления качеством данных организация доходит до осознания полезности унифицированного подхода и создает некую централизованную структуру управления качеством, например Центр компетенций. Фокус внимания переносится на обеспечение согласованного повышения качества данных на всех направлениях бизнеса или во всех приложениях, что чаще всего подразумевает переход к управлению основными данными на корпоративном уровне.

Нередко организационная система управления данными органично произрастает из первоначальной инициативы по повышению качества данных какой-то отдельно взятой категории, после того как руководство убеждается, что даже локальное вложение в повышение качества данных приносит ощутимую стоимостную отдачу в масштабах всей компании, и расширяет программу качества данных на смежные дисциплины — управление основными, справочными данными и метаданными.

Программа качества данных может развиваться параллельно с программой управления данными и даже частично с нею объединяться по причине схожести операционных моделей, однако до полной централизации функций управления качеством данных в крупных компаниях дело доходит крайне редко, поскольку практически всегда и везде имеются специфические для каждого направления бизнеса или конкретного приложения аспекты качества данных, управление которыми в централизации не нуждается. Там, где программа качества данных реализована по децентрализованной, сетевой или гибридной (с Центром компетенций) модели, нужно согласовывать компоненты операционной модели управления качеством данных с компонентами общей модели DMO во избежание конфликтов интересов, отношений, иерархий, отчетности, стандартов, процедур и даже программных средств.

6.4 Корпоративная архитектура

Группа проектирования корпоративной архитектуры разрабатывает и документирует высокоуровневые схемы построения информационных систем организации с целью оптимизации решения стратегических задач. Архитектура предприятия как область знания охватывает следующие предметы:

- технологическую архитектуру;
- архитектуру приложений;
- информационную архитектуру (или архитектуру данных);
- бизнес-архитектуру.

Без понимания архитектуры данных, как фундамента всех функций управления данными, организация управления данными попросту невозможна. Следовательно, архитекторы данных могут включаться в состав Совета по руководству данными или рабочей группы DMO с целью обеспечения пунктирной связи с группой проектирования архитектуры предприятия.

В тех случаях, когда архитекторы данных входят непосредственно в состав руководства DMO, их взаимодействие с остальными коллегами обычно осуществляется в рамках наблюдательных советов по архитектуре (Architecture Review Board, ARB) или комитетов с аналогичными функциями регулярного изучения и пересмотра архитектурных стандартов, реализуемых в различных проектах или программах и/или влияющих на их реализацию, а также обратного влияния реализуемых проектов/программ на стандарты архитектуры данных. ARB может и должен быть наделен полномочиями санкционировать или отклонять новые проекты

и системы по признаку их соответствия или несоответствия действующим архитектурным стандартам.

В случае отсутствия в DMO собственных архитекторов данных специалистам по управлению данными остается полагаться лишь на следующие возможные каналы взаимодействия с проектировщиками корпоративной архитектуры данных.

- **Через команды по руководству данными**: поскольку в программе руководства данными так или иначе участвуют представители и области управления данными и разработчики корпоративной архитектуры данных, комитет или рабочая группы по руководству данными могут служить рабочей площадкой для согласования целей, ожиданий, стандартов, направления и содержания работ.
- **Через ARB**: поскольку при наличии такового любые проекты в области управления данными подлежат согласованию с ARB, на этой стадии архитекторы как раз и могут давать наставления, отзывы и рекомендации, выполнение которых позволит им санкционировать проект.
- **По мере необходимости**: при отсутствии формальных комитетов главам архитектурного направления и управления данными нужно периодически встречаться и согласовывать проекты и процессы, идущие на обоих направлениях, с целью их координирования и минимизации обоюдного негативного влияния. Со временем затруднительность такого подхода, вероятно, так или иначе побудит руководство к учреждению формальной должности или комитета, отвечающего за обеспечение согласований.

Были бы архитекторы данных, — а кому из них излагать архитектурные соображения в рамках встреч рабочей группы по руководству данными и/или играть руководящую роль на обсуждениях ARB, они и сами разберутся.

6.5 Особенности управления данными, присущие глобальным организациям

Транснациональным компаниям и международным организациям приходится сталкиваться со множеством дополнительных трудностей, обусловленных необходимостью управления данными в глобальных масштабах с учетом множества разнородных национальных законов, норм и правил, регулирующих, в частности, защиту персональных и конфиденциальных данных, а также чувствительных данных иных категорий, состав и правила обращения с которыми могут зависеть от специфики местного законодательства. Всё вышеперечисленное накладывается на сложную и без того структуру данных, характерную для глобальной организации и обусловленную объективными факторами (распределенный характер рабочей силы и систем, различия в часовых поясах и языках, и т. п.), что делает процесс решения задачи обеспечения согласованного, действенного и эффективного управления данными в общеорганизационных масштабах занятием столь же увлекательным и нескончаемым, как обучение котов строевой подготовке.

Глобальным организациям надлежит уделять особое внимание следующим аспектам управления данными:

- соблюдение стандартов;
- синхронизация процессов;
- упорядочение линий отчетности;
- подготовка кадров;
- информационно-разъяснительная работа и обратная связь;
- мониторинг и измерение;
- экономия за счет масштабирования;
- устранение дублирований.

По мере глобализации программ и организационных систем управления данными повышается привлекательность сетевых и федеративных моделей, обеспечивающих согласование на уровне отчетности и стандартов при относительной свободе выбора и варьирования систем и механизмов на региональном уровне.

7. РОЛИ В ОБЛАСТИ УПРАВЛЕНИЯ ДАННЫМИ

Роли в области управления данными могут определяться как на функциональном, так и на индивидуальном уровне. При этом и названия ролей, и их относительная значимость и нужность могут варьироваться от организации к организации.

Все ИТ-роли можно привязать к фазам или точкам жизненного цикла данных, и все они сказываются на управлении данными — либо напрямую (например, когда архитектор создает проект хранилища данных), либо косвенно (например, когда веб-разработчик оснащает сайт пользовательскими и прикладными программными интерфейсами). Многие бизнес-роли также связаны с созданием или обработкой данных. Некоторые роли (например, анализ качества данных) требуют от исполнителей сочетания технических навыков со знанием специфики бизнеса. Ниже кратко описаны лишь те роли и функции, которые непосредственно задействованы в управлении данными.

7.1 Организационные роли

Организационные системы управления данными, относящиеся к блоку ИТ, предлагают широкий спектр сервисов, начиная с проектирования ИТ-инфраструктуры и архитектуры данных/приложений и заканчивая поставкой под ключ и администрированием СУБД.

При централизованной сервисной организационной системе управления данными всё внимание уделяется исключительно управлению данными как таковому. В такую команду могут входить исполнительный директор по управлению данными, подотчетные ему менеджеры, архитекторы и аналитики данных, эксперты по качеству данных, администраторы баз данных, администраторы информационной безопасности, специалисты по метаданным, моделированию и администрированию данных, проектировщики архитектуры хранилищ и интеграции данных,

аналитики BI — иными словами, представители всех групп, имеющих прямое или косвенное отношение к сбору, обработке и хранению данных.

При частично распределенной сервисной организационной системе управления данными (федеративной модели) имеется ряд параллельно функционирующих ИТ-подразделений, каждое из которых отвечает за обеспечение своего среза работ по управлению данными. Чем крупнее организация, тем больше, как правило, децентрализуются ИТ-функции. Например, каждой бизнес-функции может быть придана собственная группа разработчиков ПО. Возможен и гибридный подход, например: разработчики приложений у каждой бизнес-функции имеются собственные, а функция администрирования БД остается централизованной.

Бизнес-функции, обеспечивающие управление данными, чаще всего ассоциируются с командами по руководству данными или корпоративному управлению информацией. Например, распорядители данных часто входят в состав организационной системы по руководству данными, что облегчает работу органов руководства данными, в частности таких, как Совет по руководству данными.

7.2 Индивидуальные роли

Индивидуальные роли могут определяться в рамках бизнес- или ИТ-функций. Часть ролей так или иначе будет иметь смешанный (гибридный) характер, поскольку требует от исполнителей знания как информационных систем, так и бизнес-процессов.

7.2.1 Руководящие роли

Исполнительное руководство управления данными в зависимости от ситуации может представлять как сферу ИТ, так и сферу бизнеса. Должности директора по ИТ (CIO) и технического директора (CTO) уже получили повсеместное распространение, а в последние годы и позиция директора по данными (CDO) начала активно приживаться в практике корпоративного управления.

7.2.2 Бизнес-роли

Бизнес-роли преимущественно относятся к функциям руководства данными. Прежде всего, речь идет о распорядителях данных по направлениям деятельности. Как правило, это признанные эксперты в предметных областях, которые отвечают за метаданные и качество данных, как по отдельным бизнес-сущностям, так и по предметным областям и, в целом, по базам данных. Роли распорядителей данных могут серьезно разниться в зависимости от специфики и приоритетов организации. Часто они отвечают за обеспечение точности определений бизнес-терминов и областей значений различных данных в соответствующих предметных областях. Но со временем во многих организациях находят целесообразным задействовать этих же распорядителей для определения критериев качества данных, бизнес-правил и атрибутов данных в профильной области, выявления и содействия разрешению проблемных вопросов. Они же незаменимы при определении стандартов, правил и процедур. Распорядители данных

могут функционировать на разных уровнях — предприятия, бизнес-подразделения или функции управления. Роли этих людей могут быть как формализованными на уровне должностей (например, ответственный за конкретные данные), так и совершенно неформальными (люди просто заботятся о том, чтобы с данными по их линии работы всё было в порядке, вне зависимости от названия должности).

Помимо формальных распорядителей данных, свой важный вклад в организацию управления данными вносят бизнес-аналитики и разработчики моделей бизнес-процессов, поскольку без их участия попросту невозможно обеспечить соответствие моделируемых процессов реальным и, как следствие, пригодность накапливаемых данных для использования их ниже по информационно-технологическому потоку.

Вносят свой посильный вклад в общеорганизационное управление данными и другие категории компетентных сотрудников со стороны бизнеса: например, потребители аналитики, оставляющие замечания по поводу публикуемых организацией данных, которые способствуют всестороннему совершенствованию управления данными.

7.2.3 ИТ-роли

В управлении данными задействуются специалисты по ИТ всевозможных профилей и уровней, включая архитекторов, разработчиков и системных администраторов баз данных и приложений, не говоря уже о многочисленном техническом персонале. Перечислим лишь самые распространенные роли.

- **Архитектор данных**: старший аналитик, отвечающий за архитектуру и интеграцию данных на уровне предприятия или функционального подразделения. В зависимости от профиля архитекторы данных могут специализироваться на построении хранилищ данных, витрин данных, процессов интеграции и т. п.
- **Разработчик модели данных** отвечает за выявление и структурное описание (моделирование) требований к данным, определение объектов и элементов данных, а также связей между ними, бизнес-правил, требований к качеству данных и, в целом, логических и физических моделей данных.
- **Администратор модели данных** отвечает за управление версиями модели данных и их своевременное и согласованное обновление.
- **Администратор базы данных** отвечает за своевременное получение и обработку массивов входящих данных, а также технологическое обеспечение их доступности.
- **Администратор информационной безопасности** отвечает за контроль доступа к данным в зависимости от уровня защиты данных и прав доступа, имеющихся у запрашивающих доступ сторон.
- **Архитектор интеграции данных** отвечает за принципиальное обеспечение совместимости и качества данных на уровне предприятия.

- **Специалист по интеграции данных** проектирует, разрабатывает и внедряет системы интеграции (копирования, извлечения, преобразования, загрузки) массивов необходимых данных в требуемых режимах (пакетной или потоковой обработки).
- **Разработчик аналитических/статистических отчетов** занимается созданием программных средств генерирования отчетов в согласованных форматах.
- **Архитектор приложений** отвечает за интеграцию прикладного ПО с системами управления данными.
- **Технический архитектор** координирует работы по интеграции портфеля новых ИТ-проектов в существующую инфраструктуру.
- **Технический инженер** отвечает за изыскание и реализацию (в пределах своей компетенции) возможностей совершенствования ИТ-инфраструктуры.
- **Администратор службы поддержки** отвечает за своевременный прием, обработку и отслеживание всех сигналов, а также контроль разрешения проблем, связанных с передачей/получением данных, работой информационных систем или ИТ-инфраструктуры.
- **ИТ-аудитор**: более чем желательно наличие внешнего или внутреннего независимого контроля соблюдения всех критериев приемлемости ИТ-обеспечения управления данными, включая контроль качества и обеспечение надлежащего уровня ИБ.

7.2.4 Гибридные роли

Гибридные роли требуют от исполнителей сплава технических навыков со знанием бизнеса. В зависимости от специфики организации такие специалисты могут формально относиться как к бизнес-подразделениям, так и к ИТ-службам. Типичные примеры:

- **Аналитик качества данных** отвечает за пригодность данных к использованию и текущий мониторинг состояния данных; участвует в анализе корневых причин выявленных проблем с данными и вырабатывает рекомендации по реорганизации бизнес-процессов и совершенствованию ИТ-решений, направленные на устранение недостатков и повышение качества данных.
- **Специалист по метаданным** отвечает за их определение и интеграцию, управление метаданными и их своевременное обновление, исполняет функции администратора хранилищ метаданных.
- **Архитектор BI**: старший бизнес-аналитик, отвечающий за выбор и интеграцию приложений, используемых в пользовательской BI-среде.
- **Администратор BI** отвечает за эффективный доступ бизнес-пользователей к средствам и результатам бизнес-анализа.
- **Руководитель программы BI** координирует на корпоративном уровне БА-требования, инициативы и проекты, обеспечивает их интеграцию в комплексную программу приоритетных BI-исследований и осуществляет оперативное планирование ее реализации.

8. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Aiken, Peter and Juanita Billings. *Monetizing Data Management: Finding the Value in your Organization's Most Important Asset*. Technics Publications, LLC, 2013. Print.

Aiken, Peter and Michael M. Gorman. *The Case for the Chief Data Officer: Recasting the C-Suite to Leverage Your Most Valuable Asset*. Morgan Kaufmann, 2013. Print.

Anderson, Carl. *Creating a Data-Driven Organization*. O'Reilly Media, 2015. Print.

Arthur, Lisa. *Big Data Marketing: Engage Your Customers More Effectively and Drive Value*. Wiley, 2013. Print.

Blokdijk, Gerard. *Stakeholder Analysis — Simple Steps to Win, Insights and Opportunities for Maxing Out Success*. Complete Publishing, 2015. Print.

Borek, Alexander et al. *Total Information Risk Management: Maximizing the Value of Data and Information Assets*. Morgan Kaufmann, 2013. Print.

Brestoff, Nelson E. and William H. Inmon. *Preventing Litigation: An Early Warning System to Get Big Value Out of Big Data*. Business Expert Press, 2015. Print.

Collier, Ken W. Agile *Analytics: A Value-Driven Approach to Business Intelligence and Data Warehousing*. Addison-Wesley Professional, 2011. Print. Agile Software Development Ser.

Dean, Jared. *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. Wiley, 2014. Print. Wiley and SAS Business Ser.

Dietrich, Brenda L., Emily C. Plachy and Maureen F. Norton. *Analytics Across the Enterprise: How IBM Realizes Business Value from Big Data and Analytics*. IBM Press, 2014. Print.

Freeman, R. Edward. *Strategic Management: A Stakeholder Approach*. Cambridge University Press, 2010. Print.

Gartner, Tom McCall, contributor. «Understanding the Chief Data Officer Role». 18 February 2015, http://gtnr.it/1RIDKa6

Gemignani, Zach, et al. *Data Fluency: Empowering Your Organization with Effective Data Communication*. Wiley, 2014. Print.

Gibbons, Paul. *The Science of Successful Organizational Change: How Leaders Set Strategy, Change Behavior, and Create an Agile Culture*. Pearson FT Press, 2015. Print.

Harrison, Michael I. *Diagnosing Organizations: Methods, Models, and Processes*. 3rd ed. SAGE Publications, Inc, 2004. Print. Applied Social Research Methods (Book 8).

Harvard Business Review, John P. Kotter et al. *HBR's 10 Must Reads on Change Management*. Harvard Business Review Press, 2011. Print. HBR's 10 Must Reads.

Hatch, Mary Jo and Ann L. Cunliffe. *Organization Theory: Modern, Symbolic, and Postmodern Perspectives*. 3rd ed. Oxford University Press, 2013. Print.

Hiatt, Jeffrey and Timothy Creasey. *Change Management: The People Side of Change*. Prosci Learning Center Publications, 2012. Print.

Hillard, Robert. *Information-Driven Business: How to Manage Data and Information for Maximum Advantage*. Wiley, 2010. Print.

Hoverstadt, Patrick. *The Fractal Organization: Creating sustainable organizations with the Viable System Model*. Wiley, 2009. Print.

Howson, Cindi. *Successful Business Intelligence: Unlock the Value of BI and Big Data*. 2nd ed. Mcgraw-Hill Osborne Media, 2013. Print.

Kates, Amy and Jay R. Galbraith. *Designing Your Organization: Using the STAR Model to Solve 5 Critical Design Challenges*. Jossey-Bass, 2007. Print.

Kesler, Gregory and Amy Kates. *Bridging Organization Design and Performance: Five Ways to Activate a Global Operation Model*. Jossey-Bass, 2015. Print.

Little, Jason. *Lean Change Management: Innovative practices for managing organizational change*. Happy Melly Express, 2014. Print.

National Renewable Energy Laboratory. *Stakeholder Analysis Methodologies Resource Book*. BiblioGov, 2012. Print.

Prokscha, Susanne. *Practical Guide to Clinical Data Management*. 2nd ed. CRC Press, 2006. Print.

Schmarzo, Bill. *Big Data MBA: Driving Business Strategies with Data Science*. Wiley, 2015. Print.

Soares, Sunil. *The Chief Data Officer Handbook for Data Governance*. MC Press, 2015. Print.

Stubbs, Evan. *The Value of Business Analytics: Identifying the Path to Profitability*. Wiley, 2011. Print.

Tompkins, Jonathan R. *Organization Theory and Public Management*. Wadsworth Publishing, 2004. Print.

Tsoukas, Haridimos and Christian Knudsen, eds. *The Oxford Handbook of Organization Theory: Meta-theoretical Perspectives*. Oxford University Press, 2005. Print. Oxford Handbooks.

Verhoef, Peter C., Edwin Kooge and Natasha Walk. *Creating Value with Big Data Analytics: Making Smarter Marketing Decisions*. Routledge, 2016. Print.

Willows, David and Brian Bedrick, eds. *Effective Data Management for Schools*. John Catt Educational Ltd, 2012. Print. Effective International Schools Ser.

Управление данными и управление организационными изменениями

1. ВВЕДЕНИЕ

Большинству организаций для совершенствования практики управления данными требуется изменение методов совместной работы и переосмысление сотрудниками роли данных в их деятельности, а также влияния использования данных и структуры ИТ-решений на рабочие процессы. Успешное управление данными в современной практике требует, среди прочего, соблюдения следующих условий:

- восприятие и усвоение парадигмы горизонтального управления данными через согласование каналов передачи ценной информации с цепочками создания добавленной стоимости;
- переориентация с раздельных вертикалей подотчетности на распределенное информационное обслуживание общих нужд;
- превращение качества информации из предмета заботы каждого отдельно взятого бизнес-подразделения или побочной функции ИТ-отдела в фундаментальную ценность организации;
- переосмысление понятия «обеспечение качества информации» с повышением его статуса с уровня «проверки и очистки данных» до уровня основного функционала организации;
- внедрение процессов для измерения стоимости халатного отношения и ценности дисциплинированного подхода к управлению данными.

Изменений подобного масштаба одними информационно-технологическими новшествами не достичь, хотя надлежащий выбор программного обеспечения может этому способствовать. Но достигается желаемый результат прежде всего за счет тщательно продуманного и выстроенного

подхода к управлению изменениями в организации. Изменения потребуются на всех уровнях, так же как и скоординированное управление ими во избежание тупиковых инициатив и подрыва доверия как к функции информационного управления, так и к ее руководству.

Специалисты по управлению данными, наученные еще и грамотно управлять изменениями, сумеют быстрее и успешнее помочь своим организациям начать извлекать полноценную выгоду из имеющихся данных. Для этого им важно четко понимать:

- факторы, приводящие к провалу планируемых изменений;
- механизмы запуска успешных изменений;
- препятствия на пути изменений;
- характер субъективного восприятия изменений и психологических реакций на них.

