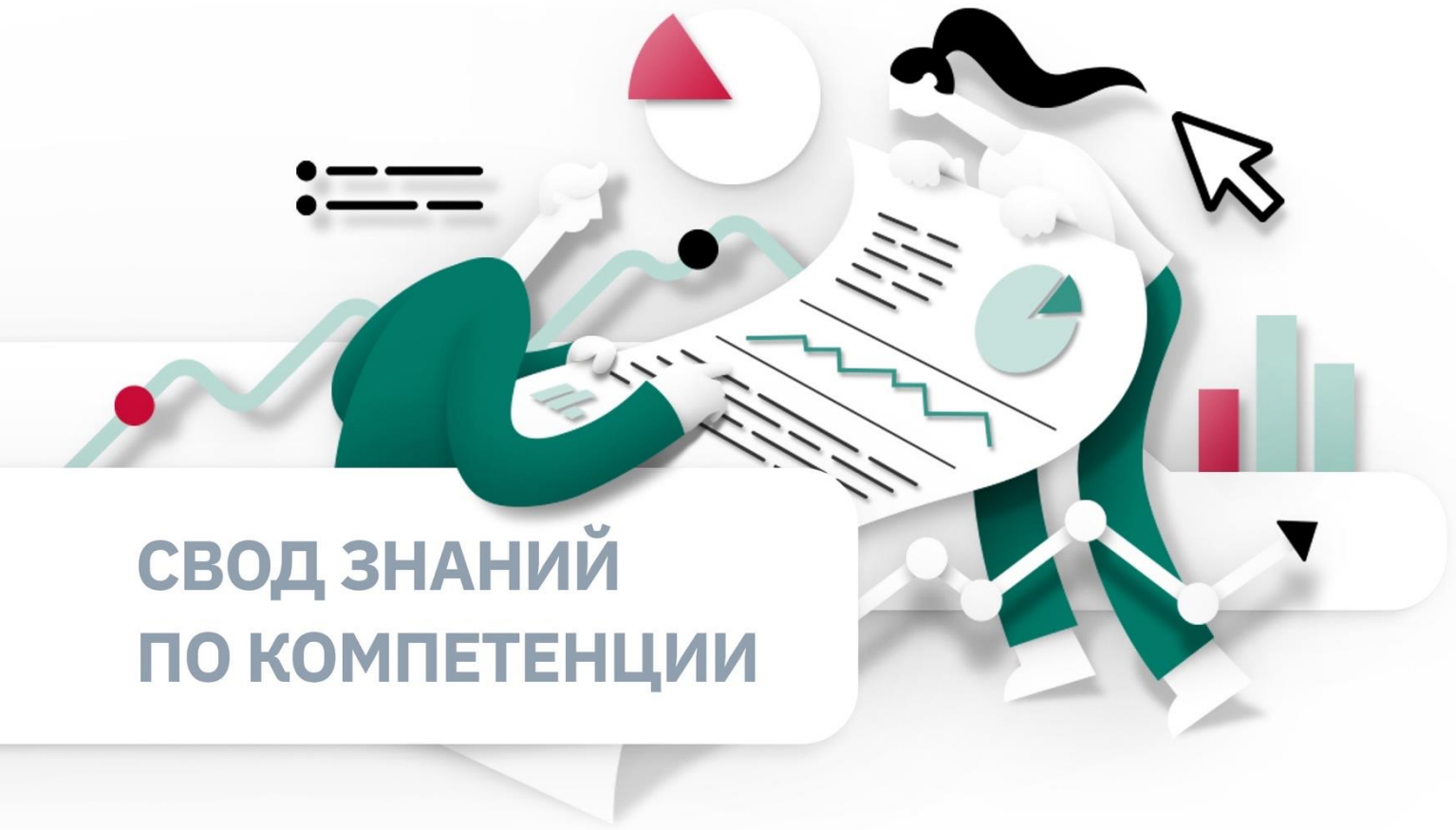
РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВШГУ Высшая Школа Государственного Управления
РАНХиГС