2. ЭМПИРИЧЕСКИЕ ЗАКОНЫ ПРАКТИКИ ИЗМЕНЕНИЙ

Эксперты в области управления организационными изменениями вполне единодушны в формулировке свода «фундаментальных законов», объясняющих, почему перемены даются непросто. Нужно с самого начала планирования перенастройки организации отдавать себе полный отчет в незыблемости кратко сформулированных ниже эмпирических правил, без учета которых трудно рассчитывать на успешное проведение в жизнь каких бы то ни было серьезных изменений.

- **Изменяются не организации, а люди.** От объявления об учреждении новой организации или внедрении новой системы ничего не меняется. Перемены происходят лишь после того, как люди начинают вести себя иначе, нежели раньше, по причине осознания ими ценности изменений. Процесс совершенствования практик управления данными и реализации принципов формального распоряжения данными влечет далеко идущие последствия для организации. Людям будет предложено изменить привычное обращение с данными и взаимодействие друг с другом по всем направлениям работы с данными.
- **Люди не противятся изменениям, они упорствуют в нежелании меняться.** Никто лично не воспримет и не поддержит изменений, если усмотрит в них произвол, волюнтаризм или диктат. Люди с гораздо большей вероятностью согласятся изменить своим привычкам, если сами участвовали в определении необходимых изменений, видят, в чем их смысл и мотивы, и знают, как они будут внедряться и какие плоды принесут. Часть работы в рамках инициатив по реорганизации управления данными как раз и заключается в терпеливом разъяснении различным группам смысла изменений с целью выработки общеорганизационного понимания ценности отлаженной практики управления данными.
- **Сложившаяся ситуация не обусловлена ничем, кроме стечения обстоятельств.** Почему всё делается так, как делается, а не иначе? Вероятно, кто-то когда-то решил, что так будет лучше. Покопавшись в истории организации, даже можно отыскать поворотные точки принятия

тех или иных решений, утверждения бизнес-требований, определения процессов, разработки систем, составления правил или выбора бизнес-модели, которые теперь нуждаются в изменении. Изучение истоков текущей ситуации в практике управления данными бывает полезным исключительно с точки зрения минимизации риска повторения прежних ошибок. Если сотрудникам будет предоставлено право голоса в процессе согласования изменений, они лучше поймут, по каким причинам статус-кво нельзя считать приемлемым, и яснее увидят, что новые инициативы нужны для исправления ситуации.

- **Изменения требуют внешнего стимула.** Если хотите улучшений, нужно что-то менять. Эйнштейн здраво заметил: «Невозможно решить проблему на том же уровне мышления, на котором она возникла».
- **Перемены давались бы проще простого, если бы не затрагивали людей.** «Технология» изменений обычно особой сложности не представляет. Основные трудности проистекают от необходимости иметь дело с живыми людьми и их особенностями.

Для успешных изменений требуются проводники изменений, то есть люди, внимательно работающие с людьми, а не только с системами. Именно проводники изменений активно прислушиваются к мнению сотрудников, клиентов и других заинтересованных лиц, чтобы уловить назревающие проблемы до того, как они разразятся, и сгладить волны недовольства изменениями.

Ну и, конечно же, для успеха изменений требуется ясное ***ви́дение*** конечной цели и умение четко и регулярно доносить его до заинтересованных сторон, чтобы обеспечить участие и активную, а главное, долгосрочную и не ослабевающую перед лицом трудностей поддержку.

3. УПРАВЛЯТЬ НЕ ИЗМЕНЕНИЯМИ, А ПРОЦЕССОМ ПЕРЕХОДА

Общепризнанный эксперт в области управления изменениями Уильям Бриджес[1] подчеркивает центральное место именно переходного периода в процессе управления изменениями. В его понимании *переход* — это процесс психологической адаптации людей к новым ситуативным условиям. В то время как многие склонны осмысливать изменившуюся ситуацию исключительно в терминах нового начала, Бриджес утверждает, что изменение — ярко выраженный трехфазный процесс. Первая фаза — окончание — заключается в осознании того, что существующее положение вещей исчерпало себя. Труднее всего на этом этапе людям дается отказ от привычных условий существования. Мысленно покончив с прошлым, они оказываются на нейтральной полосе, где существующее состояние отпустило их еще не до конца, а новое — по-настоящему еще не началось. Изменение завершается с окончательным переходом в новое состояние — новым

[1] Уильям Бриджес (*англ.* William Bridges, 1933–2013) — историк американской культуры и литературы, в 1974 году сменивший профессию и в качестве бизнес-консультанта сформулировавший основы управления переходным периодом в жизни людей и организаций. — *Примеч. пер.*

началом (табл. 34). Из трех этих фаз самой непредсказуемой и коварной является этап пересечения нейтральной полосы, поскольку там в сознании и психике случаются самые причудливые смешения старых и новых понятий и представлений. Если члены организации не осилят переход через нейтральную полосу, высок риск быстрого скатывания к старым привычкам, и изменения надолго не приживутся.

Таблица 34. Фазы перехода по Бриджесу

Фаза перехода	Описание психологических признаков
Окончание	◆ Мы признаём, что есть вещи, которые нужно решительно отбросить. ◆ Мы признаём, что чего-то безвозвратно лишаемся. ◆ **Пример:** смена места работы. Человек увольняется со старого места, при этом непременно чего-то лишается — например, возможности тесно общаться с близкими друзьями по прежней работе.
Нейтральная полоса	◆ Со старым покончено, а по-новому всё еще не складывается. ◆ Складывается впечатление, будто никто толком и не знает, как быть и что делать дальше. ◆ Всё не на своих местах, кругом беспорядок. ◆ **Пример:** переезд на новое место жительства. В первые дни или месяцы в новом доме чувствуешь себя не «как дома», а по-прежнему «на чемоданах»: непонятно, что где лежит, а иногда царит полный хаос.
Новое начало	◆ Освоились и зажили по-новому: никакого дискомфорта. ◆ **Пример:** рождение ребенка. После первых месяцев волнений на родителей нисходит светлое успокоение: вот же оно, наше драгоценное чадо, — и как мы только без него жили?

Бриджес настаивает, что главной, а то и единственной причиной провала организационных изменений является нежелание или неспособность инициаторов перемен задумываться о проблеме окончания старого и смягчении влияния на людей скоропостижной кончины привычного порядка вещей. Он утверждает: «В большинстве организаций стартуют с нового начинания вместо того, чтобы приберечь его до финишной прямой. Внимания окончаниям не уделяют, существования нейтральной полосы не признают, а потом еще удивляются, почему это люди не желают меняться» (Bridges, 2009).

Переживая изменения, всякая личность проходит через все три фазы, но скорость перехода у каждого человека своя. Прогресс зависит от таких факторов, как накопленный жизненный опыт, образ жизни, предпочтения и привычки, степень заинтересованности и участия в выявлении и решении проблем, ну и, наконец, степени добровольности/принудительности движения по направлению к поставленным целям в субъективном восприятии тех, кого эти изменения затрагивают.

Бриджес подчеркивает, что, хотя первоочередной задачей управляющего изменениями была и остается выработка понимания и представления о пункте назначения, конечной точке маршрута — иными словами, формирования ви́дения конечной цели и путей ее достижения, — на переходном этапе наивысшей целью управления изменениями становится убеждение людей в неизбежной необходимости отправиться в путь к этой цели. На переходной фазе высокую

роль играют проводники изменений и, в целом, авторитетные менеджеры и лидеры мнений, способные помочь людям осознать, что процесс и этапы переходного периода являются совершенно естественными.

Рисунок 113. Фазы перехода по Бриджесу (иллюстрация)

Далее приведен проверочный лист в помощь управляющему изменениями с обобщенным описанием основных пунктов, которые следует учитывать, чтобы помочь людям поэтапно справиться со всеми проблемами переходного периода.

- Окончание
  - Помогите каждому осознать имеющиеся проблемы и причины, по которым требуются изменения.
  - Выявите, кто именно и что именно рискует потерять. Не забывайте, что для кого-то самым болезненным является расставание с друзьями и коллегами по работе, а для кого-то — формальное понижение в должности, статусе или влиянии.
  - Ощущение утраты — вещь субъективная. Что для одного горе, для другого — сущий пустяк. Просто принимайте субъективизм оценок потерь как данность. Не спорьте ни с кем, не пытайтесь никого переубедить относительно кажущейся значимости потерь и не удивляйтесь неадекватным реакциям.
  - Приготовьтесь чутко улавливать признаки болезненных переживаний по поводу утрат и открыто выражать сочувствие и поддержку.

- Определите поэлементно, с чем именно придется распрощаться безоговорочно, а что из старого багажа можно взять с собой в дорогу. К определенному моменту все должны порвать с пережитками прошлого, чтобы не затягивать отправку в путь и не травить душу муками расставания со старыми добрыми временами.
- Относитесь к прошлому с уважением. Всё-таки люди, вероятно, трудились не покладая рук в крайне трудных условиях. Признавайте их заслуги и давайте понять, что оцениваете их былые труды по достоинству.
- Подчеркивайте элементы преемственности. По возможности демонстрируйте, что окончание этапа — не конец всему: напротив, самое ценное и значимое для людей будет не просто сохранено, а улучшено и приумножено.
- Держите людей в курсе происходящего. Информируйте их раз за разом, при всяком удобном случае, разными способами — и с помощью материалов для чтения на досуге, и в личных неформальных беседах.
- Используйте анализ ключевых фигур для планирования правильного подхода к различным сотрудникам. Старайтесь оценивать ситуацию с их позиции и их глазами, — это весьма помогает в выборе правильных средств привлечения людей на свою сторону и своевременном выявлении очагов сопротивления.

- Нейтральная полоса
  - Признавайте, что фаза перехода — самая трудная (по причине смешения старого с новым), но пройти через это должен каждый.
  - Увлеките людей совместной работой; дайте им время и место для экспериментов.
  - Продолжайте давать людям понять, что их по-прежнему ценят и уважают.
  - Публично благодарите людей за любые идеи и предложения, даже если реализовать их не получается. Для апробирования инноваций и извлечения уроков лучше всего использовать классическую модель Шухарта — Деминга (цикл PDCA, см. главу 13).
  - Продолжайте постоянно информировать людей всевозможными способами.
  - Сообщайте в порядке обратной связи о результатах изучения и апробирования представленных предложений.

- Новое начало
  - Не торопите события и не давайте отмашку на запуск полностью новой системы раньше времени.
  - Убедитесь, что все люди четко знают свои роли в новой системе.
  - Убедитесь в четкости формулировок политик, правил, процедур и приоритетов. Никаких двусмысленных посланий от вас исходить не должно.
  - Спланируйте торжественный запуск новой системы с чествованием и награждением особо отличившихся в процессе реализации изменений.
  - Продолжайте постоянно информировать людей всевозможными способами.

4. ВОСЕМЬ ОШИБОК УПРАВЛЕНИЯ ИЗМЕНЕНИЯМИ ПО КОТТЕРУ

Авторитетнейший исследователь в области управления изменениями Джон П. Коттер в книге «Проведение изменений»[1] сформулировал восемь главных причин, по которым организации терпят неудачу на пути осуществления реформ. В рамках модели Коттера очень наглядно раскрываются и основные ошибки реорганизации информационного управления данными.

4.1 Ошибка № 1: самонадеянность

Величайшей ошибкой реформаторов Коттер считает резкий рывок вперед без оглядки на готовность личного состава, не подкрепленный предварительной выработкой у коллег и вышестоящих руководителей понимания причин безотлагательности изменений[2]. Анализ по Коттеру помогает управляющим изменениями учиться на чужих ошибках и не повторять их. Проводники изменений часто допускают следующие просчеты:

- переоценка своей способности подвигнуть организацию на серьезные изменения;
- недооценка трудности решения задачи снятия людей с привычных мест;
- непонимание риска цементирования статус-кво вследствие естественных защитных реакций людей на резкие движения и непродуманные подходы;
- эффект слона в посудной лавке: масштабные изменения затеваются без разъяснения их причин, смысла и назначения (то есть без формирования у людей ви́дения);
- Суетливость, встревоженность и назойливость вследствие мнимой неотложности и, как следствие, усиление опасений и неприятия у фигурантов изменений, из-за чего те и запираются (иногда даже в буквальном смысле) каждый в своей функциональной нише.

Хочется думать, что перед лицом организационного кризиса толика наглой самоуверенности не повредит, однако практика свидетельствует об обратном. Заинтересованные стороны часто мертвой хваткой цепляются за сохранение статус-кво, когда их обкладывают требованиями внесения всяческих изменений (часто еще и несовместимых между собой), что ведет к выработке естественной реакции отторжения любых перемен по принципу: «Если важно абсолютно всё, значит, на самом деле не важно ничто».

---

[1] Джон Пол Коттер (*англ.* John Paul Kotter, p. 1947) — инженер-электротехник по первому образованию (MIT, 1968), продолживший карьеру на ниве научной организации управления в Гарвардской бизнес-школе, автор двух десятков монографий. Цитируемая книга (*Leading Change*, Harvard Business School Press, 1996) в 2011 г. включена в «топ-25 влиятельнейших книг по бизнес-управлению всех времен» по версии журнала *TIME*. — *Примеч. пер.*

[2] То есть речь идет о невыполнении критерия накопления критической массы недовольства статус-кво, если описывать ситуацию по формуле перемен Глейчера (см. раздел 6).

4.1.1 Примеры в контексте информационного управления

Таблица 35 описывает примеры проявлений излишней самонадеянности в контексте информационного управления.

Таблица 35. Сценарии проявлений самонадеянности

Сценарий	Пример
Реакция на изменение внешних требований	«Нас по старым правилам не штрафовали — и по новым не тронут».
Реакция на изменение направления бизнеса	«Мы годами успешно справлялись с информационным обеспечением. Какие проблемы? Справимся и теперь».
Реакция на появление новой технологии	«А зачем она нам? Во-первых, новая — значит непроверенная. А во-вторых, у нас все системы работают стабильно и с подстраховкой».
Реакция на сбои, ошибки и проблемы	«Надо бы бригаду наладчиков туда, пусть посмотрят. Вроде бы в [название отдела] есть свободные люди? Вот пусть и разберутся».

4.2 Ошибка № 2: неспособность создать достаточно мощную поддержку сверху

Коттер постулирует практическую невозможность крупных изменений без активной поддержки не только со стороны главы организации, но и со стороны коалиции руководителей всех основных направлений ее деятельности. Особенно важна заинтересованность высшего руководства в тех случаях, когда речь идет об усилиях по совершенствованию руководства данными, поскольку они немыслимы без серьезных изменений в поведении. Без поддержки со стороны высших руководителей краткосрочные локальные и личные интересы быстро перевесят любые доводы в пользу упорядоченного руководства данными с прицелом на долгосрочную перспективу.

Руководящая коалиция — мощная команда энтузиастов-единомышленников из числа высокопоставленных представителей всех функциональных подразделений и бизнес-направлений, которая помогает вырабатывать и претворять в жизнь новые стратегии развития и за счет этого преображать организацию. Самое сложное — определить, кого именно следует привлечь к участию в руководящей коалиции (см. раздел 5.2).

4.3 Ошибка № 3: недооценка фактора наглядности при формулировке ви́дения

Сильная команда энтузиастов-единомышленников бесполезна без четкого понимания неотложности изменений. Нужно иметь еще и ясное, вразумительное и зримое представление о направлениях, содержании и результатах перемен. Именно ви́дение замысла позволяет привязать усилия к контексту и разъяснить людям смысл различных компонентов изменений. Хорошо сформулированное и переданное ви́дение способствует мобилизации энергии, необходимой для надлежащего претворения замысла в жизнь. Без публично заявленной формулировки ви́дения как руководства к принятию решений любая развилка чревата затяжными дебатами, а любое неверное действие или решение — подрывом авторитета, а то и торпедированием всей инициативы.

Главное, не путать ви́дение со стратегическим планированием или программным управлением. Ви́дение — это не план проекта, не устав проекта и не детализированное покомпонентное изложение программы изменений.

Ви́дение — это ясное, лаконичное и убедительное описание конечного результата изменений.

Донесение его до людей — залог обеспечения связи с ними. Применительно к инициативам по совершенствованию управления данными формулировка ви́дения должна кратко сообщать об имеющихся трудностях, пользе от их преодоления и пути перехода в желаемое будущее состояние.

4.3.1 Пример из практики информационного управления

Формулировка ви́дения при обосновании того или иного проекта нередко сводится к банальному описанию дополнительных функциональных возможностей предлагаемой к внедрению новой технологии. Однако при всей важности совершенствования ИТ-функционалов сами по себе картину будущего они никак не меняют. Формулировка ви́дения должна описывать, *что именно* будет достигнуто организацией за счет внедрения новых технологий.

Например, заявление вроде: «К 1 апреля мы завершим запланированное на I квартал внедрение нового интегрированного комплекса средств автоматизации финансовой отчетности и бизнес-аналитики на платформе [вставить бренд и версию]» — формулировка вполне разумной, достижимой и измеримой цели. Но к ви́дению такое заявление отношения не имеет, поскольку не дает убедительной и наглядной картины того, к чему именно приведет внедрение новинки.

А можно сформулировать перспективу иначе, например: «Мы делаем всё возможное для повышения точности, оперативности и доступности финансовой отчетности. Также мы работаем и над совершенствованием обмена текущими данными, прекрасно понимая, насколько вам всем важно иметь устойчивый и бесперебойный доступ к достоверной и актуальной информации, особенно в конце квартала, когда нудно срочно готовить отчетность. Поэтому мы для начала к концу I квартала внедрим систему [вставить название], — она вам очень поможет, но есть и еще заготовки…», — и картина прояснится. Люди поймут, что и зачем делается. Помогите им увидеть пользу от организационных изменений — и вы получите необходимую поддержку.

4.4 Ошибка № 4: недостаточная повторяемость (x10, x100, x1000) внушения ви́дения

Даже при единодушном согласии относительно неудовлетворительности текущей ситуации люди не начнут менять себя и свое поведение до тех пор, пока не увидят явных выгод от изменений по сравнению с текущим состоянием.

Последовательное, поступательное разъяснение ви́дения, подкрепленное действенными стимулами и примерами, — критически важный компонент успешного управления изменениями. Коттер, в частности, рекомендует подкреплять слово делом в части разъяснительной работы. Понимание всеми неотвратимости принципа «сказано — сделано» — критический фактор успеха. Главное, чтобы и лидеры строго придерживались заявленных принципов, ибо нет ничего убийственнее для перемен, чем сообщение: «Делай то, что я говорю, а не то, что я делаю».

4.5 Ошибка № 5: потеря ви́дения цели из-за неумения обходить препятствия

Новые инициативы обречены на провал, если люди беспомощно опускают руки, уткнувшись в кажущееся непреодолимым препятствие на пути к реализации замысла, даже если они вполне сознают необходимость продолжения движения в выбранном направлении по пути перемен. В процессе трансформации организация должна научиться выявлять крупные препятствия и умело лавировать между ними либо, если обходного пути нет, устранять корневые причины их возникновения. Препятствия на пути перемен:

- **Психологические барьеры** — в головах, и решать их нужно на уровне поиска и устранения первопричин. Откуда в людях нерешительность? Чем она обусловлена — страхом, невежеством или каким-то иным фактором?
- **Структурные препятствия** обычно носят организационный характер и могут происходить, например, от слишком узкого определения должностных обязанностей, не позволяющего налаживать продуктивное сотрудничество между различными категориями работников, или от кривизны систем оценки эффективности работы, вынуждающих людей делать трудный выбор между движением к воплощению ви́дения и сиюминутными эгоистичными интересами. Управление переменами должно включать еще и согласование структуры позитивных и негативных стимулов с замыслом преобразований.
- **Активное неприятие.** Что еще заставляет людей противиться и даже противодействовать переменам? В чем причины отказа адаптироваться к новому набору условий? Что побуждает выдвигать требования, несовместимые с преобразованиями? Особое внимание нужно уделять поведению ключевых членов организации: если на словах они всецело за реформы и озвучивают правильные вещи, вполне соответствующие ви́дению, а на деле продолжают работать и вести себя по-старому, никак не стимулируют подчиненных к изменению поведения или не приводят организацию рабочих процессов в соответствие с новой моделью, исполнение замысла неизбежно потерпит неудачу.

Коттер призывает организации полагаться на самых смекалистых в деле устранения или обхода подобных препятствий. Но если уж и они не справятся — остальные и подавно разуверятся в своих силах, и изменения не будут осуществлены.

4.6 Ошибка № 6: пренебрежение созиданием краткосрочных побед

По-настоящему глубокие и фундаментальные изменения требуют времени. Каждый, кто отваживался заняться фитнесом или лечебным голоданием ради похудения, знает этот секрет: мотивация продолжать подобные занятия подпитывается ежедневным взвешиванием, фиксирующим малые достижения и подвижки на пути к цели. Любая деятельность, накладывающая долгосрочные обязательства и вложение сил и ресурсов, требует дополнения каким-то элементом обратной

связи, позволяющим с первых шагов регулярно подтверждать отдачу от трудов и затрат, выражающуюся в малых успехах.

Потому и в комплексных усилиях по преобразованию организации обязательно должны присутствовать краткосрочные цели, способствующие решению долгосрочных задач. Достижение очередной промежуточной цели — повод для маленькой совместной радости и дополнительный стимул к продолжению движения вперед. Эти локальные победы должны ощущаться как плоды **созидания**, а не подарки судьбы в виде случайно сбывшихся надежд. При успешных преобразованиях менеджеры всегда активно ставят ближайшие цели, быстро достигают их и вознаграждают команду. Без систематических усилий, гарантирующих череду малых достижений, шансы на успех крупных реформ резко снижаются.

4.6.1 Примеры в контексте информационного управления

В практике информационного управления краткосрочные цели и победы чаще всего заключаются в успешном разрешении своевременно выявленных проблем. Например, если речь идет о проекте разработки бизнес-глоссария в рамках инициативы по внедрению общеорганизационного управления данными, краткосрочной победой можно считать разрешение противоречия между трактовками одного и того же термина двумя бизнес-подразделениями (к примеру, исправление ситуации, при которой значения одного и того же ключевого показателя у двух разных отделов не сходились по причине использования ими разных расчетных формул).

Выявление проблемы, ее решение и включение найденного решения в общее долгосрочное ви́дение преобразований позволяет команде отметить решение локальной задачи как демонстрацию работоспособности ви́дения. А также проиллюстрирует понимание ви́дения на практике и закрепит его в коллективном понимании сотрудников организации.

4.7 Ошибка № 7: преждевременное объявление о победе

Слишком часто проекты изменений, особенно растянутые на годы, чреваты соблазном заявления о полном и безоговорочном успехе после первого же ощутимого крупного достижения. Быстрые победы — мощное средство поддержания боевого духа и стимула к продолжению преобразований. Однако любое предположение, что дело сделано, как правило, оказывается серьезной ошибкой. До тех пор, пока изменения прочно не закрепились и не вросли в культуру организации, новые подходы остаются хрупкими, а старые привычки и практики способны в любой момент проявиться заново. Коттер полагает, что на полное и бесповоротное изменение всей компании целиком уходит от трех до десяти лет.

4.7.1 Пример в контексте информационного управления

Классический пример синдрома «миссия выполнена» — сценарий, при котором внедрение новой технологии рассматривается в качестве самодостаточного пути совершенствования информационного управления или разрешения проблемы с качеством или надежностью данных. После полномасштабного развертывания технологии в производственной среде бывает трудно обеспечить

продолжение целенаправленных усилий по продвижению к конечной цели проекта, особенно если ее общее ви́дение было сформулировано недостаточно четко. Таблица 36 содержит несколько примеров, связанных с последствиями поспешных объявлений о победе.

Таблица 36. Сценарии преждевременных объявлений о победе

Сценарий	Пример
Внедрение современной ИТ-системы управления качеством данных	«Мы наконец купили этот замечательный пакет средств контроля качества [TQM v. Pro+]! Все проблемы решены!» ◆ В результате в организации никто больше и не заглядывает в отчеты с метриками качества данных, включая генерируемые [TQM v. Pro+], не говоря уже о принятии мер по исправлению проблем, требующих вмешательства в ручном режиме
Непонимание того факта, что после реализации функционала он требует конфигурирования и эксплуатационного сопровождения	«Мы внедрили модуль автоматизации отчетности по линии [X-надзора]. Больше никаких нарушений по этой части у нас не выявят!» ◆ Требования изменятся, никто не озаботится перенастройкой модуля отчетности ◆ Никто не заглядывает в генерируемую отчетность, не замечает проблем, о которых она свидетельствует, и мер не принимает
Перенос данных при миграции на новую систему	«Взяли данные из *системы X*, перенесли в *систему Y*. Сообщений об ошибках не получаем». ◆ Число записей совпадает, но в *систему Y* они импортированы некорректно (например, по причине неполноты или усечения в процессе миграции). Алгоритмы преобразования нужно было определить вручную

4.8 Ошибка № 8: Пренебрежение закреплением перемен в корпоративной культуре

Меняются не организации, а люди. До тех пор, пока новые схемы поведения не внедрятся в организационную культуру, не станут неотъемлемой частью социальных норм и общепринятых ценностей, они подвержены распаду и деградации, и эти диссипативные процессы начнутся сразу же, как только будут приостановлены настойчивые усилия по продвижению перемен извне. Коттер пишет: «Вступая на путь преобразований, вы игнорируете сложившуюся культуру и делаете это на свой страх и риск».

Два ключевых момента, способствующие закреплению изменений в культуре организации:

- Раз за разом наглядно демонстрируйте людям, каким именно образом конкретные элементы их поведения и отношения к работе влияют на ее эффективность и результаты.
- Не жалейте времени на поэтапное внедрение изменений в основу культуры нового поколения менеджеров.

4.8.1 Пример в контексте информационного управления

Этот риск с особой яркостью подчеркивает важность комплексного учета всех аспектов влияния человеческого фактора на проведение общеорганизационных изменений, включая

совершенствование руководства и управления данными, практик использования и обеспечения качества данных (риск, обусловленный человеческим фактором, присутствует во всех областях).

Например, организация ввела требование обязательной маркировки всех документов метаданными в строго определенных форматах с целью автоматизации процессов архивирования в системе управлениям контентом. Поначалу все исправно сопровождают создаваемые и редактируемые документы всеми необходимыми метаданными, но через месяц-другой постепенно возвращаются к старым привычкам маркировать документы, что приводит к накоплению большого объема неклассифицированной документации, требующей повторной ручной обработки и классификации, чтобы привести их в соответствие с техническими требованиями новой системы управления контентом.

Этот пример иллюстрирует простой, но не всегда и не для всех очевидный факт: повышение качества информационного управления достигается за счет согласованного совершенствования работы процессов, людей и технологий. В указанной триаде чаще всего упускают из виду центральное звено — человеческий фактор, — а это ведет к недоработкам, снижению качества выдаваемой информации и постепенному откату назад по шкале прогресса. Вот почему так важно при внедрении новых технологий или реорганизации рабочих процессов заранее продумывать механизмы обеспечения необратимости перемен.

5. ВОСЕМЬ СТАДИЙ ПРОВЕДЕНИЯ КРУПНОЙ РЕФОРМЫ ПО КОТТЕРУ

В дополнение к восьми ошибкам управления изменениями Коттер выделяет также и восемь препятствий на пути изменений[1]:

- замкнутость и разобщенность культур;
- бюрократия;
- политика;
- низкий уровень доверия;
- несработанность команд;
- высокомерие;
- слабость или некомпетентность лидеров;
- страх перед неизвестностью.

Для успешных масштабных преобразований Коттер предлагает использовать модель восьми стадий преодоления вышеперечисленных препятствий (см. рис. 114). В рамках модели Коттера

[1] Традиция отождествления числа 8 (восемь) с благополучием и процветанием восходит к древнекитайской «Книге перемен» (易經), определяющей модель прогнозирования будущего в формате матрицы триграмм 8×8. — *Примеч. пер.*

каждая из восьми проблем решается таким образом, чтобы обеспечить долгосрочную устойчивость результатов. При этом каждой из восьми стадий соответствует одна из восьми фундаментальных ошибок управления переменами, способных привести к краху всего начинания.

Первые четыре шага направлены на преодоление обороны защитников статус-кво. По словам Коттера, все усилия на этих стадиях нужны прежде всего для осознания того, насколько нелегкие и масштабные перемены предстоит реализовать.

Затем следуют три этапа (стадии 5–7) внедрения и апробирования новых практик. И, наконец, последний шаг — закрепление изменений в культуре организации в качестве прочной платформы будущих достижений и усовершенствований.

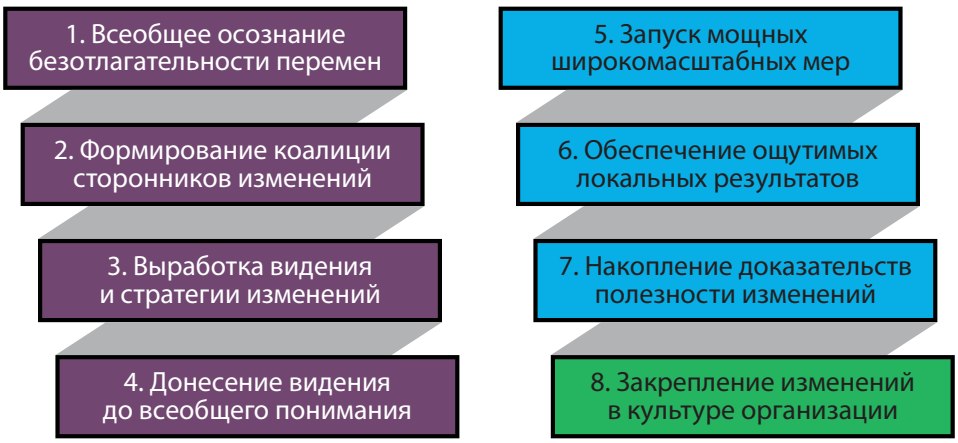


Рисунок 114. Восемь стадий крупного преобразования по Коттеру

Коттер не рекомендует перепрыгивать через стадии. Все успешные преобразования должны последовательно пройти через все восемь этапов. Всегда велик соблазн сфокусироваться на стадиях 5–7, то есть сразу же приступить к внедрению и оценке результатов. Однако, поторопившись и не заложив предварительно прочный фундамент новой модели работы (стадии 1–4), обеспечить устойчивость результатов реформы в долгосрочной перспективе не получится. Особое место у восьмой ступени: память о результатах, достигнутых на каждом из семи этапов, должна в процессе движения вперед регулярно освежаться, а ви́дение конечной цели — закрепляться и обрастать деталями в коллективном сознании организации. Это возвращает нас к напоминанию о необходимости пошагового движения вперед по пути малых успехов, каждый из которых лишь подчеркивает, что текущее состояние по-прежнему неидеально, и подсвечивает очередные проблемы, стоящие на повестке дня организации.

5.1 Выработка всеобщего понимания ситуации и безотлагательности перемен

Люди найдут тысячу причин и отговорок, чтобы увильнуть от участия в том, что им лично не нужно. Следовательно, нужно сделать так, чтобы каждый почувствовал острую потребность

личного участия в проведении изменений. Активная поддержка со стороны критической массы затрагиваемых лиц — непременное условие успеха любого преобразования.

Прямая противоположность чувству острой неотложности перемен — удовлетворенность. Когда подавляющему большинству людей вполне комфортно живется и работается в имеющихся рамках, крайне трудно — а то и невозможно — хоть как-то их расшевелить, чтобы собрать достаточно мощную группу поддержки изменений, необходимую для формирования ви́дения нового будущего и проведения реформ. В редких случаях могут появляться группы энтузиастов внутри организации, но результаты их инициатив почти никогда не бывают устойчивыми.

В контексте информационного управления создать ощущение неотложности изменений могут следующие факторы:

- изменения в законодательстве;
- угрозы информационной безопасности;
- риски приостановки видов деятельности;
- изменения в бизнес-стратегии;
- слияния и поглощения;
- проверки со стороны надзорных органов;
- угроза подачи судебных или арбитражных исков;
- серьезные технологические изменения в отрасли;
- резкое усиление позиций конкурентов;
- критика в СМИ состояния информационного управления в организации или отрасли.

5.1.1 Источники инертности

Коттер выделяет девять возможных причин инертности людей и организаций (см. рис. 115).

- При отсутствии явных симптомов кризиса трудно вывести людей из благодушного состояния и пробудить в них ощущение неотложности изменений.
- Череда локальных успехов хороша на пути к новому ви́дению организации, а в условиях застоя, напротив, лишь притупляет остроту восприятия и нередко загоняет организацию всё глубже в яму, мешая людям увидеть за чередой текущих дел критичность складывающейся ситуации, требующей срочных изменений.
- Оценка работы сотрудников по заниженным меркам или продуктов по стандартам, не соответствующим мировому уровню, чревата развитием внутри организации крайне губительных долгосрочных тенденций.
- Слишком узкофункциональная формулировка целей и задач, да еще и в сочетании с несогласованностью между собой систем измеримых показателей оценки работы подразделений различного профиля может привести к ситуации, когда *никто* не несет ответственности за эффективность и качество работы организации в целом. В результате бизнес разваливается на фоне формально исправного функционирования каждого отдельно взятого подразделения.

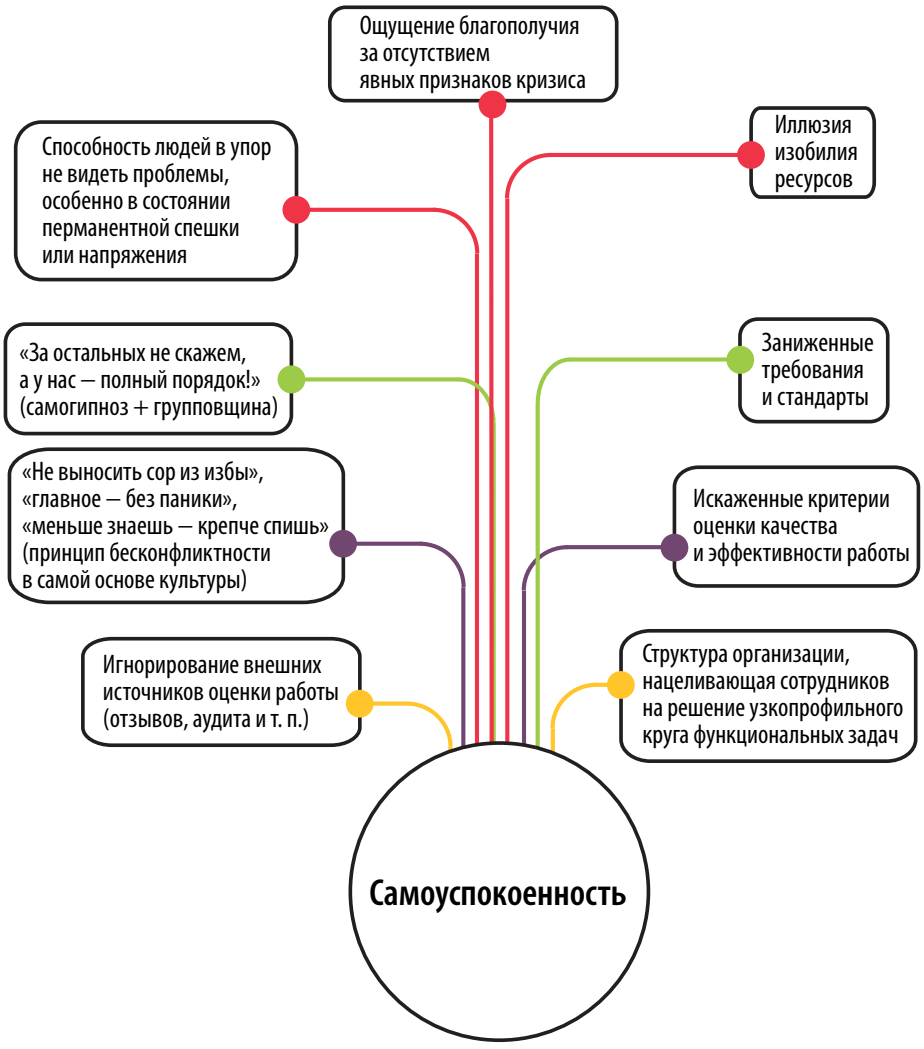


Рисунок 115. Источники самоуспокоенности

- Если внутриорганизационные системы планирования, оперативного управления и контроля заточены на легкое и простое выполнение всеми и каждым личных планов и достижение целевых показателей, инертность не заставит себя долго ждать.
- Если единственным источником обратной связи в оценке работы являются неисправные внутренние системы, ни о какой вменяемой проверке обоснованности всеобщего благодушия не может быть и речи.
- Но даже и там, где проблемы выявляются или принимаются к сведению внешние отзывы о работе организации, о них зачастую предпочитают умалчивать или — хуже того — просто затыкают рот всякому, кто о них осмеливается заикнуться, поскольку считают любую критику подрывом духа сплоченности, источником взаимных обид или просто «разговорами в строю». Проще говоря, вместо того чтобы принять информацию о недостатках к сведению

и заняться их устранением, организационная культура побуждает «заметать сор под ковер» и руководствоваться принципом «нет человека — нет проблемы» в отношении «критиканов».

- Врожденные инстинкты и благоприобретенные рефлексы самосохранения заставляют человеческую психику работать на отрицание очевидного, и люди попросту не воспринимают нелицеприятных для них вещей. Даже после того, как проблема разрослась, многие подсознательно продолжают ее игнорировать и вести себя так, будто ничего плохого не происходит, или, в лучшем случае, приуменьшают значимость проблемы и интерпретируют ее наименее болезненным лично для себя образом («от меня тут ничего не зависит», «не мне об этом судить» и т. п.).
- И даже в организациях, где восемь первых трудностей преодолены или почти не дают о себе знать, всегда имеется риск ухода от проблем, в том числе авторитетными лидерами, безостановочно вещающими об успехах и достижениях и создающими иллюзию полного благополучия. Особенно часто подобный «самогипноз» встречается в организациях с богатой историей былых успехов, придающей людям особое чувство собственной значимости и способствующей формированию культуры высокомерия. И то и другое способно погасить всякую тягу к изменениям.

Хороший эмпирический принцип: в рамках любой инициативы, направленной на организационные изменения и преобразования, нельзя недооценивать силы противодействия новому, направленные на сохранение и закрепление статус-кво. С самоуспокоенностью и самодовольством можно и обязательно нужно бороться — и бороться успешно. Ни одна организация не сможет принять и провести в жизнь по-настоящему важные решения без эффективной борьбы с реально имеющимися у нее проблемами.

5.1.2 Нагнетание ощущения неотложности изменений

Для резкого роста уровня осознания срочной необходимости перемен достаточно бывает устранить источники инертности или ослабить их. Для формирования понимания их острой неотложности от лидеров требуются еще и настойчивость, и даже смелость и отвага, включая рискованные действия на грани фола. Тут стоит вспомнить Эдвардса Деминга с его увещеваниями в адрес руководителей, включая призыв «учредить лидерство», в знаменитой программе «14 пунктов преобразования управления»[1].

Смелость и решительность, в частности, требуют действий, необходимых для того, чтобы переломить негативные тенденции, отучив подопечных от дурных привычек методами, о которых не рассказывают в рекламных буклетах. Иными словами, требуется выработать и взять на вооружение целую *новую философию* (опять же, заимствование из Деминга). Решимость в деле пробуждения сознания от инертной спячки чревата конфликтами и нервозностью, но лишь

[1] Описываемые 14 пунктов преобразования управления сформулированы в опубликованном в 1982 г. знаменитом труде Деминга *Out of the Crisis.* [Рус. пер.: *Эдвардс Деминг.* «Выход из кризиса». — М.: Альпина Паблишер, 2011]. — *Примеч. пер.*

в краткосрочной перспективе. К тому же и сами конфликты, и встревоженность можно направить в продуктивное русло, используя их для переформатирования ви́дения будущего организации, что позволяет мудрому лидеру извлекать дивиденды из устранения текущих поводов для беспокойства и капитализировать их в обоснования долгосрочных целей.

Смелые и решительные шаги, однако, немыслимы без наличия у лидеров массовой поддержки снизу, а такая поддержка обеспечивается авторитетом, который зарабатывается успешными результатами предыдущих решительных действий лидеров, — то есть круг замыкается. Излишняя осмотрительность руководства и пренебрежение постоянным и безоглядным нагнетанием и эскалацией ощущения неотложности преобразований негативно сказываются на способности организации к восприятию и реализации изменений.

5.1.3 Выработка осознания безрадостности перспективы надвигающегося кризиса

Весьма действенный путь обеспечения понимания неотложности изменений заключается в заострении всеобщего внимания на явных признаках надвигающегося кризиса. Иногда бывает полезно взбудоражить аудиторию апокалиптическими картинами того, что ожидает организацию в случае непринятия адекватных мер по выживанию в резко изменяющихся условиях, чреватых всяческими рисками. Однако и в случае согласия было бы ошибочно рассчитывать на автоматическое приведение организации в состояние готовности противостоять глобальным рискам. Финансово-экономический кризис организации любого масштаба чаще всего обусловлен не только и не столько внутриорганизационными факторами, сколько истощением прежних источников ресурсов любого рода. Как следствие, организация нередко гонится за миражами вместо сосредоточения остатков имеющихся ресурсов на решении зримых задач, определяемых ви́дением конечного результата.

Поэтому и нужно бомбардировать организацию сообщениями об успехах и достижениях (пусть даже и кажущихся), проблемах и возможностях (реальных, мнимых и потенциальных), а главное — ставить перед организацией амбициозные цели, даже если они и идут вразрез со сложившимися традициями.

Коттер предлагает использовать нонконформизм и даже умышленную конфронтацию в качестве эффективных средств пресечения проблем, что называется, в зародыше.

5.1.4 Роль менеджеров среднего и нижнего звена

В зависимости от масштабности перемен и их целевых объектов (например, отдел или вся организация) изменяется и уровень ключевых игроков. В любом случае к таковым относятся непосредственные руководители реформируемой бизнес-структуры, поскольку именно от них зависит выведение подчиненных из состояния инертности. Если подразделение достаточно автономно, у руководства есть хороший шанс расшевелить его без оглядки на темпы роста сознательности и проведения преобразований в остальной организации.

При слабой автономии попытки реформировать подразделение в отрыве от основной структуры изначально обречены, поскольку любые инициативы затухнут в силу внешней инерции.

Для устранения инерционности мышления часто бывает нужно задействовать вышестоящее руководство. Однако менеджеры среднего и нижнего звена могут выступить в роли движущей силы перемен на уровне своих подразделений, если будут мыслить и действовать стратегически. Например, они могут использовать имеющиеся навыки и средства бизнес-анализа для наглядной демонстрации неотвратимости провала какого-нибудь ключевого бизнес-проекта в случае непринятия предлагаемого ими комплекса мер по изменению статус-кво. Особенно эффективен в таких случаях вынос вопроса о необходимости изменений на третейский суд или дебаты с участием авторитетных внешних экспертов, которые, во-первых, помогут с анализом ситуации, а во-вторых, повлияют своим авторитетом на мнение излишне консервативных высших руководителей, убедив их в необходимости проведения изменений.

5.1.5 Критическая масса сознательных сторонников — это сколько?

Понимание неотложности начала поиска решения проблемы автоматически подразумевает согласие с недопустимостью сохранения статус-кво. Но для устойчивой реализации инициативы требуется ее сознательная поддержка некой критической массой менеджеров реформируемой структуры. Но каков должен быть численный перевес сторонников реформ над противниками и пассивным балластом? Коттер предлагает оценку на уровне 75% «за». Однако не следует излишне усердствовать с агитацией: переизбыток активных сторонников может сыграть с инициаторами злую шутку, если начнутся прения, в ходе которых недовольные разделятся на непримиримые противоборствующие лагеря сторонников различных методов «пожаротушения».

В целом, для запуска процесса трансформации обычно бывает достаточно просто убедительного большинства согласных с тем, что без перемен не обойтись. А степень осознания неотложности реформ должна обеспечивать вовлечение в *руководящую коалицию* достаточного числа авторитетных высокопоставленных лиц. Наконец, понимание необходимости трансформации должно быть в достаточной мере закреплено в коллективном сознании, чтобы после первых же успехов на начальной стадии преобразований энтузиазм не иссяк. Есть и еще один критически важный аспект: умение слушать клиентов, а для этого переговорить с внешними потребителями, поставщиками, акционерами или иными заинтересованными сторонами и выяснить их мнение насчет степени срочности и насущности инициируемых у вас преобразований, — обычно услышанное стороннее мнение весьма подстегивает к скорейшим действиям.

5.2 Руководящая коалиция

Не бывает людей, способных единолично выработать ви́дение, равно как не бывает и людей, обладающих всеми необходимыми связями, чтобы до всех, кому нужно, это ви́дение эффективно донести. Для успешного проведения изменений следует всячески избегать двух сценариев из разряда деструктивных крайностей:

- гордый одиночка во главе реформ (будь то CEO или неформальный лидер);
- комитет с низким уровнем доверия.

В первом случае судьба реформ оказывается всецело в руках единственного человека. Вот только темпы изменений в современных условиях таковы, что в одиночку с ними в большинстве организаций справиться нереально. В результате важные решения будут приниматься со всё большим запаздыванием либо поспешно и непродуманно. Оба варианта — путь к провалу.

Под комитетом с низким уровнем доверия понимается придание пусть даже и вполне толковому лидеру такой рабочей группы, где представлены узкие специалисты из совершенно разнородных функциональных подразделений (плюс, не исключено, еще и внешние консультанты). А вот авторитетных руководителей достаточно высокого уровня там явно недостает (а чаще и вовсе нет). Если преобразования считаются «важными, но *не настолько*», никто не будет мотивирован разбираться с истинным положением дел и поиском приемлемых в складывающейся ситуации решений. Провал такой рабочей группы неизбежен[1].

Жизненно необходимо создание компетентной руководящей коалиции, имеющей реальный авторитет и уровень полномочий, прямо указывающий всем на неотложный характер проводимых ею реформ. Кроме того, эта команда должна еще и демонстрировать способность к эффективному принятию административно-управленческих решений, — а для этого требуется еще и высокий уровень взаимного доверия между членами. При условии работы в режиме слаженной команды руководящая коалиция способна будет в кратчайшие сроки переваривать всевозрастающие объемы информации и ускорять темпы реализации идей за счет сформированной команды по-настоящему сильными, решительными и ответственными людьми с достаточными полномочиями, уровнем информированности об истинном положении вещей и заинтересованности в проведении в жизнь ключевых решений.

Четыре ключевые характеристики эффективной руководящей коалиции следующие.

- **Сила позиций**: достаточно ли ключевых игроков в команде, особенно из состава оперативного руководства основных направлений деятельности? Ведь, оставшись за бортом, они могут с легкостью затормозить процесс трансформации.
- **Профессиональный опыт**: адекватно ли представлены экспертами все важные аспекты деятельности и относящиеся к ним прикладные дисциплины? Ведь без всесторонней оценки трудно рассчитывать на принятие информированных решений.
- **Авторитетность**: достаточно ли в команде членов, пользующихся в масштабах всей организации репутацией надежных и проверенных людей? Ведь без таких никто рабочую группу всерьез воспринимать не будет.
- **Лидерские качества**: достаточно ли среди привлеченных к участию руководителей истинных и проверенных лидеров? Иначе некому будет повести организацию за собой по пути перемен.

Ключевым предметом озабоченности является как раз проблема лидерства. В руководящей коалиции необходимо обеспечить оптимальную сбалансированность между навыками управления

[1] То есть речь идет о ситуации, в равной мере хорошо описываемая сразу двумя баснями И. А. Крылова — «Квартет» и «Лебедь, рак и щука». — *Примеч. пер.*

и лидерскими качествами. Грамотные менеджеры призваны обеспечивать управляемость всего переходного процесса, а лидеры — вести людей за собой и придавать импульс переменам. Ни без первого слагаемого, ни без второго устойчивых результатов достигнуто не будет.

Кроме того, в контексте формирования руководящей коалиции неизбежно приходится решать и ряд других важнейших вопросов, включая следующие.

Сколько человек привлечь к определению и проведению в жизнь программы реформ?

Стандартный ответ бизнес-консультантов на подобные вопросы: «по обстоятельствам» — хотя и звучит как форменная издевка, но полностью отражает реальное положение дел. Можно лишь уточнить, что число членов коалиции, в целом, возрастает с ростом общей численности затрагиваемых преобразованиями целевых групп лиц. Плюс к тому: даже при самых крупномасштабных реформах число членов координационного совета не должно выходить за разумные рамки управляемости, а при маломасштабных и затрагивающих весьма ограниченную группу лиц изменениях важно следить за тем, чтобы ненароком не обойти вниманием кого-то из ключевых заинтересованных сторон, которые могут еще и обидеться из-за того, что их «не приняли в игру».

Кого именно привлекать или приглашать в состав руководящей коалиции?

Руководящая коалиция тем и отличается от формального оргкомитета проекта или координационного совета программы, что призвана стать платформой влияния на всю организацию. А раз так, то необходимо и представительство в составе коалиции всех заинтересованных сообществ. Однако коалиция на то и называется руководящей, что не может ограничиваться ролью представительного форума, созванного исключительно для высказывания и обсуждения всевозможных мнений. Всесторонне выясняйте мнения и представления людей — как влияющих на информационный обмен, так и затрагиваемых предполагаемыми изменениями в потоках данных в рамках всей организации.

Ключевой характеристикой членов руководящей коалиции является реальное влияние на свои подразделения — неважно, за счет обладания формальными полномочиями или за счет авторитета и опыта работы в организации.

Характер личности — главный ключ к подбору участников руководящей коалиции.

При формировании руководящей коалиции лидерам реформ нужно всячески избегать привлечения в свои ряды лиц с чертами характера и элементами поведения, подрывающими или ослабляющими эффективность, функциональность или популярность кампании. В частности, постарайтесь обойтись без участия личностей, относящихся к ярко выраженным деструктивным типажам.

- **Отъявленные нигилисты, скептики и пессимисты.** Своей негативистской демагогией такие типы способны подавлять любые попытки конструктивного и откровенного диалога, необходимого руководящей коалиции для выработки творческих идей и динамично оттачиваемого по ходу реализации ви́дения будущего организации и возможностей для роста.

- **Рассредоточенные фантазеры и любители абстрактных теорий.** Руководящей коалиции важно фокусироваться на текущих и предстоящих изменениях во всей их конкретике. Растекающиеся мыслью по древу участники сбивают команду с курса, уводя обсуждение в сторону, что приводит к задержкам с реализацией планов и упущенным возможностям развить первоначальные успехи.
- **Карьеристы и люди, преследующие собственные цели.** Все усилия руководящей коалиции должны быть направлены на обеспечение благополучия организации в целом и касаться всех и каждого. Наличие у кого-то из членов коалиции скрытых планов достижения собственных целей и удовлетворения корыстных потребностей за счет общеорганизационных преобразований — недопустимая роскошь.

5.2.1 Важность обеспечения эффективного лидерского руководства коалицией

Важнейшая грань различия лежит между оперативным управлением («менеджментом») и высокоуровневым руководством («лидерством»). Руководящая коалиция, целиком составленная из эффективных менеджеров, но лишенная признанных лидеров, успеха не добьется. Вакантные позиции лидеров, конечно, можно заполнить за счет привлечения авторитетных специалистов со стороны, но всё-таки лучше ориентироваться на подготовку и продвижение собственных воспитанников, попутно демонстрируя сотрудникам наличие перспектив карьерного роста, а застоявшимся лидерам — готовность смены.

При создании коалиции Коттер рекомендует остерегаться «эгоистов», «змей» и «недовольных участников».

- **Эгоисты**, будучи озабочены исключительно самоутверждением за счет других, не оставляют остальным участникам поля для высказывания.
- **Змеи** сеют раздоры, провоцируют и раздувают конфликты и, в целом, создают и удобряют почву для всеобщего взаимного недоверия и подозрительности.
- **Недовольные участники** — это, как правило, высокопоставленные лица, поддержкой которых вроде бы удалось заручиться, но смысл, назначение и необходимость изменений до их понимания явно не доходит.

Как минимум, подобных личностей следует остерегаться и не подпускать близко к руководству преобразованиями во избежание их подрыва. Желательно также и всячески оградиться от их влияния на оперативное управление программой изменений, а еще лучше держать их на коротком поводке и быть в курсе каждого их шага.

5.2.2 Пример в контексте информационного управления

Когда речь идет об инициативе по изменению информационного управления, руководящая коалиция помогает организации выявлять возможности для привязки этой инициативы к контексту различных областей деятельности и организационным реформам в целом.

Например, в случае реакции на изменение внешних законодательных требований юридический отдел сразу же начинает прорабатывать оперативный план перераспределения потоков данных и реорганизации процессов их обработки и надзора. А параллельным курсом может начаться реализация инициативы по приведению в состояние готовности к новым требованиям программного обеспечения хранилища данных, включая системы отслеживания происхождения, проверки и обеспечения качества, точности и достоверности данных.

В такой ситуации председатель совета по распоряжению данными, отвечающий за скоординированное проведение преобразований, может привлечь в руководящую коалицию и главу юридического отдела, и главу отдела отчетности с целью совместной проработки вопросов совершенствования документооборота и управления информационными процессами в контексте распоряжения данными. А это, в свою очередь, может повлечь необходимость получения вводных от менеджеров, отвечающих за создание, обработку данных и документов на местах, и привлечения их в команду с целью согласования планируемых изменений. Так, шаг за шагом, и вырабатывается полное понимание цепи создания ценной информации, что позволяет выявить и подключить к участию в руководящей коалиции всех значимых игроков.

5.2.3 Основы эффективности командного взаимодействия

Эффективная командная работа основана на двух принципах — доверии и общности цели. Недоверие обычно возникает вследствие взаимного недопонимания и недостаточной информированности, но может подпитываться еще и специфическими факторами, такими как неуместное соперничество. Например, противопоставление интересов «бизнеса» интересам «айтишников», — на этой линии фронта идут мощные внутриорганизационные битвы за ресурсы и влияние. Для выстраивания доверия регулярно занимайтесь совместными проработками важных вопросов: так выработаются устойчивое взаимопонимание, взаимоуважение и привычка учитывать интересы всех сторон. Ну и, конечно же, важнейшим условием взаимопонимания является избавление от узости и стереотипности мышления.

5.2.4 Борьба с шаблонным мышлением узкогрупповыми категориями

В психологии этот феномен принято называть *групповым мышлением*: в сплоченных и однородных по составу и интересам группах — особенно изолированных от внешних источников информации, противоречащей стойкому мнению большинства, — быстро вырабатывается всеобщая склонность к единомыслию строго в рамках понятий, диктуемых главным местным авторитетом, особенно если этот лидер не терпит возражений и не приветствует открытых обсуждений.

При групповом мышлении все единогласно поддерживают выдвинутые лидером предложения, а собственное мнение или возражения, если таковые имеются, оставляют при себе и стараются о них поскорее забыть. Признаки группового мышления:

- Никто не высказывает возражений.
- Альтернативные варианты не предлагаются.

- Противоречащие господствующему мнению точки зрения не высказываются и быстро забываются.
- Информация, способная пошатнуть единомыслие, не приветствуется.

Во избежание группового мышления важно:

- стимулировать всех участников к строгому соблюдению научной методологии сбора данных с целью лучшего понимания истинной природы и причин проблем;
- разработать и применять на практике объективные критерии оценки любых решений;
- научиться основам эффективной совместной работы, чтобы штамповка готовых решений по шаблонам группового мышления перестала казаться заманчивым средством ускорения работы;
- поощрять коллективный поиск решений методом мозгового штурма;
- придерживаться правила «лидеры высказываются последними»;
- активно изыскивать внешние знания и излагать их в процессе обсуждений;
- определившись с решением, озадачивать команду разработкой также резервного «плана Б» (чтобы еще раз переосмыслить исходные посылки, положенные в основу «плана А», и, возможно, найти даже лучшее решение).

5.2.5 Примеры в контексте информационного управления

Групповое мышление может проявляться в самых разнообразных контекстах. Наиболее характерным примером служит традиционное размежевание по линии «бизнес vs ИТ», при котором каждая из сторон упорно противится любым предложениям противоположной стороны. Другой потенциальный сценарий проталкивания групповых интересов в ущерб общеорганизационным — пренебрежение нормами информационной безопасности, защиты конфиденциальных данных и даже элементарной этики вследствие азартного увлечения сбором и анализом данных.

Имеется множество причин для применения жесткой внутриорганизационной дисциплины в сфере распоряжения данными. Одна из ключевых — необходимость обеспечения ясности и прозрачности моделей данных и методов их получения и обработки. Подобная ясность позволяет оперативно и в рамках четко определенных правил разрешать многие проблемы, возникающие как вследствие размежевания бизнеса и ИТ, так и конфликтов или конкуренции приоритетов.

5.2.6 Общность целей

Если каждый член руководящей коалиции будет упорно гнуть свою собственную линию, от взаимного доверия вскоре не останется и следа.

Типичными целями, на которых сходятся интересы представителей различных внутриорганизационных сообществ, являются совершенствование или максимально возможное повышение эффективности работы организации в четко определенных областях. Главное — не путать эти

цели с ви́дением будущего трансформированной организации, поскольку они служат всего лишь дополнением глобальной стратегии изменений.

5.3 Выработка ви́дения и стратегии

Распространенная ошибка в управлении изменениями — полагаться как на авторитарно-приказные методы, так и на ручное управление всеми деталями изменений. Ни то ни другое в сложных ситуациях переходного периода не дает действенных результатов.

Если целью ставится переучивание или изменение привычного поведения сотрудников, а босс не обладает непререкаемым авторитетом, административно-командные методы малоэффективны даже в простейших ситуациях. Таким образом, властными распоряжениями всех барьеров сопротивления сломить не удастся. С требованиями проводников изменений также особо не считаются, а по большей части их игнорируют или ловко обходят.

Обойти это слабое место нередко пытаются посредством ручного управления на микроуровне с детальными предписаниями, что и как делать каждому сотруднику, и последующим пристальным контролем соблюдения этих предписаний. Некоторые препятствия на пути изменений таким способом преодолеваются, но постепенно подобный подход начинает отнимать у менеджеров непомерно много времени и сил, затрачиваемых на детальную проработку всех рабочих инструкций и методологий, поскольку практика всё больше усложняется, увлекая за собой на качественно новый уровень сложности и рабочие процессы, регулируемые в ручном режиме.

Единственный подход, стабильно помогающий проводникам изменений осуществлять прорыв через вязкую трясину статус-кво, заключается в закладывании под изменения прочного фундамента — ясного и убедительного ви́дения будущего, которое придает всем мощный стимул к его скорейшему воплощению в реальность и импульс к движению вперед.

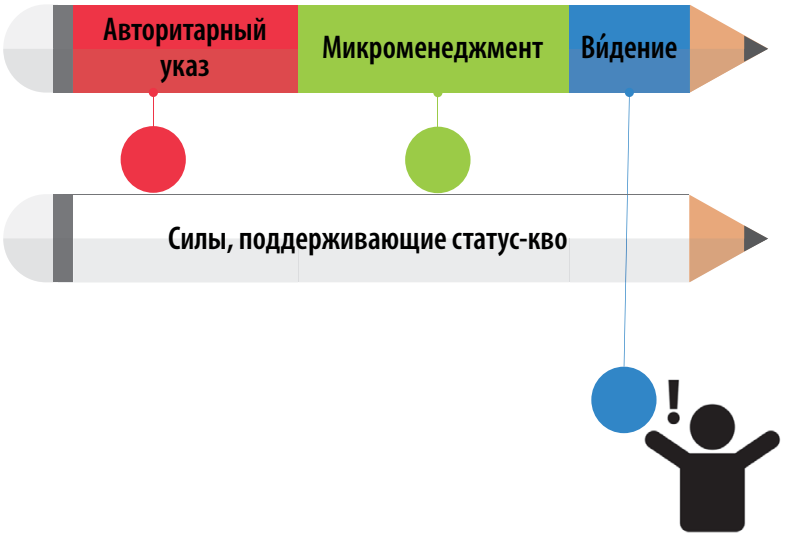


Рисунок 116. Ви́дение будущего сквозь настоящее

5.3.1 Причины жизненной важности ви́дения

Ви́дение — это, по сути, четкая картина будущего с явными или неявными объяснениями причин, по которым людям за это будущее можно и нужно бороться. Хорошее ви́дение выполняет сразу три важные задачи, обеспечивая ясность, мотивацию и согласованность целей и действий.

- **Ясность.** Хорошее ви́дение снимает любые вопросы относительно общего направления и конечной цели изменений, а также весьма упрощает принятия широкого спектра более детализированных решений, поскольку задает ключевые критерии и параметры. Действенное ви́дение (будучи подкреплено верными стратегиями) помогает разрешать практически любые разногласия относительно направления или неясности относительно мотивировки и драйверов изменений. Можно избежать бесконечных дебатов, просто задав вопрос: «Предлагаемая мера укладывается в общую линию ви́дения?» Аналогичным образом можно использовать ви́дение и для быстрой очистки от всякого хлама, чтобы у команды появилась возможность сосредоточить все усилия на приоритетных проектах преобразований.
- **Мотивация.** Ясное ви́дение побуждает людей двигаться в верном направлении, даже если первые шаги даются им лично нелегко или даже болезненно. Особенно это относится к организациям, где людей регулярно сдергивают с насиженных мест. Перед лицом мрачных, безрадостных и деморализующих ближайших перспектив лишь картина светлого будущего способна придать людям сил не прекращать борьбу за его материализацию.
- **Согласованность.** Убедительное и неотразимое в своей привлекательности ви́дение будущего помогает людям держать равнение на него, — то есть, по сути, обеспечивает согласованность и скоординированность коллективных действий на уровне мотивации. Альтернативой такому решению может стать разве что унылая череда детализированных указаний или бесконечных совещаний. Опыт показывает, что без разделяемого всеми интуитивного понимания верного направления группа взаимозависимых людей рискует погрязнуть в конфликтах и зациклиться на бесконечных согласованиях.

5.3.2 Природа и характер эффективного ви́дения

Ви́дение может быть весьма простым и приземленным. Оно не должно быть грандиозным или всеобъемлющим. Ви́дение — всего лишь одно из средств, имеющихся в арсенале системной технологии проведения изменений, наряду со стратегиями, планами, бюджетами и т. п. Тем не менее наличие ви́дения имеет крайне важное значение, поскольку именно оно заставляет команды фокусироваться на осязаемых улучшениях.

Ключевыми характеристиками эффективного ви́дения являются следующие.

- **Образная наглядность**: передается убедительная картина желанного будущего.
- **Заманчивость**: ви́дение апеллирует не к сиюминутным прихотям, а к *долгосрочным* интересам сотрудников, клиентов, акционеров и прочих заинтересованных лиц.

- **Осуществимость**: подразумеваемые ви́дением цели реалистичны и достижимы.
- **Сфокусированность**: картина достаточно отчетлива, чтобы ею можно было руководствоваться в принятии решений.
- **Гибкость**: картина представлена в достаточно общем плане, чтобы не стеснять личной инициативы и не препятствовать выработке альтернативных решений в случае изменения условий или возникновения затруднений.
- **Доходчивость**: поделиться ви́дением или передать его словами можно за считаные минуты.

Первый ключевой критерий проверки ви́дения на действенность: насколько просто представить себе такое состояние и достаточно ли оно заманчиво? Хорошее ви́дение может даже требовать жертв, но обязано обеспечивать соблюдение интересов *всех* людей, вовлеченных в процесс, в долгосрочной перспективе. Ви́дение, не сфокусированное на долгосрочных выигрышах для всех, рано или поздно будет отвергнуто. Помимо этого, ви́дение должно уходить корнями в реалии рынка продуктов или услуг, на которых специализируется организация. На большинстве рынков главная реалия заключается в необходимости постоянно заботиться об удовлетворении нужд конечных потребителей.

Ключевые вопросы, которыми надлежит задаваться:

- Если это сбудется, как это скажется на потребителях (внутренних и внешних)?
- Если это сбудется, как это скажется на акционерах? Станут ли они счастливее? Получат ли дополнительные дивиденды в долгосрочной перспективе?
- Если это сбудется, как это скажется на сотрудниках? Станет ли обстановка на рабочих местах лучше, благоприятнее и спокойнее? Откроются ли дополнительные возможности для роста и самореализации? Удастся ли стать более привлекательным работодателем?

Другим ключевым критерием является стратегическая осуществимость ви́дения. Для реализуемости ви́дения одних благих намерений недостаточно. Реализация может потребовать напряжения сил и ресурсов, но люди должны знать, что желанная цель достижима. Осуществимость отнюдь не подразумевает легкости реализации. Напротив, ви́дение должно ставить перед людьми весьма трудные задачи и тем самым побуждать их к фундаментальному переосмыслению происходящего. Вне зависимости от выбора целей, они должны быть, с одной стороны, достаточно амбициозными, а с другой — обоснованными, то есть организация должна согласовывать ви́дение с рациональным пониманием рыночных тенденций и пределов своих ресурсных возможностей.

Ви́дение должно быть достаточно сфокусированным, чтобы задавать людям ориентиры, но при этом не слишком детерминированным, иначе оно будет связывать сотрудников по рукам и ногам, провоцируя на всё более иррациональные модели поведения. Зачастую оптимальным подходом является сохранение простоты общего ви́дения как такового, но с внедрением в его формулировку множественных конкретизирующих привязок основных ситуаций, требующих

принятия решений, к внешне простой, но основополагающей ценности ви́дения, которое тем самым превращается в главную точку отсчета. Пример:

Мы ставим целью за пять лет выйти в мировые лидеры в своей отрасли. Под лидерством мы понимаем поднятие эффективности управления информацией до уровня, обеспечивающего неуклонное наращивание оборота и прибылей, а также превращения работы у нас в самое благодарное в человеческом понимании занятие. Для достижения столь амбициозной цели нам потребуется прочный фундамент доверия к нашей способности к принятию решений, ясность внутренних и внешних коммуникаций, улучшение понимания информационного ландшафта среды, в которой мы работаем, и разумно обоснованные инвестиции в надлежащие инструменты и технологии поддержки взращиваемой нами информационно-ориентированной культуры и этики, — культуры, вселяющей доверие и вызывающей восхищение у акционеров, клиентов, сотрудников и сообществ.

5.3.3 Создание эффективного ви́дения

Коттер рекомендует вырабатывать формулировку ви́дения итерационным методом, что позволяет эффективно и четко сформулировать все основные элементы. Основные этапы, компоненты и принципы таковы.

- **Первый черновой вариант.** Кому-то одному, в достаточной мере владеющему ситуацией и языком, предлагается составить исходный проект формулировки лаконичного заявления, отражающего потребности рынка.
- **Вклад руководящей коалиции.** Руководящая коалиция совместными усилиями выверяет заготовку таким образом, чтобы переработанный вариант в полной мере вписывался в более широкие стратегические перспективы.
- **Коллективная доработка.** Процессы, затрагивающие широкий круг лиц, без командной работы успешной реализации не поддаются. Стимулируйте людей подключаться к работе и вносить свой вклад.
- **Приятие и умом, и сердцем.** Аналитическое мышление и «заоблачные мечтания» должны гармонично дополнять друг друга, а не являться взаимоисключающими на протяжении всей проработки ви́дения.
- **Неупорядоченность.** Никаких регламентированных процедур, а, напротив, открытые дебаты, столкновения мнений, коррективы, изменения и даже полные переработки концепции преобразований. Если этого нет, значит, что-то не так с ви́дением или командой.
- **Временны́е рамки.** Выработка ви́дения — не разовое мероприятие. Могут потребоваться недели, месяцы и даже годы, чтобы прийти к единому мнению. А в идеале ви́дение должно еще и динамично (но не резко) эволюционировать сообразно развитию текущей ситуации.
- **Конечный продукт.** Четкая и лаконичная, емкая и гибкая формулировка желательного и достижимого будущего состояния, которую можно изложить и разъяснить максимум за пять минут.

Рисунок 117. Лидерство и управление: контрастные разграничения функций

5.4 Донесение ви́дения изменений до всеобщего понимания

Ви́дение начинает по-настоящему работать лишь после того, как все заинтересованные в переменах стороны вырабатывают общее понимание их целей, направления и общей картины желаемого конечного результата. Широко распространенные проблемы с донесением ви́дения до всеобщего понимания включают следующее.

- **Недостаточное внимание**, уделяемое информационно-разъяснительной работе.
- **Скудость разъяснений.** Расплывчатые формулировки вымывают из описания планируемых изменений всякое представление об их срочности и актуальности, а в результате все разъяснения проходят впустую.
- **Недостаточная разноплановость.** Менеджеры по определению обучены спрашивать с подчиненных и отчитываться перед начальством, то есть мыслить категориями вертикального подчинения. Лидерам же нужно выходить за рамки строгой иерархии и налаживать двустороннюю связь с самыми широкими кругами интересантов. А для поддержки столь разноплановой коммуникации от них требуется еще и четкое понимание круга актуальных проблем и возможных способов их решения.

Еще одна трудность возникает при неизбежном решении вопросов согласования позиций различных заинтересованных сторон, включая руководящую коалицию и рабочую группу, отвечающую за практическую реализацию планируемых реформ. Руководящая коалиция зачастую тратит

массу времени на проработку предварительных вариантов ответов на всевозможные вопросы понапрасну, поскольку окончательное решение так или иначе спускается на низший уровень (например, страницы сбора отзывов, вопросов и ответов, комментариев к проектам решений и т. п.). Результатом подобного подхода становится перегруженность информацией, что затмевает ви́дение цели и, как следствие, сеет панику и всплеск сопротивления изменениям (к счастью, эти эффекты носят преходящий характер).

Если принять во внимание тот факт, что в среднестатистической организации разъяснения, касающиеся изменений, занимают немногим более 0,5% от общего объема сообщений, доходящих до сотрудника, становится вполне понятным, почему недопустимо ограничиваться простой рассылкой информации об изменениях: она попросту затеряется среди общей массы и не сработает. Поэтому сообщение с ви́дением изменений нужно коммуницировать по-особенному и предельно доходчиво.

Коттер выделяет семь ключевых условий действенного донесения ви́дения до целевой аудитории.

- **Простота.** Убирайте любые профессиональные жаргонизмы, канцеляризмы, сложноподчиненные предложения.
- **Метафоричность, образность и иллюстративность.** Лучше один раз увидеть, чем сто раз услышать. Следует обрисовать словами (или даже инфографикой) наглядную картину будущего, используя по мере надобности аналогии, параллели и примеры.
- **Многоканальность.** Сообщение нужно доносить через всевозможные каналы, используя любые средства коммуникации — от агитации в лифте до объявлений по громкой связи, от мини-собраний до общеорганизационных форумов.
- **Повторение — мать учения.** Идеи воспринимаются и усваиваются, лишь будучи услышанными множество раз подряд.
- **Лидеры как образец для подражания.** Поведение авторитетных людей должно в полной мере соответствовать заявляемому ви́дению, то есть они должны быть достойным примером для подражания. Если у лидеров слова расходятся с делом, это разом перечеркивает результаты всех усилий по донесению ви́дения до сознания целевой аудитории.
- **Разъяснение причин явных несоответствий.** Алогизмы и несвязности, двусмысленности и противоречия подрывают доверие ко всему посланию целиком.
- **Двусторонняя связь.** Умение не только высказываться, но и прислушиваться к мнению целевой аудитории многократно усиливает эффективность коммуникации.

5.4.1 Примеры в контексте информационного управления

В информационном управлении нечеткость определения, неясность или неубедительность формулировки и неумение доносить ви́дение изменений до понимания аудитории — регулярно наблюдаемые явления, особенно в рамках инициатив по развертыванию новых систем или функционалов с акцентом прежде всего на технологические аспекты новинок. При отсутствии

понимания практических достоинств или потенциальных выгод от новых средств, технологий или методов обработки информации не исключено активное сопротивление сотрудников переходу на новые средства и методы работы.

Например, если в организации внедряются процессы управления документами и контентом на основе метаданных, пользователи в бизнес-подразделениях могут попросту саботировать выполнение требований, если предварительно не объяснить им, зачем нужно маркировать документы тегами метаданных или классифицировать записи. То есть нужно донести до их понимания ясное ви́дение пользы управления документами и контентом на основе метаданных для организации и лично для них. В противном случае ценная со всех других точек зрения инициатива не будет поддержана ввиду неприятия и невыполнения новых требований.

5.4.2 Чем проще, тем лучше

Трудно испытать эмоциональный подъем от ознакомления с формулировками на запутанном или невнятном языке.

Не нужно далеко ходить за примерами проблем со взаимопониманием, проистекающими от неумения выражаться просто и ясно. Пример подобной формулировки ви́дения:

«Наша цель заключается в снижении сроков устранения неполадок, с тем чтобы продемонстрировать, что они существенно ниже, чем у всех наших основных конкурентов на ключевых географических и демографических рынках. Аналогичным образом, мы настроены на снижение целевых показателей времени разработки новых продуктов, сроков обработки заказов и, в целом, на совершенствование всех рабочих показателей по другим относящимся к нашим клиентам направлениям для изменения».

Перевод на простой язык: «Мы будем работать быстрее и лучше всех в отрасли, действуя в интересах потребителей».

Чем проще формулировка ви́дения, тем понятнее и работникам, и заинтересованным лицам, и клиентам смысл предстоящих изменений, их влияние лично на них и роль, которая им отводится в новой картине. К тому же, усвоив смысл грядущих позитивных изменений, они еще и помогут распространить хорошие новости в своих кругах.

5.4.3 Многообразие коммуникационных форумов и каналов

Чтобы действенно коммуницировать ви́дение, лучше использовать побольше различных каналов связи. Причин тому много, начиная с перегруженности некоторых каналов иной информацией или «багажом наследия» предыдущих реформаторских инициатив и заканчивая тем фактом, что разные люди по-разному воспринимают и интерпретируют информацию, поступающую из различных источников. Если люди будут получать одно и то же сообщение по всем каналам, вероятность того, что ви́дение будет услышано, усвоено и воспринято как руководство к действию, возрастет многократно. В связи с этим «многоканальный/мультиформатный» подход просто необходим для вдалбливания ви́дения в коллективное сознание методом повторения вплоть до полного усвоения.

5.4.4 Повторение — всему голова

Во многих отношениях ви́дение изменений и идеи для изменений подобны речному потоку, запертому перекрывшей русло обрушившейся скальной породой. Вода далеко не сразу прорвется через такую плотину (если только сверху не сойдет мощнейший паводок, но в таком случае и последствия наводнения бывают разрушительными), однако со временем в результате множества циклов эрозии вода непременно подточит камень и пробьет себе новое русло в обход скалы.

В точности так же и реформаторские инициативы должны включать множество итераций повторения и пересказа ви́дения изменений на всевозможных площадках, в различных форматах и вариациях, пока перед всеобщим мысленным взором не предстанет зароненный в сознание по-настоящему прочный образ неизбежных изменений.

Ниже описаны сценарии двух возможных подходов. Который из них, по-вашему, является более эффективным?

- **Сценарий 1.** Высшее руководство выложило в корпоративную сеть «Видеообращение ко всем сотрудникам», дополнив массовой рассылкой сообщений на мобильные телефоны с просьбой «непременно ознакомиться с этим важным анонсом грядущих изменений». В самом видеообращении гендиректор уведомляет сотрудников о том, что планируется реорганизация, а все дополнительные детали и инструкции будут им передаваться через непосредственных начальников. После этого за полгода в сети публикуется три статьи о ви́дении изменений, а кроме того, на ежеквартальной конференции руководства проводится специальное инструктивное заседание. Итого: шесть сеансов односторонней связи без прорисовки осязаемых деталей.
- **Сценарий 2.** Высшее руководство будет четыре раза в день собираться для краткого обсуждения и обмена мнениями по поводу хода реформ и отображения достигнутого прогресса в контексте ви́дения. Каждый из членов руководящего совета, в свою очередь, поставит перед прямыми подчиненными задачу также четыре раза в день находить время для обсуждения и потребовать такого же четырехразового ежедневного обсуждения от собственных подчиненных, — и так далее до самых низовых сотрудников. Так что теперь Фрэнк на всяком совещании своего отдела разработки продуктов просит подчиненных отчитываться о проделанной работе и представлять ближайшие оперативные планы в контексте ви́дения. И работающая у Фрэнка в отделе технолог Мэри всякий раз, отчитываясь о текущем статусе своего проекта, делает это исключительно в привязке к ви́дению. И внутренний аудитор Гарри, докладывая о выявленных нарушениях, особо указывает на их негативное влияние на перспективы претворения в жизнь ви́дения. И так на всех направлениях работы и уровнях управления: у любого менеджера имеется бессчетное число возможностей на протяжении всего календарного года высказываться по поводу текущей ситуации в привязке к направлению, определяемому ви́дением[1].

[1] Собственно, в этом и состоит «принятие новой философии» и «учреждение института лидерства» — то есть два ключевых *пункта трансформации* в *управлении качеством* в модели У. Эдвардса Деминга. — *Примеч. пер.*

5.4.5 Слово и дело

Силу личного примера, демонстрируемого руководством, заменить нечем. Только будучи подкрепленными делами, заявления о нужности изменений, затрагивающих различные аспекты организационной культуры, обретают ценность. Даже если бы и не было для того иных причин, соблюдение старшими по должности менеджерами принципа «сказано — сделано» порождает легенды о притягательности ви́дения с переходом в обсуждение достоинств замысла, а это исключительно мощное орудие воздействия на менталитет.

Верно и обратное: если ваши поступки противоречат вашим же словам, вы тем самым дискредитируете не только и не столько себя, сколько замысел реформ, посылая людям явное сообщение, что лично для вас ви́дение и планы преобразований не важны, а значит, и они вправе относиться к этому наплевательски. Ничто другое не подрывает ви́дение изменений и усилия по его материализации столь же сильно и необратимо, как пример со стороны кого-либо из членов руководящей коалиции, поступающего несообразно словам.

5.4.6 Сила примера в контексте информационного управления

В сфере управления данными «пустословие» руководства может проявляться в самых элементарных вещах. Пример: начальник отдела отправляет персональные данные клиентов прямо по электронной почте во вложенном файле безо всякого шифрования или защиты паролем в нарушение действующих правил информационной безопасности, причем никаким санкциям не подвергается.

И простейший пример противоположного свойства: инициативный комитет по внедрению руководства данными всем составом строго придерживается принципов и правил, соблюдения которых требует от остальной организации, и наравне со всеми подвергается проверкам в рамках собственными силами реализованных мер по надзору за соблюдением правил обращения с информацией, отчетности и по своевременному выявлению и устранению проблем и сбоев.

Рассмотрим также пример реализации проекта внедрения управления метаданными. Прежде всего рабочая группа должна применить разработанные стандарты метаданных к собственной проектной документации. Зачем? Во-первых, просто для того, чтобы убедиться в практичности и удобстве применения собственных разработок перед распространением нововведения на всю организацию; а во-вторых, и еще для демонстрации другим полезности, удобства и прочих преимуществ надлежащей маркировки и классификации всех записей и документов, хранящихся в информационных системах.

5.4.7 Разъяснение противоречий

В некоторых случаях логические противоречия, рассогласованность или нестыковки оказываются неизбежными. Может, к примеру, возникнуть ситуация, когда по тактическим или оперативным соображениям, или же просто ради того, чтобы хоть как-то сдвинуть дело с мертвой точки в рамках общеорганизационной системы, проводнику изменений приходится предпринимать шаги, идущие вразрез с заявленным ви́дением. Когда такое происходит, проблема требует

обязательного, но предельно аккуратного решения во избежание ущерба для ви́дения, — возможно, и не без элемента «театрализации». Примером несоответствия, вызывающего вопросы, может служить приглашение внешних консультантов на фоне заявленного намерения полагаться на собственные силы или, хуже того, курса на экономию и сокращение численности штатных сотрудников. «Как так? — спросят люди. — Нам выделяют пачку офисной бумаги в неделю на весь отдел, поскольку денег нет. А на этих экспертов из [название консалтинговой фирмы] в дорогих костюмах деньги нашлись?!» Из столь неловкой ситуации есть два выхода. Первый убьет ви́дение, причем гарантированно, а второй оставит неплохие шансы побороться за возвращение ситуации в конструктивное и контролируемое русло.

Первый вариант — игнорировать проблему как таковую или же пытаться парировать и гасить недоуменное недовольство защитными реакциями. Неизбежной развязкой при таком сценарии рано или поздно станет унизительная сдача позиций путем устранения зримого противоречия, — и далеко не факт, что этим удастся компенсировать урон, нанесенный долгосрочным целям преобразований.

Второй вариант — живо заинтересоваться вопросом и подобрать достаточно разумное и веское обоснование необходимости временно примириться с кажущимся нарушением логики. Объяснение должно быть простым, ясным и честным. Например, всё в том же случае с дорогими консультантами на фоне дефицита ресурсов можно предложить людям следующее объяснение.

«Мы вполне отдаем себе отчет в том, что выглядим странно, не жалея денег на услуги консультантов на фоне экономии на всём остальном ради достижения нашего ви́дения устойчиво прибыльной коммерческой фирмы. Однако для того, чтобы научиться стабильной экономии средств не на бумаге, а на деле, нам нужно вырваться из плена старых представлений и привычного образа мыслей — и освоить новые навыки, и это требует инвестиций в знания. А поскольку внутри организации мы, увы, не располагаем источниками требующихся знаний, нам приходится в краткосрочной перспективе нести дополнительные расходы на их приобретение, а заодно использовать эту возможность для накопления собственного запаса знаний на будущее. Каждый консультант привлечен к строго определенному проекту. И каждой проектной группе дано задание выведать как можно больше полезного о своей новой функции у консультанта. За счет этого мы и обеспечим себе задел устойчивого совершенствования работы по всем направлениям на годы вперед!»

Главный ключ — дать прямое и открытое объяснение противоречия, из которого следует, что оно носит локальный характер и вполне обоснованно с точки зрения долгосрочных интересов.

5.4.8 Пример в контексте информационного управления

На примере объяснения несоответствий очень удобно иллюстрировать важность роли тщательно проработанных моделей руководства данными, дополненных согласованными всеми заинтересованными сторонами протоколами принятия решений, и продвигать идею обязательного административного оформления любых временных исключений из правил с указанием сроков действия таких поблажек.

Например, установлен административный запрет на любые испытания в производственной среде, но проекту без этого никак не обойтись, поскольку ни в какой «песочнице» объем и структуру «живых» данных, которые необходимы для проверки алгоритмов сопоставления или очистки, воспроизвести не удастся. В таком случае должно последовать ясное, четкое и открытое объяснение причины согласия на разовое нарушение установленного стандарта. Реализуемо это только через надлежащие механизмы высокоуровневого надзора. Если же подобный проект начнет экспериментировать в режиме реального времени с рабочими данными *без* предварительного согласования и официальной санкции после надлежащей оценки рисков, проектировщики должны понести неотвратимое дисциплинарное наказание (по принципу соответствия «слова и дела») или быть официально оправданы, а причина освобождения их от санкций ясно и четко разъяснена и положена в основу нового правила, явным образом разрешающего тестирование в рабочей среде во всех подобных ситуациях.

5.4.9 Умейте слушать — и будете услышаны

Стивен Кови[1] советует всем, кто стремится к высокой эффективности: «Старайтесь сперва понять и лишь затем быть понятыми». Иными словами, возьмите за правило прислушиваться к мнению других, если хотите, чтобы они прислушивались к вашему (Covey, 2013).

Зачастую у руководящей группы формируется не вполне верное ви́дение — и возникают неожиданные препятствия или затруднения, которых вполне можно было бы избежать при лучшей информированности о реальном положении вещей. В целом, недостаточная информированность чревата крайне дорогостоящими ошибками и ослаблением привлекательности ви́дения и нацеленности на его претворение в реальность. Двусторонние диалоги — жизненно важный метод выявления и устранения предметов озабоченности, мешающих людям принять изменения и/или поверить в достижимость предлагаемого им ви́дения будущего. В связи с этим голос клиента должен иметь не менее решающее значение при определении и развитии ви́дения, чем, скажем, при оценке качества данных. Если каждая беседа, любой обмен мнениями, всякий контакт и диалог рассматриваются еще и в качестве очередной возможности для обсуждения ви́дения и сбора отзывов и предложений, то можно и без формальных собраний заручиться результатами тысяч человеко-часов обсуждения перспектив развития и эффективных средств воплощения выработанного на их основе ви́дения в реальность.

5.4.10 Пример в контексте информационного управления

В интересующем нас контексте двусторонний обмен мнениями лучше всего иллюстрируется сценарием, при котором функция ИТ-обеспечения искренне считает, что доступ ко всем данным, необходимым ключевым бизнес-подразделениям, предоставляется своевременно и в полном

[1] Стивен Кови (*англ.* Stephen R. Covey, 1932–2012) — американский консультант по вопросам руководства и управления, а также личностной самоорганизации. Цитируемая здесь книга «Семь навыков высокоэффективных людей» в 2011 г. составила достойную компанию конспективно представленному в разделах 4 и 5 труду Дж. Коттера в «топ-25 влиятельнейших книг по бизнес-управлению всех времен» по версии журнала *TIME*. — *Примеч. пер.*

объеме, однако руководители этих бизнес-подразделений постоянно жалуются на задержки с получением жизненно необходимых для работы данных и от отчаяния даже занялись самодеятельностью и своими усилиями построили худо-бедно пригодные времянки для отправления нужд формирования отчетности и выставления цифр в витрины данных, работающие по принципу импорта из неведомо откуда взятых, зато «свежих» электронных таблиц.

Выработка ви́дения перспектив совершенствования информационного управления и распоряжения данными невозможна без выявления, изучения и устранения разрывов в восприятии происходящего функцией ИТ-обеспечения и бизнес-пользователями. Без выработки общего понимания реалий информационной среды невозможно избежать череды неминуемых сбоев, ошибок и отказов, не говоря уже о том, чтобы заручиться широкой базой всеобщей поддержки долгосрочной программы модернизации всего комплекса ИТ-систем, необходимой для устойчивого развития бизнеса.

6. ФОРМУЛА ИЗМЕНЕНИЙ

Один из самых популярных методов обеспечения эффективности изменений — так называемая *формула Глейчера*[1] — задает необходимое условие преодоления сопротивления изменениям и выглядит следующим образом:

$$C - (D \times V \times F) > R$$

Согласно данному уравнению, изменения (Change) становятся возможными лишь при условии достаточности векторного произведения движущих импульсов — неудовлетворенности (Dissatisfaction) текущим состоянием, ви́дения разумной и привлекательной альтернативы (Vision) и понимания первых шагов (First steps), необходимых для выхода из неблагоприятной ситуации, — для преодоления силы сопротивления изменениям (Resistance) в организации.

Способность влиять на значение любой из четырех переменных в правой части уравнения Глейчера в сторону их приращения способствует повышению эффективности изменений. Однако, как бывает в случае задействования сложных и не до конца исследованных механизмов, важно сознавать риски бездумного нажимания кнопок и дергания рычагов во избежание получения совершенно непредсказуемых эффектов. Для этого следует внимательно

[1] «Формула» (которую логичнее было бы назвать матричным уравнением с неустановленным числом неизвестных и параметров) названа в честь некоего автора по имени Дэвид Глейчер (*англ.* David Gleicher) из консалтинговой фирмы Arthur D. Little, якобы подсказавшего в 1960-х годах эту формулу впоследствии опубликовавшему ее одному из основоположников теории организационного управления профессору бизнес-школы MIT Ричарду Бекхарду (*англ.* Richard Beckhard, 1918–1999). Приведенная версия является продуктом творческой доработки исходной ($C = A \times B \times D > X$), где буквенные обозначения параметров (кроме C = Change) особой смысловой нагрузки не несли. — *Примеч. пер.*

и систематически вести работу еще и по следующим направлениям управления изменениями и контроля их результатов.

- Выявление как локальных всплесков недовольства, так и устойчивых тенденций к росту неудовлетворенности положением дел внутри организации — мощное средство обнаружения проблем, равно как и удовлетворения насущных нужд подопечных; главное тут — предельная аккуратность и дозированный характер уступок; иначе сопротивление переменам будет только нарастать и усиливаться.
- Для выработки согласованного ви́дения будущего требуется четкое, конкретное и живое ви́дение того, что именно людям придется делать совершенно по-новому, от чего придется отказаться, в чем придется частично переучиваться, что придется освоить с нуля и т. д. Сделайте же так, чтобы люди в полной мере сознавали новые знания, навыки, методы и подходы к работе, которых потребуют от них новые реалии. Но при этом важно представить им новые требования таким образом, чтобы они не выглядели пугающими и непреодолимыми, дабы не порождать тяги к выстраиванию политических препятствий на пути изменений во имя сохранения статус-кво.
- При описании первых шагов на пути изменений убедитесь в их реализуемости и достижимости результатов в прямой привязке к ви́дению.
- Принимайте все необходимые меры по снижению и эскалации сопротивления изменениям. Изучайте интересы всех заинтересованных сторон.

7. ДИФФУЗИЯ ИННОВАЦИЙ И ПОДДЕРЖАНИЕ ИЗМЕНЕНИЙ

Наконец, нужно наладить постоянно действующую систему информационно-разъяснительной работы, обучения и подготовки кадров с целью обеспечения стабильно высокого качества информации и управления данными в масштабах организации. Для реализации этого компонента общеорганизационной реформы требуется понимание динамики и механизмов распространения новых идей. Этот аспект изменений в теории управления известен под названием «диффузия инноваций».

Теория диффузии инноваций ставит своей задачей объяснить, как, почему и какими темпами распространяются новые идеи и технологии в различных культурных средах. Понятие диффузии инноваций было сформулировано в 1962 году Эвереттом Роджерсом[1], а массовую популярность приобрело после публикации книги Сета Година[2] «Идея-вирус». С тех пор моделирование

[1] Эверетт Роджерс (*англ.* Everett M. «Ev» Rogers, 1931–2004) — американский социолог и специалист по теории средств массовой коммуникации, декан журфака Университета штата Нью-Мексико. — *Примеч. пер.*

[2] Сет Годин (*англ.* Seth Godin, p. 1960) — американский предприниматель и экономист, специалист в области информатики и сетевого маркетинга, автор концепции доверительного маркетинга. — *Примеч. пер.*

диффузии инноваций широко применяется в самых различных сферах человеческой деятельности — от фармакологии до животноводства, не говоря уже о всяческой бытовой радиоэлектронике и компьютерной технике.

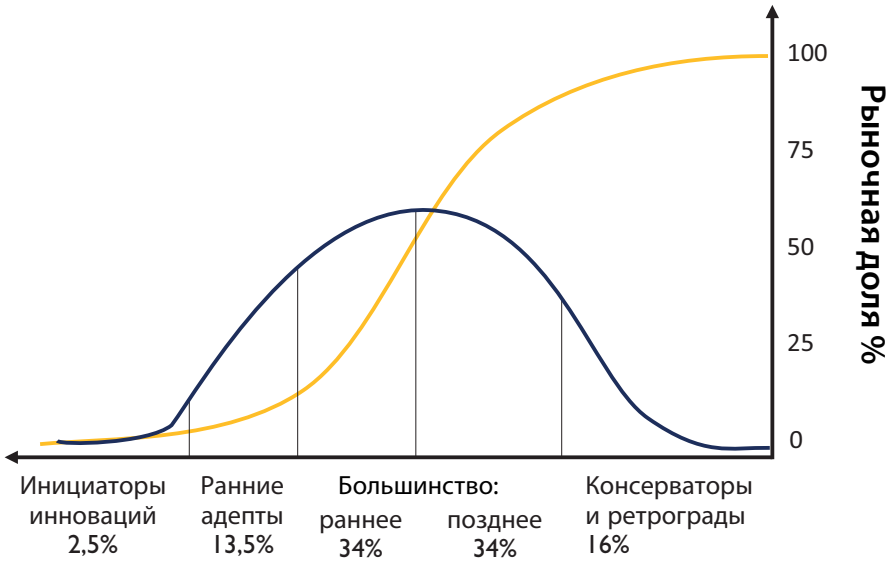


Рисунок 118. Диффузия инноваций по Эверетту Роджерсу

Теория диффузии инноваций утверждает, что изменения начинаются по инициативе мизерного меньшинства популяции (2,5%). Инициаторы инноваций — это, как правило, молодежь с высоким социальным статусом и без финансовых проблем (то есть значительно моложе, образованнее и богаче среднего уровня по исследуемой популяции). Принадлежность к высокому социальному классу и надежные финансовые тылы позволяют новаторам не опасаться возможных последствий и убытков вследствие ошибочного выбора. Действуют они в прямом контакте с разработчиками инновационных технологий, что придает им дополнительное ощущение защищенности от риска. Следом за ними новшество воспринимают ранние последователи, доля которых составляет 13,5% популяции. По социально-демографическим характеристикам ранние последователи мало отличаются от инициаторов инноваций, но более уязвимы и чувствительны к рискам, а потому и выжидают некоторое время, прежде чем присоединиться к первопроходцам. При этом ранние последователи прекрасно отдают себе отчет в том, что правильный выбор поможет им оказаться в центре всеобщего внимания и снискать дополнительное уважение в обществе. Ну а после этого накатывает и основная волна — и новшество воспринимается ранним и поздним большинством (68%). Последними переходят на инновационное решение — отстающие (рис. 118 и табл. 37).

Таблица 37. Восприятие нового в рамках модели диффузии инноваций применительно к информационному управлению[1]

Категория	Определение (в контексте информационного управления)
Инициаторы инноваций	Первыми определяют новый подход к решению проблем качества информации. Не опасаясь рисков, смело профилируют и анализируют данные, экспериментируют со шкалами оценок и учатся переводить описания проблем, испытываемых бизнесом, на язык информационного управления. Новаторы вкладывают личные средства в приобретение необходимых информационных ресурсов и обучение недостающим для совершенствования работы профессиональным навыкам
Ранние последователи	Это вторая по скорости восприятия инноваций категория лиц, но именно ранние последователи обладают высочайшим авторитетом среди коллег и являются лидерами мнений, формируя тем самым условия для нарастания следующей волны массового восприятия новшеств. Пользуются репутацией «менеджеров-визионеров» (бывалых, многоопытных, берущих на себя ответственность за развитие экспериментальных областей бизнеса) и прекрасно сознают, что проблемы с качеством информации — их личные, поскольку препятствуют достижению успеха. Эта категория сотрудников берет за основу подход инициаторов инноваций, превращает его в полноценную бизнес-инициативу и переходит к формальному закреплению новой практики информационного управления
Раннее большинство	Первая волна массового восприятия инновации существенно запаздывает по сравнению с ранними последователями. Помимо замедленного по сравнению с предыдущей группой восприятия, для представителей раннего большинства характерны: социальный статус выше среднего, контакты с ранними последователями, а кое-кому и статус неформальных авторитетов в масштабах системы. Многие представители этой категории работают в «традиционных главных сферах деятельности» организации, где принято списывать убытки вследствие низкого качества данных на «издержки производства»
Позднее большинство	Представителями позднего большинства являются люди с весьма скептическим отношением к любым инновациям, лично воспринимающие их лишь после того, как на новинку перейдет как минимум половина коллег или соседей. К позднему большинству относятся обычно лица с невысоким социальным статусом, незавидным финансовым положением, контактами среди себе подобных и ранним большинством, а также крайне редко обладающие высоким авторитетом и репутацией лидеров. В переводе на язык информационного управления здесь могут подразумеваться организационные подразделения с ограниченным бюджетом, укомплектованные разочарованными или малообразованными сотрудниками, а это весьма скептичная и невосприимчивая к инновациям среда. Возможны и упорное неприятие, и даже саботаж изменений
Отстающие	Последние, кто принимает новые подходы. Ярко выраженного собственного мнения у этих людей, как правило, не бывает, как и склонности кому-то его навязывать, да и авторитетом они не пользуются вовсе. Но проводников изменений эти по преимуществу пожилые люди не переносят на дух, а всё новое воспринимают в штыки. «Традиции» — их главная (а нередко и единственная) ценность. В информационном управлении речь часто идет о сотрудниках подразделений, которые сопротивляются новым подходам

7.1 Главные трудности на пути распространения инноваций

Труднее всего дается преодоление двух ключевых барьеров на пути распространения инноваций на всю организацию. Первый прорыв состоит в том, чтобы перейти наконец со стадии восприятия новшеств одними ранними последователями на стадию массового восприятия. Для этого

[1] © 2014 Daragh O Brien. Используется с разрешения правообладателя.

требуется тщательное управление изменениями, чтобы помочь инициаторам и ранним последователям выявить достаточное число недовольных статус-кво среди коллег и настоятельно порекомендовать, а то и буквально навязать им инновационный подход. Без этого перелома не наступит, поскольку необходимо набрать «критическую массу» сторонников инновации для запуска цепной реакции ее массового восприятия и доминирования.

Вторая ключевая трудность приходится на точку фазового перехода от завершения массового восприятия поздним меньшинством к началу переубеждения отстающих. Команде реформаторов нужно смириться с тем, что наставить 100% популяции на новый путь — задача невыполнимая. Всегда остается какой-то процент целевой группы, который будет отчаянно сопротивляться изменениям до конца. Единственное, что остается делать руководству на данной стадии, — решить, каким именно образом положить конец пребыванию этой группы в организации.

7.2 Ключевые элементы диффузии инноваций

Рассматривая процессы распространения инноваций в рамках организации, следует учитывать четыре ключевых элемента.

- **Инновация**: идея, практика, технология или иной предмет, воспринимаемый как нечто новое или непривычное на индивидуальном /групповом уровне.
- **Коммуникационные каналы**: любые средства обмена информационными сообщениями между лицами и группами лиц.
- **Фактор времени**: темпы восприятия инновации членами социальной системы.
- **Социальная система**: множество взаимосвязанных штатных и структурных единиц, задействованных в совместном решении задач по достижению общей цели.

В контексте информационного управления инновацией может быть нечто весьма простое: например, сама идея введения в каждом подразделении должности распорядителя данных и его функции ответственного за данные, а на общеорганизационном уровне — совета распорядителей данных, с тем чтобы они регулярно решали общие проблемы управления данными коллегиально, а не в рамках традиционного узкофункционального мышления.

Процесс информирования сотрудников об этой инновации подразумевает необходимость выбора наиболее эффективных коммуникационных каналов и налаживание управления этими каналами и потоками сообщений.

Ну и, наконец, идея социальной системы как набора взаимосвязанных индивидуальных и групповых участников совместного предприятия живо напоминает понятие системы в трактовке У. Эдвардса Деминга, считавшего, что только систему и можно оптимизировать, поскольку по отдельности ее компоненты оптимизации не поддаются. Инновацию, не распространяющуюся за пределы бизнес-подразделения или проектной группы, никак нельзя считать хорошо диффундировавшим изменением.

7.3 Пять стадий восприятия инновации

На индивидуальном уровне восприятие изменений, как правило, происходит в пять шагов. Узнав об инновации (шаг 1), людям нужно убедиться в ее ценности и пригодности для их целей (шаг 2) и принять решение о целесообразности инновации (шаг 3). Если решено от инновации не отказываться, далее следуют: 4) внедрение или реализация; 5) подтверждение успешности восприятия изменений (табл. 38 и рис. 119).

Таблица 38. Стадии восприятия инноваций по Роджерсу (1964)

Стадия	Определение
1. Осведомленность	Человек узнаёт о существовании инновации, но не имеет достаточной информации о том, в чем именно она заключается, да и не особо стремится получить такую информацию
2. Убеждение	На стадии убеждения человек, заинтересовавшись инновацией, активно ищет и изучает информацию о ней
3. Решение	На этой стадии человек взвешивает все «за» и «против» использования инновации и выбирает решение — принять или отклонить ее. Из-за субъективности такого выбора эта стадия, по мнению Роджерса, труднее всего поддается эмпирическому изучению
4. Реализация	На стадии реализации происходит практическая проверка пригодности и полезности инновации, а также сбор недостающей дополнительной информации о ней с целью доработки прикладного решения
5. Подтверждение	По завершении доработки принимается окончательное решение о целесообразности продолжения использования инновации и возможности ее дополнительных усовершенствований с целью полнейшего раскрытия потенциала.

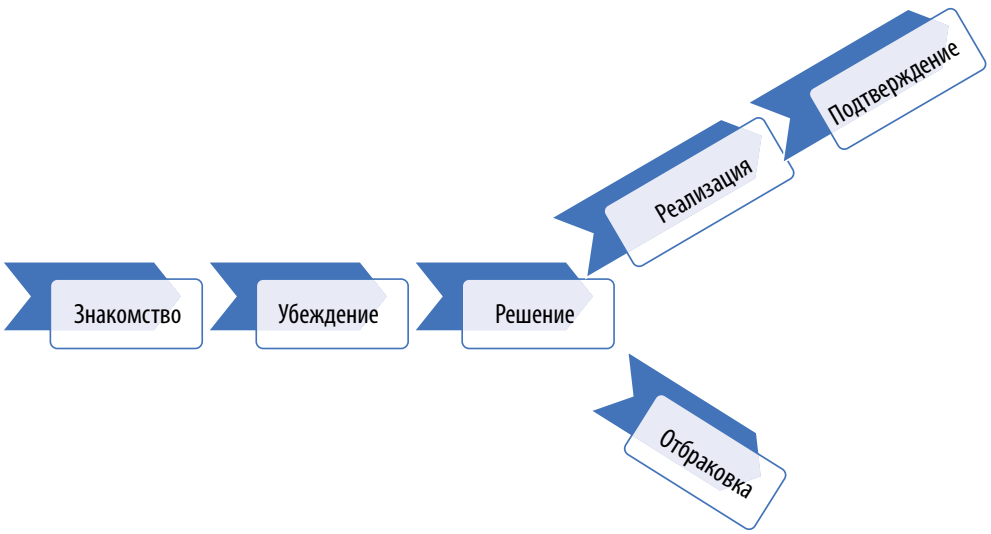


Рисунок 119. Стадии восприятия инноваций

Понятно, что психологически от новой идеи всегда проще отказаться, забраковав ее, нежели трудиться над ее внедрением в собственную практику. Именно поэтому столь большое значение в модели Роджерса придается переломному моменту достижения критической массы ранних последователей и раннего большинства.

7.4 Субъективные причины неприятия или отторжения инноваций и изменений

По большей части люди всё-таки руководствуются достаточно рациональными доводами в пользу принятия либо отклонения инновации или изменения. Ключевой критерий — это наличие или отсутствие у нового решения осязаемых преимуществ над старым.

Возьмем смартфон последнего поколения. Понятно, что он в чем-то лучше смартфонов предпоследнего поколения — функциональнее или проще в использовании, либо моднее и более стильно выглядит, либо поддерживает новейшие приложения, быстрее и проще обновляется. То же самое и с программными средствами, технологиями, платформами и методологиями управления данными: мы всё дальше уходим от такой архаики, как ввод с клавиатуры, «наколенное» программирование, трудоемкий поиск или выявление данных в ручном режиме.

Например, во многих организациях вполне возможно неприятие достаточно простых изменений в управлении документами и контентом просто из нежелания возиться с маркировкой файлов метаданными с целью поддержки контекстного поиска. А значит, нужно доходчиво и терпеливо разъяснять, что внедрение метаданных дает преимущество за счет повышения уровня информационной безопасности и защиты данных, автоматизации управления сроками хранения и плановым архивированием, упрощения и ускорения обработки поисковых запросов и т. д. Пусть сопоставят затраты времени на тегирование файлов с экономией времени на поиск или масштабами последствий утечки какой-нибудь закрытой информации, — это помогает по достоинству оценить сравнительные преимущества нового подхода.

Убедившись, что предложение об усовершенствовании будет принято, люди начинают задаваться вопросом о том, насколько нововведение вообще совместимо с их привычками, характером, образом жизни, привычным стилем работы и т. п. Возвращаясь к примеру со смартфоном, сочетание функций плеера, электронной почты, чата и т. д. с функцией обычного мобильного телефона означает изначально заложенную в эту концепцию совместимость с образом и стилем жизни целевых пользователей.

Оценивая совместимость, потребитель (сознательно или бессознательно) учитывает целый ряд факторов. Например: насколько сложно освоить новшество? Если инновация слишком сложна в практическом использовании, вероятность ее позитивного восприятия резко снижается. Опять же, на пути эволюции платформ смартфонов и планшетов были неудобства слишком сложного пользовательского интерфейса, пока цель в простоте использования всё же не была достигнута. А первые же удачи указали верный путь всем прочим разработчикам, переопределив господствующие на рынке ожидания и стандарты пользовательских интерфейсов мобильных устройств.

Следует учитывать также и еще два важных фактора.

Наличие возможности апробации позволяет потребителю оценить, насколько новое средство или технологическое решение просто и удобно в обращении, испытать его функциональность и поэкспериментировать с возможностями. Отсюда и массовое предложение пробных бесплатных версий, льготных периодов подписки и т. п. Возможность протестировать решение существенно повышает вероятность того, что пользователь оценит новое программное средство по достоинству. Значимость такого подхода еще и в том, что он позволяет понять и оценить относительные плюсы и минусы, протестировать новые технологии на совместимость не только с другими программными и аппаратными средствами, но и со своим образом жизни и стилем работы, и сделать заключение о целесообразности перехода. В качестве первых шагов на пути к ви́дению новой, преображенной организации вполне годятся методы итеративного прототипирования и тестовых испытаний новых решений совместно с ключевыми игроками.

Наблюдаемость — мера заметности и зримости изменений. Сделавшись видимой невооруженным глазом, инновация начинает сама себя продавать через сети формального и неформального личного общения. А затем может последовать и запуск цепной реакции — либо приятия, либо отторжения. Заранее планируйте дальнейшие действия на случай негативной обратной связи. Наблюдение за тем, *как именно* люди используют новую технологию или работают с информацией (допустим, какие средства визуализации традиционно считавшихся «сухими» цифр предпочитают), может подсказать, например, каким средствами лучше воспользоваться для сбора отзывов.

8. ОБЕСПЕЧЕНИЕ ПОДДЕРЖКИ ИЗМЕНЕНИЙ

Для запуска преобразований требуется ясное и убедительное ви́дение целей, незамедлительные первые шаги в задуманном направлении, чувство неотложности изменений или неудовлетворенности существующим положением вещей, руководящая коалиция и план лавирования между ловушками и препятствиями, подстерегающими проводников изменений в начале пути.

Однако в случае инициатив по модернизации информационного управления (например, по внедрению программ управления данными) имеется еще одна специфическая проблема: как правило, подобные проекты инициируются в качестве ответной реакции на появление какого-то конкретного нового стимула или, напротив, симптома неблагополучия в организации. А вот после устранения симптома или решения бизнес-задачи чувство неотложности дальнейших изменений и неудовлетворительности текущей ситуации быстро улетучивается, а значит, программе совершенствования информационного управления становится всё труднее заручаться устойчивой политической или финансовой поддержкой, особенно в условиях конкуренции за бюджетные средства с другими проектами.

Детальный анализ подобных сложных ситуаций и их возможных разрешений выходит за рамки настоящей работы. Единственное, что уместно посоветовать в контексте свода знаний об управлении данными: обратитесь еще раз к принципам управления изменениями, достаточно

подробно описанным в настоящей главе, и глубоко проанализируйте текущую ситуацию и перспективы, — это наведет вас на верное решение.

8.1 Острота чувства неотложности или неудовлетворенности

Крайне важно поддерживать ощущение безотлагательности изменений. Следовательно, не менее важно оперативно улавливать очаги разгорающегося недовольства положением дел в организации и анализировать, как можно исправить или хотя бы нейтрализовать ситуацию средствами информационного управления.

Например, инициативу по обеспечению защиты конфиденциальных данных в соответствии с законодательными требованиями можно расширить, дополнив ее мерами по обеспечению качества и совершенствованию управления персональными данными. Новый компонент, в свою очередь, замыкается обратно на истоки инициативы, поскольку большинство законодательных актов о защите конфиденциальных данных включают и положения, регулирующие качество и права доступа, то есть, по сути, предписывают выявлять данные низкого качества. Однако именно этот факт позволяет вывести ви́дение программы управления данными на новый виток восходящей спирали, осознав, что после устранения базовых рисков утечки и разглашения чувствительных данных в нее можно и нужно «второй волной» интегрировать методы и практики обеспечения качества информации.

8.2 Формирование ви́дения

Распространенная ошибка заключается в непонимании различия между объемами работ по проекту и ви́дением измененного будущего. Для достижения ви́дения может потребоваться реализация множества проектов. Важно, чтобы ви́дение было достаточно широким для обеспечения возможности полномасштабных действий, а не создавало для лидеров реформ патовую ситуацию после сбора первого урожая «самых вкусных и низко висящих плодов».

Важно понимать колоссальную разницу между двумя типами формулировок ви́дения. Пример формулировки первого типа:

«Мы намерены разработать и внедрить тщательно структурированную модель управления персональными данными в строгом соответствии с „Общим регламентом“ ЕС».

Пример формулировки второго типа:

«Мы выйдем в отраслевые лидеры по числу воспроизводимых и масштабируемых методов управления критически важными информационными ресурсами, необходимыми для обеспечения рентабельности, минимизации риска, повышения качества обслуживания и выполнения наших этических обязательств по защите персональных данных».

Первая формулировка больше похожа на проектное задание. А вторая отражает именно ви́дение, поскольку задает общее направление и очерчивает рамки перспектив развития организации.

8.3 Состав руководящей коалиции

Ограничение членства в руководящей коалиции лишь теми заинтересованными фигурами, на которых организация управления данными сказывается самым непосредственным образом,

приведет к сужению рамок ви́дения и, как следствие, ограничению эффективности изменений. Опять же, как и в случае с формулировкой ви́дения, важно понимать разницу между проектными группами, отвечающими за получение в установленные сроки заданных результатов, и руководящей коалицией, осуществляющей общую координацию всех этих проектов и дальнейшую эволюцию ви́дения будущего организации.

8.4 Объективность и осязаемость улучшений

Даже в случае узкой фокусировки инициативы на решении некой прикладной задачи или совершенствовании отдельного аспекта управления данными в большинстве случаев выработанные в ходе ее реализации принципы, практические методы и инструменты могут быть затем перенесены на другие инициативы. Способность продемонстрировать работоспособность используемых подходов и методов с точки зрения получения объективных и осязаемых преимуществ по сравнению с результатами других организационных инициатив помогает руководящей коалиции расширять поле деятельности за счет выявления новых областей, требующих неотложного вмешательства или служащих источником неудовлетворенности клиентов, где предыдущие наработки вполне могут пригодиться.

Например, в обслуживающей компании ЖКХ методы профилирования и контроля качества данных, используемые при формировании ЕПД, можно без особых переделок перенести в систему контроля соблюдения правил начисления и оплаты услуг поставщиков коммунальных услуг. А затем, если интегрировать две эти системы, управляющая компания получит готовую систему управления качеством данных предприятия ЖКХ с системой балльных оценок и привязанных к ним инициатив по устранению недостатков. Немалая и очень наглядная польза, особенно если в той же конторе всего-то годом ранее все калькуляции, выписки и проводки счетов по умолчанию производились, проверялись и подчищались вручную.

9. ДОНЕСЕНИЕ ЦЕННОСТИ УПРАВЛЕНИЯ ДАННЫМИ ДО ВСЕОБЩЕГО ПОНИМАНИЯ

Для того чтобы помочь организации в полной мере осознать всю степень важности управления данными, нередко приходится вырабатывать формализованный комплексный план реорганизации информационного управления, учитывающий все аспекты преобразований, описанные в данной главе. Подобный план помогает организации научиться ценить как имеющиеся в ее распоряжении данные, так и вклад специалистов по управлению ими. Однако и после того, как программа управления данными сложилась и устоялась, она нуждается в устойчивой поддержке. Выработке понимания и получению устойчивой поддержки со стороны сотрудников способствует регулярная информационно-разъяснительная работа. Если каналы связи между программой и различными подразделениями организации работают в режиме двустороннего

обмена мнениями, планы обмена информацией могут поспособствовать еще и налаживанию и укреплению партнерских взаимоотношений между ключевыми фигурами на разных концах линий связи, позволяя им оперативно обмениваться мнениями и идеями. Но для реализации столь комплексного подхода к информационному обмену требуется потратить определенные усилия на его планирование.

9.1 Базовые принципы коммуникаций

Назначение любых средств и каналов связи — поддерживать передачу информативных сообщений получателям. Планируя обмен информацией, нужно учитывать содержание сообщения, средство или канал передачи, формат и целевую аудиторию. Для обеспечения соблюдения этого базового набора правил на любой формальный план информационного обмена — вне зависимости от тематики — должны распространяться общие принципы. Причем особо важное значение такая политика играет как раз таки во всем, что касается распространения информации об управлении данными, поскольку многие попросту не понимают критической роли управления данными в обеспечении успешной работы всей организации. Общий план коммуникаций и каждое отдельно взятое сообщение должны:

- преследовать ясную цель и иметь четко сформулированный желаемый результат;
- состоять только из ключевых посылов к получению желаемого результата;
- быть специальным образом адоптированы к потребностям целевой аудитории/получателей;
- доставляться по каналам, соответствующим специфике аудитории/получателей.

Тематика сообщений может быть самой разнообразной, но цели выхода на связь сводятся к решению одной из следующих задач:

- проинформировать;
- просветить или научить;
- поставить цели или сформулировать ви́дение;
- определить решение проблемы;
- стимулировать к изменениям в нужном направлении;
- повлиять на характер действий или мотивировать к действиям;
- запросить отзывы;
- заручиться поддержкой.

Самое важное во избежание недопонимания — существенность формулировок. Общие сведения об управлении данными будут значительно лучше доходить до понимания адресатов, если группа информационного управления возьмет на себя обязательство составить четкое представление о текущем состоянии практики управления данными на местах, о требуемых улучшениях, а также формулировку ви́дения, увязывающего задачи по совершенствованию практики управления

данными со стратегическими целями организации. Сообщения, касающиеся управления данными, должны:

- способствовать развитию и укреплению понимания материальной и нематериальной ценности инициатив по совершенствованию распоряжения и управления данными;
- описывать, каким именно образом функции управления данными вносят вклад в выработку и реализацию бизнес-стратегии и получение ощутимых результатов;
- демонстрировать на конкретных примерах из практики, как совершенствование управления данными приводит к снижению издержек, росту доходности, выявлению и нейтрализации риска, повышению качества принимаемых решений и т. п.;
- разъяснять фундаментальные принципы управления данными с целью повышения базового уровня понимания основ информационного управления всеми сотрудниками организации.

9.2 Оценка информированности и подготовка целевой аудитории

Планирование коммуникаций должно включать выявление целевых аудиторий для будущих сообщений. По результатам такого анализа появляется возможность выборочно и адресно компоновать контент сообщений, чтобы они выглядели и звучали осмысленно, актуально и значимо с точки зрения уровня, интересов и потребностей различных целевых аудиторий. Например, если план кампании заключается в том, чтобы заручиться поддержкой инициативы в верхах, то и формулировки должны быть максимально нацеленными именно на самое высшее руководство, где обычно более всего интересуются итоговыми показателями окупаемости и отдачи от вложений в финансируемые программы развития.

Тактика же убеждения людей в необходимости прислушаться к тому, что вы им предлагаете, в значительной степени зависит и от целевой аудитории, и от характера проблем, и от соотношения интересов адресных групп с целями и задачами программы, — поэтому ориентируйтесь на следующие общие рекомендации.

- **Предлагайте решения проблем**: транслируемые вами сообщения должны описывать, каким образом усилия по совершенствованию управления данными помогут решить проблемы, имеющие прямое отношение к удовлетворению нужд всех заинтересованных сторон, и помогут их как-то объединить. Примеры: нужды вкладчиков сильно отличаются от нужд банкиров; интересы разработчиков ИТ-решений расходятся с интересами бизнеса.
- **Учитывайте болевые точки**: у каждой из заинтересованных сторон они свои. Всесторонний учет проблем и больных мест различных интересантов в информационно-разъяснительных материалах помогает аудитории проникнуться ценностью ваших предложений. Примеры: службе ИБ будет интересно узнать, как и чем именно программа управления данными поспособствует снижению рисков утечек данных и облегчит ее работу, а отделу маркетинга — как она поможет отыскивать новых клиентов и развивать каналы сбыта.

- **Представляйте изменения как улучшения**: в большинстве случаев введение практик управления данными требует от людей изменения привычного ритма и порядка работы. Нужно разъяснять людям, что́ они за это получат, дабы пробудить в них желание реализовать предлагаемые изменения. Иными словами, нужно сделать так, чтобы они усмотрели в переменах позитив и пользу лично для себя.
- **Делитесь ви́дением успеха**: старайтесь красочно и образно описывать, как всё будет устроено в будущем по достижении желаемого состояния, чтобы адресаты послания поняли, как именно программа скажется лично на них. Нужно передать эту картину так, чтобы помочь аудитории понять и оценить преимущества программы управления данными.
- **Избегайте профессионального жаргона**: далеко не все понимают специфические термины из области ИТ и управления данными, поэтому не стоит злоупотреблять описанием технических аспектов предлагаемых решений, если вы обращаетесь к общей, а не профессиональной аудитории, чтобы не отвлекать внимание от основного смысла послания.
- **Рассказывайте об интересных случаях, приводите примеры**: занятные аналогии и невыдуманные истории — действенный способ иллюстрации смысла программы управления данными. Кроме того, на образных примерах людям проще понять и усвоить ее цели и задачи.
- **Не пренебрегайте запугиванием**: есть люди, для которых страх — лучший, а иногда и единственный мотиватор. Поэтому не забывайте рассказывать о возможных страшных последствиях неуправляемой циркуляции данных (например, штрафах, санкциях, административных и уголовных статьях), — это, кстати, дополнительно подчеркивает ценность данных и качественного управления ими. Впечатляющие примеры того, как отсутствие должного управления данными приводило к краху бизнес-подразделений и даже целых корпораций, также находят живой отклик у аудитории.

Эффективное выступление требует внимательного отслеживания реакции публики. Если чувствуется, что послание не доходит, смените тактику и попробуйте осветить вопрос под другим углом.

9.3 Задействование элементов неформального общения

Факты, примеры и рассказы о программе управления данными — не единственное средство формирования у заинтересованных лиц ее позитивного восприятия и понимания ценности. Люди прислушиваются к мнениям и находятся под влиянием не только своих коллег и начальства, но и неформальных лидеров. По этой причине при планировании коммуникаций нужно использовать методы анализа заинтересованных сторон для выявления устойчивых неформальных групп по интересам с целью воздействия на их лидеров, а через них — и на контингент тех, кто привык к ним прислушиваться. По мере расширения базы поддержки можно переходить к тактике использования новообретенных сторонников для распространения и донесения сообщения до понимания их коллег и руководителей[1].

[1] То есть использовать маркетинговый прием «вирусного» распространения. — *Примеч. пер.*

9.4 План коммуникаций

Коммуникации должны осуществляться слаженно и требуют продуманного планирования всех элементов. Таблица 39 описывает обязательные элементы хорошего плана или дорожной карты, необходимые для достижения поставленных целей.

Таблица 39. Элементы плана коммуникаций

Элемент	Описание
Сообщение	Информация, которую нужно донести до понимания целевой аудитории
Цель/Задача	Желаемый результат распространения сообщения (то есть зачем вообще понадобилось информационное воздействие на аудиторию?)
Аудитория	Целевая группа лиц или персона, на которую оказывается воздействие. Для каждой аудитории формулируется собственная цель кампании и разрабатывается отдельный план
Стиль	Уровень формальности/неформальности сообщений и глубина детализации сообщаемой информации подбираются индивидуально для каждой аудитории. Высшему руководству достаточно общей картины, а будущим участникам реализации проектов важны мельчайшие подробности. Выбор стилистики также зависит и от организационной культуры
Канал, метод, средство передачи	Способ и формат передачи месседжа (веб-страница, блог, электронная почта, личные встречи, малая группа, серия выступлений перед большой аудиторией, поучительные застольные беседы, семинары и т. д. и т. п.). Эффект воздействия зависит от средства подачи информации
Календарный план	Восприятие сообщений зависит еще и от их своевременности. Сотрудники с гораздо большей вероятностью будут читать полученные по электронной почте письма, если запланировать рассылку на начало рабочего дня в понедельник, а не на пятничный вечер. Если нужно убедить руководство выделить средства на следующий квартал или финансовый год, согласуйте отправку корреспонденции с циклом бюджетного планирования. Уведомления о предстоящих изменениях в рабочих процессах нужно рассылать заранее, а не накануне их вступления в силу и тем более не задним числом
Частота	Любое сообщение нужно транслировать регулярно и многократно, чтобы все адресаты его гарантированно услышали и усвоили. План-график рассылки сообщений и проведения информационно-разъяснительных мероприятий должен обеспечивать достаточную для усвоения целевой аудиторией частоту повторения каждого сообщения. По возможности разнообразьте каналы связи, а также используйте для оперативного информирования целевых групп периодические публикации (бюллетени, блоги и т. п.)
Материалы	План-график подготовки и выпуска информационных материалов — обязательный компонент любой информационно-разъяснительной кампании. Примеры: полные и краткие версии презентаций, пресс-релизы, заготовки текстов спонтанных обращений (в лифтах, столовых и т. п.), аннотации, анонсы и, конечно же, всевозможные маркетинговые материалы (плакаты, кружки, календари и прочие средства наглядной агитации)

Элемент	Описание
Коммуникаторы	Нужно тщательно подбирать кандидатуры тех, кто будет лично вести информационно-разъяснительную работу среди членов целевой группы. На роль «коммуникаторов сообщения» требуются люди, пользующиеся авторитетом и влиянием среди представителей целевой аудитории. К словам высокопоставленного спонсора программы управления данными или всеми уважаемого члена совета директоров большинство людей прислушается с куда большим вниманием, чем к разъяснениям рядового менеджера. Решения о том, кто именно будет озвучивать те или иные сообщения в адрес тех или иных адресных групп, должны приниматься сообразно целям, преследуемым этими разъяснениями
Ожидаемая реакция	План информационной кампании должен предвосхищать реакции различных групп, а иногда и отдельных лиц на полученные сведения. В принципе, не так уж и сложно спрогнозировать возможные вопросы и возражения, а значит, и подготовиться к грамотной реакции на них. В целом, мышление категориями потенциальных ответов и реакций — хороший способ дополнительного уяснения целей и формулировки убедительных сообщений в их поддержку
Метрики	План информационно-разъяснительной работы должен обязательно включать измеримые показатели ее эффективности. Ведь главный интерес представляет достижение цели, которая заключается в обеспечении понимания людьми содержания сообщений и согласия принять их за руководство к действию. Это проверяется через анкетирование, опросы, собеседования, фокусные группы и иные механизмы обратной связи. Ну и конечный и неоспоримый показатель успеха разъяснений — наблюдаемые реальные изменения в поведении
Бюджет и план ресурсного обеспечения	План информационной кампании должен учитывать потребности в ресурсах и расходах на его реализацию и ориентироваться на достижение поставленных целей в пределах выделенного бюджета

9.5 Продолжение осуществления коммуникаций по завершении внедрения программы управления данными

Управление данными осуществляется в рамках постоянно действующей программы, а не разовой инициативы или проекта. Соответственно, и усилия по информационной поддержке программы управления данными с целью обеспечения ее устойчивых успехов не должны приостанавливаться.

Старые сотрудники уходят, на их место приходят новые, да и внутри организации наблюдается ротация. По мере кадровых изменений нужно освежать в бизнес-подразделениях понимание необходимости управления данными. Требуют обновления и сами планы коммуникаций. Нужды заинтересованных сторон со временем меняются, да и программа управления данными не стоит на месте. Людям нужно какое-то время на то, чтобы до конца впитать смысл посланий, и выслушивание раз за разом повторений знакомых формулировок как раз и помогает им закрепить их понимание в долгосрочной памяти. В постепенной адаптации к изменяющимся реалиям и возросшему уровню сознательности сотрудников нуждаются как методы информирования, так и сами сообщения.

Не ослабевает со временем и конкуренция за финансирование. Одна из целей плана дальнейших информационных мероприятий — регулярно напоминать всем ключевым фигурам о финансовой отдаче и прочих преимуществах программы управления данными. Явный прогресс и громкие успехи — единственный залог дальнейшей финансово-организационной поддержки со стороны высшего руководства.

Эффективное планирование и постоянное информационное взаимодействие призваны обеспечить демонстрацию устойчивого усиления позитивного влияния совершенствования практики управления данными на организацию. Со временем накапливающееся у организации понимание важности данных приведет к окончательному переосмыслению их роли и значения. А до этого требуется постоянная разъяснительная работа по донесению до сознания организации понимания того факта, что лишь совершенствование управления данными позволяет бизнесу использовать их как ценнейший актив, извлекать прибыль из информационных ресурсов и гарантировать организации процветание.

10. ЦИТИРУЕМАЯ И РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Ackerman Anderson, Linda and Dean Anderson. *The Change Leader's Roadmap and Beyond Change Management*. Two Book Set. 2nd ed. Pfeiffer, 2010. Print.

Ackerman Anderson, Linda, Dean Anderson. *Beyond Change Management: How to Achieve Breakthrough Results Through Conscious Change Leadership*. 2nd ed. Pfeiffer, 2010. Print.

Ackerman Anderson, Linda, Dean Anderson. The *Change Leader's Roadmap: How to Navigate Your Organization's Transformation*. 2nd ed. Pfeiffer, 2010. Print.

Barksdale, Susan and Teri Lund. *10 Steps to Successful Strategic Planning*. ASTD, 2006. Print. 10 Steps.

Becker, Ethan F. and Jon Wortmann. *Mastering Communication at Work: How to Lead, Manage, and Influence*. McGraw-Hill, 2009. Print.

Bevan, Richard. *Changemaking: Tactics and resources for managing organizational change*. CreateSpace Independent Publishing Platform, 2011. Print.

Bounds, Andy. *The Snowball Effect: Communication Techniques to Make You Unstoppable*. Capstone, 2013. Print.

Bridges, William. *Managing Transitions: Making the Most of Change*. Da Capo Lifelong Books, 2009. Print.

Center for Creative Leadership (CCL), Talula Cartwright, and David Baldwin. *Communicating Your Vision*. Pfeiffer, 2007. Print.

Contreras, Melissa. *People Skills for Business: Winning Social Skills That Put You Ahead of The Competition*. CreateSpace Independent Publishing Platform, 2013. Print.

Covey, Stephen R. Franklin Covey Style Guide: *For Business and Technical Communication*. 5th ed. FT Press, 2012. Print.

Covey, Stephen R. *The 7 Habits of Highly Effective People: Powerful Lessons in Personal Change*. Simon and Schuster, 2013. Print. [Рус. пер.: *Стивен Р. Кови*. Семь навыков высокоэффективных людей: мощные инструменты развития личности. — М.: Альпина Паблишер, 2012.]

Franklin, Melanie. *Agile Change Management: A Practical Framework for Successful Change Planning and Implementation*. Kogan Page, 2014. Print.

Garcia, Helio Fred. Power of Communication: The: Skills to Build Trust, Inspire Loyalty, and Lead Effectively. FT Press, 2012. Print.

Godin, Seth and Malcolm Gladwell. *Unleashing the Ideavirus*. Hachette Books, 2001. [Рус. пер.: *Сет Годин*. Идея-вирус? Эпидемия! — СПб.: Питер, 2005.]

Harvard Business School Press. *Business Communication*. Harvard Business Review Press, 2003. Print. Harvard Business Essentials.

HBR's 10 Must Reads on Change Management. Harvard Business Review Press, 2011. Print.

Hiatt, Jeffrey, and Timothy Creasey. *Change Management: The People Side of Change*. Prosci Learning Center Publications, 2012. Print.

Holman, Peggy, Tom Devane, Steven Cady. *The Change Handbook: The Definitive Resource on Today's Best Methods for Engaging Whole Systems*. 2nd ed. Berrett-Koehler Publishers, 2007. Print.

Hood, J. H. *How to book of Interpersonal Communication: Improve Your Relationships*. Vol. 3. WordCraft Global Pty Limited, 2013. Print. «How to» Books.

Jones, Phil. *Communicating Strategy*. Ashgate, 2008. Print.

Kotter, John P. *Leading Change*. Harvard Business Review Press, 2012. Print.

Locker, Kitty, and Stephen Kaczmarek. *Business Communication: Building Critical Skills*. 5th ed. McGraw-Hill/Irwin, 2010. Print.

Luecke, Richard. *Managing Change and Transition*. Harvard Business Review Press, 2003. Print. Harvard Business Essentials.

Rogers, Everett M. *Diffusion of Innovations*. 5th Ed. Free Press, 2003. Print.

Выражение признательности

Второе издание DAMA-DMBOK — плод бескорыстного труда многих людей. Работа началась в конце 2011 года с разработки документа, содержавшего описание первоначальной версии пересмотренной рамочной структуры, который был опубликован в 2012 году. После этого редакционный комитет DAMA-DMBOK потратил много времени и усилий на подготовку к выпуску проекта DMBOK2. В состав комитета вошли:

Патриция Куполи (Patricia Cupoli, DAMA, Филадельфия) являлась главным редактором на протяжении большей части работы, осуществляя деятельность по поиску авторов и оказанию им помощи в написании глав. К нашему глубочайшему сожалению, Пат ушла из жизни летом 2015 года. Она оставалась на своем посту до последних дней.

Дебора Хендерсон (Deborah Henderson, IRMAC [канадский аффилированный член DAMA], Торонто), директор программы разработки продуктов DAMA-DMBOK с момента ее запуска в 2005 году. Всячески способствовала продвижению проекта и обеспечила его успешное завершение после ухода Пат.

Сьюзен Эрли (Susan Earley, DAMA, Чикаго) разработала эскизный вариант рамочной структуры DAMA-DMBOK2 и осуществляла основную редакторскую правку проекта DMBOK2. Она приводила в порядок и организовывала весь обильный и разнообразный контент DMBOK2, а также принимала, рассматривала и учитывала множество отзывов и предложений, поступавших от членов DAMA.

Эва Смит (Eva Smith, DAMA, Сиэтл), координатор проекта и менеджер средств совместной работы, прекрасно справилась с задачей логистического обеспечения проекта, включая организацию онлайн-форума для обсуждения предварительных версий глав DMBOK2 членами DAMA.

Елена Сикора (Elena Sykora, IRMAC, Торонто), библиограф-исследователь, составила исчерпывающую библиографию для каждой из семнадцати глав DMBOK2.

Неоценимую поддержку редакционному комитету также оказали Саньяй Шируд (Sanjay Shirude), Кэти Нолан (Cathy Nolan), Эмери Поуп (Emarie Pope) и Стив Хоберман (Steve Hoberman), которым мы и выражаем здесь самую теплую и сердечную благодарность.

Лаура Себастьян-Коулман (Laura Sebastian-Coleman, DAMA, Новая Англия), глава службы публикаций DAMA и выпускающий редактор DMBOK2, довела до совершенства книгу перед публикацией. В этом ей помогали члены экспертного редакционного совета, в состав которого вошли Питер Айкен (Peter Aiken), Крис Брэдли (Chris Bradley), Ян Хендерикс (Jan Henderyckx), Майк Дженнингс (Mike Jennings), Дара О'Брайен (Daragh O'Brien) и автор этих строк, а также наши внештатные помощники — Лайза Олинда (Lisa Olinda) и Данетт Макгилврэй (Danette McGilvray).

И, конечно же, DMBOK2 никогда не получилась бы без вклада основных авторов ее глав, воплотивших замысел книги и облачивших ви́дение рамочной структуры в словесные образы. Все они работали над созданием книги на добровольных началах, щедро вкладывая в общее дело не только свои познания, но и драгоценное время. Всем им мы в заключение выражаем свою признательность поименно. Кроме того, мы искренне благодарны и сотням членов DAMA, поделившихся с нами своими отзывами и предложениями. Имена всех этих неравнодушных помощников мы также публикуем.

Проект DMBOK спонсировался DAMA International, DAMA International Foundation и Советом председателей региональных отделений DAMA. Без их дальновидности, прозорливости, терпения и неустанной поддержки этот проект едва ли оказался бы столь успешным.

Ну и напоследок хотелось бы выразить признательность родным и близким всех добровольцев — участников проекта, — которые согласились пожертвовать своими интересами и позволили им выкроить личное время на созидание и завершение данного труда.

Сью Гьюенс (Sue Geuens), президент DAMA International

Основные соавторы

№	Глава	Основные соавторы
1	Управление данными	Экспертный редакционный совет, редакционный комитет DMBOK
2	Этика обращения с данными	Крис Брэдли (Chris Bradley), Кен Кринг (Ken Kring)
3	Руководство данными	Джон Лэдли (John Ladley), Марк Коуэн (Mark Cowan), Саньяй Шируд (Sanjay Shirude)
4	Архитектура данных	Хакан Эдвинссон (Håkan Edvinsson)

№	Глава	Основные соавторы
5	Моделирование и проектирование данных	Стив Хоберман (Steve Hoberman)
6	Хранение и операции с данными	Саньяй Шируд (Sanjay Shirude)
7	Безопасность данных	Давид Шлезингер (David Schlesinger, CISSP)
8	Интеграция и интероперабельность данных	Эйприл Рив (April Reeve)
9	Управление документами и контентом	Патриция Куполи (Patricia Cupoli)
10	Справочные и основные данные	Джин Бумер (Gene Boomer), Мехмет Орун (Mehmet Orun)
11	Ведение хранилищ данных и бизнес-аналитика	Мартин Сикора (Martin Sykora), Криш Кришнан (Krish Krishnan), Джон Лэдли (John Ladley), Лайза Нельсон (Lisa Nelson)
12	Управление метаданными	Саад Яцу (Saad Yacu)
13	Качество данных	Россано Таварес (Rossano Tavares)
14	Большие данные и наука о данных	Роберт Абате (Robert Abate), Мартин Сикора (Martin Sykora)
15	Оценка зрелости управления данными	Марк Коуэн (Mark Cowan), Дебора Хендерсон (Deborah Henderson)
16	Организация управления данными и ролевые ожидания	Келле О'Нил (Kelle O'Neal)
17	Управление данными и управление организационными изменениями	Мишелин Кейси (Micheline Casey), Андреа Томсен (Andrea Thomsen), Дара О'Брайен (Daragh O'Brien)
	Библиография (главы 1–17)	Елена Сикора (Elena Sykora)

Рецензенты и комментаторы

Список лиц, поделившихся с проектом DMBOK2 ценными отзывами, замечаниями и предложениями на разных стадиях работы:

Khalid Abu Shamleh
Gerard Adams
James Adman
Afsaneh Afkari
Mike Beauchamp
Chan Beauvais
Glen Bellomy
Stacie Benton
Susan Burk
William Burkett
Beat Burtscher
Ismael Caballero

Zaher Alhaj
Shahid Ali
Suhail Ahmad AmanUllah
Nav Amar
Samuel Kofi Annan
Ivan Arroyo
Nicola Askham
Juan Azcurra
Richard Back
Carlos Barbieri
Ian Batty
Steve Beaton
Cynthia Dionisio
Shaun Dookhoo
Janani Dumbleton
Lee Edwards
Jane Estrada
Adrianos Evangelidis
William Evans
Mario Faria
Gary Flye
Michael Fraser
Carolyn Frey
Alex Friedgan
Lowell Fryman
Shu Fulai
Ketan Gadre
Oscar Galindo
Alexandre Gameiro
Jay Gardner
Johnny Gay
Sue Geuens
Sumit Gupta
Gabrielle Harrison
Kazuo Hashimoto
Andy Hazelwood
Muizz Hassan
Leon Bernal
Luciana Bicalho
Pawel Bober
Christiana Boehmer
Stewart Bond
Gene Boomer
Taher Borsadwala
Antonio Braga
Ciaran Breen
LeRoy Broughton
Paul Brown
Donna Burbank
Nicholene Kieviets
Jon King
Richard King
Bruno Kinoshita
Yasushi Kiyama
Daniel Koger
Katarina Kolich
Onishi Koshi
Edwin Landale
Teresa Lau
Tom LaVerdure
Richard Leacton
Michael Lee
Martha Lemoine
Melody Lewin
Chen Liu
Manoel Francisco Dutra Lopes Jr
Daniel Lopez
Karen Lopez
Adam Lynton
Colin Macguire
Michael MacIntyre
Kenneth MacKinnon
Colin Maguire
Zeljko Marcan
Peter Campbell
Betty (Elizabeth) Carpenito
Hazbleydi Cervera
Indrajit Chatterjee
Bavani Chaudhary
Denise Cook
Nigel Corbin
James Dawson
Elisio Henrique de Souza
Patrick Derde
Tejas Desai
Swapnil Deshmukh
Susana Navarro
Gautham Nayak
Erkka Niemi
Andy O'Hara
Katherine O'Keefe
Hirofumi Onozawa
Mehmet Orun
Matt Osborn
Mark Ouska
Pamela Owens
Shailesh Paliwal
Mikhail Parfentev
Melanie Parker
John Partyka
Bill Penney
Andres Perez
Aparna Phal
Jocelyn Sedes
Mark Segall
Ichibori Seiji
Brian Phillippi
R. Taeza Pittman
Edward Pok
Emarie Pope
David Quan

David Hay
Clifford Heath
Jan Henderyckx
Trevor Hodges
Mark Horseman
Joseph Howard
Monica Howat
Bill Huennekens
Mark Humphries
Zoey Husband
Toru Ichikura
Thomas Ihsle
Gordon Irish
Fusahide Ito
Seokhee Jeon
Jarred Jimmerson
Christopher Johnson
Wayne Johnson
Sze-Kei Jordan
George Kalathoor
Alicia Slaughter
Eva Smith
Tenny Soman
José Antonio Soriano Guzmán
Donald Soulsby
Erich Stahl
Jerry Stembridge
James Stevens
Jan Stobbe
Santosh Subramaniam
Motofusa Sugaya
Venkat Sunkara
Alan Sweeney
Martin Sykora
Satoshi Matsumoto
George McGeachie
Danette McGilvray
R. Raymond McGirt
Scott McLeod
Melanie Mecca
Ben Meek
Steve Mepham
Klaus Meyer
Josep Antoni Mira Palacios
Toru Miyaji
Ademilson Monteiro
Danielle Monteiro
Subbaiah Muthu Krishnan
Mukundhan Muthukrishnan
Robert Myers
Dean Myshrall
Krisztian Nagy
Kazuhiro Narita
Mohamad Naser
Akira Takahashi
Steve Thomas
Noriko Watanabe
Joseph Weaver
Christina Weeden
Alexander Titov
Steven Tolkin
Toshimitsu Tone
Juan Pablo Torres
David Twaddell
Thijs van der Feltz
Elize van der Linde
Peter van Nederpelt
Peter Vennel
K Rajeswar Rao
April Reeve
Todd Reyes
Raul Ruggia-Frick
Scott Sammons
Pushpak Sarkar
John Schmidt
Nadine Schramm
Toshiya Seki
Rajamanickam Senthil Kumar
Sarang Shah
Gaurav Sharma
Vijay Sharma
Stephen Sherry
Jenny Shi
Satoshi Shimada
Sandeep Shinagare
Boris Shuster
Vitaly Shusterov
Abi Sivasubramanian
Roy Verharen
Karel Vetrovsky
Gregg Withers
Michael Wityk
Marcin Wizgird
Benjamin Wright-Jones
Teresa Wylie
Hitoshi Yachida
Saad Yacu
Hiroshi Yagishita
Harishbabu Yelisetty
Taisei Yoshimura

Предметный указатель

Именной указатель

Издательство «Олимп–Бизнес»
121170, Москва, Кутузовский проезд, 16
Тел./факс: (495) 917-85-66 (многоканальный)
Интернет-магазин: www.olbuss.ru; E-mail: es@olbuss.com

Как купить наши книги:

- В интернет-магазине издательства: www.olbuss.ru
- Сделать заказ по телефону (495) 917-85-66
- Приехать в офис издательства «Олимп–Бизнес»

Спрашивайте книги нашего издательства в магазинах вашего города

Издательство «Олимп–Бизнес» приглашает к сотрудничеству оптовиков, книготорговые организации и магазины.
Информацию об условиях работы можно получить в отделе продаж издательства

Мы в социальных сетях:

Facebook: @OlympBusiness
VKontakte: @olimpbusiness
Instagram: @olimp_business

DAMA-DMBOK
СВОД ЗНАНИЙ ПО УПРАВЛЕНИЮ ДАННЫМИ
Второе издание

Издатель *Ирина Седакова*

Перевод *Григорий Агафонов*
Выпускающий редактор *Роман Герасимов*
Литературный редактор *Диана Василёнкайтите*
Корректор *Наталья Стахеева*
Оригинал-макет и верстка *Светлана Опарина*
Художник *Роман Рузавин*

Подписано в печать 09.12.2019.
Формат 84×108 1/16. Бумага офсетная.
Гарнитура «Minion Pro». Усл. печ. л. 86,94.

Издательство «Олимп–Бизнес»
121170, Москва, Кутузовский проезд, 16

Напечатано в России

Знак информационной продукции
(Федеральный закон № 436-ФЗ от 29.12.2010 г.)

12+

DAMA-DMBOK2 — настоящий must-have для любого специалиста в области управления данными: от начинающего аналитика до CDO. Данное руководство позволяет лидерам цифровизации развивать системное мышление, обозначать ясные ориентиры и вырабатывать четкие принципы работы. Книга написана настолько подробно, что специалист любого уровня найдет в ней для себя что-то новое.

DAMA-DMBOK2 — не просто свод энциклопедических знаний; руководство содержит практические рекомендации по внедрению новой корпоративной функции — руководства данными с организационной структурой и процессной моделью. Рекомендуем каждому руководителю по цифровой трансформации обязательно прочитать DMBOK, прежде чем инвестировать в длительные проекты по Data Governance.

DAMA-DMBOK2 — безусловно лучшая инвестиция в цифровое будущее бизнеса и в собственные компетенции.

Юрий Клочко,
Генеральный директор компании BSSG

Андрей Ларионов,
к. т. н., MBA, Руководитель направления консалтинга компании BSSG

DAMA-DMBOK2 — незаменимое руководство, которое поможет каждому специалисту не потеряться на бескрайних просторах управления данными. Оно поможет вам расставить нужные акценты и определить направление работы. Уверен: знакомство с книгой станет исключительно полезным и интересным!

Сергей Кузнецов,
Генеральный директор компании «Юнидата»

Задачи «DAMA-DMBOK2»:

- Выработка общепринятого согласованного представления об областях знаний по управлению данными (выделено 11 таких областей)
- Определение руководящих принципов управления данными
- Предоставление стандартных определений для наиболее часто используемых понятий (общих и по областям знаний)
- Обзор общепринятых лучших практик, широко распространенных методов и методик, а также наиболее известных альтернативных подходов
- Краткий обзор общих организационных и культурных вопросов
- Уточнение границ сферы управления данными

ОЛИМП БИЗНЕС

интернет-магазин
www.olbuss.ru

9 785969 304048