ЦЕНТР ПОДГОТОВКИ РУКОВОДИТЕЛЕЙ И КОМАНД ЦИФРОВОЙ ТРАНСФОРМАЦИИ



СВОД ЗНАНИЙ ПО КОМПЕТЕНЦИИ

п.4. Управление данными

ЦЕНТР ПОДГОТОВКИ
РУКОВОДИТЕЛЕЙ И КОМАНД
ЦИФРОВОЙ ТРАНСФОРМАЦИИ

СВОД ЗНАНИЙ ПО КОМПЕТЕНЦИИ
П.4. УПРАВЛЕНИЕ И ИСПОЛЬЗОВАНИЕ ДАНЫХ

ОГЛАВЛЕНИЕ

Введение 5

П.4.1. Управление данными, архитектура и моделирование 7

4.1.1. Управление данными 7

Цели организации по управлению данными 7

Принципы управления данными 9

Проблемы управления данными 10

Ценность данных 11

Качество данных 12

Жизненный цикл данных 13

Стратегия управления данными 14

DAMA-DMBOK2 — фреймворк управления данными 15

4.1.2. Архитектура данных 17

Понятие и цели архитектуры данных 17

Ценности архитектуры данных 19

Корпоративная архитектура данных 20

ШАГИ ПО ВНЕДРЕНИЮ КОРПОРАТИВНОЙ АРХИТЕКТУРЫ ДАННЫХ 23

Национальная система управления данными 25

4.1.3. Моделирование и проектирование данных 27

Определение моделирования данных 27

Виды моделируемых данных 27

Концептуальные, логические и физические схемы 28

Схемы представления данных при моделировании 30

Свойства данных 34

Процесс моделирования данных 35

Методики моделирования данных 37

Универсальное моделирование данных 38

Семантическое моделирование данных 39

Инструменты работы с данными 40

Лучшие практики проектирования баз данных ... 42
КЛЮЧЕВАЯ РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА ... 43
П.4.2. Обеспечение хранения, безопасности и этичного обращения с данными ... 44
4.2.1. Хранение данных ... 44
База данных: понятие, назначение ... 44
Классификация по модели данных ... 45
Классификация по степени распределенности ... 48
Средства хранения данных ... 50
Среды баз данных ... 52
Общие процессы обработки данных в базах данных ... 54
Функции системы управления базами данных ... 57
Компоненты СУБД ... 59
Подходы к обработке данных ... 61
Жизненный цикл базы данных ... 64
Интеграция и интероперабельность данных ... 67
Инструменты интеграции и интероперабельности данных ... 78
Руководство интеграцией и интероперабельностью данных ... 81
4.2.2. Безопасность данных ... 84
Ключевые понятия ... 84
Факторы обеспечения информационной безопасности ... 86
Политика безопасности: актуальная документация ... 88
Риски информационной безопасности ... 89
Утечка данных в России: факты ... 91
Защита конфиденциальной информации ... 92
4.2.3. Этика обращения с данными ... 98
Цели организации по этике данных ... 98
Принципы защиты данных в рамках этических вопросов ... 99
Примеры неэтичного обращения с данными и риски ... 100
Формирование культуры этичного обращения с данными ... 101
КЛЮЧЕВАЯ РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА ... 103
ПК 4.3. Управление большими данными и качеством данных ... 105
4.3.1. Ключевые понятия управления большими данными ... 105
Понятие больших данных ... 105
Наука о данных (data science) ... 109
Большие данные ... 112
Компоненты архитектуры больших данных ... 114
Источники больших данных ... 115

Озера данных 116
4.3.2. Техники и методы анализа больших данных 117
Машинное обучение 118
Анализ настроений 119
Интеллектуальный анализ данных и текстов 120
Предиктивная аналитика 122
Предписывающая аналитика 123
Методы анализа неструктурированных данных 124
Операционная аналитика 125
Визуализация данных 126
Объединение данных 128
4.3.3. Управление качеством данных 130
Основные понятия 130
Принципы управления качеством данных 133
Условия управления качеством основных данных 134
Измерение качества данных 135
Роль метаданных в обеспечении качества данных 139
Структура управления качеством основных данных 140
Базовые стандарты, регламентирующие управление качеством данных 142
Инструменты управления качеством данных 143
Методология оценки и повышения качества данных на примере нсуд 144
КЛЮЧЕВАЯ РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА 149

Введение

Почему важна грамотность работы с данными

1. **Данных становится больше**

Человечество производит все больше и больше данных, они практически удваиваются каждый год. Хотя мы просим наших сотрудников использовать и даже монетизировать эти данные, но у них просто-напросто нет навыков для этого.

2. **Данные становятся более ценными**

Компания делает свои данные доступными для сотрудников, чтобы они сами могли строить отчеты или работать с данными. К счастью, уже многие компании знакомы с SelfService аналитикой[1] и не боятся отдать данные рядовым сотрудникам для самостоятельного анализа.

3. **Данные становятся доступнее**

Поскольку объем данных в последние годы растет экспоненциально, аналитика коренным образом изменила методы работы предприятий. Компании пытаются перевести все процессы в цифровой вид, все оцифровать и тегировать, и это генерирует все больше и больше данных. Для организаций очень важно использовать все эти данные для принятия правильных решений, но без обученного персонала сделать это практически невозможно.

По данным ежегодного опроса Gartner, низкая грамотность по работе с данными считается вторым по величине внутренним препятствием на пути к успеху проектов по оптимизации бизнес-аналитики, проводимых под руководством директора по данным - Chief Data Officer (CDO). По мере того, как данных становится все больше и принятие решений в организации становятся все более зависимым от полноты и правильной интерпретации данных, низкий уровень работы с данными станет препятствием для роста компании. В 2021 году 50% организаций испытывают недостаток сотрудников, умеющих работать с данными и искусственным интеллектом, необходимым для достижения целей бизнеса.

Грамотная работа с данными – это умение понимать данные, работать с ними и анализировать их, а также правильно интерпретировать и использовать. Необходимо уметь обосновывать свое мнение и выводы, беря данные за основу.

В бизнесе выводы, полученные на основе работы с данными являются «языком» современной компании. Для современной компании грамотная работа с данными так же важна, как умение читать и писать.

Навыки работы с данными полезны не только в бизнес-компаниях. Сегодня каждый человек должен знать, как быстро преобразовать данные в идеи и действия.

К ключевым навыкам, которые должны быть в организации, для того чтобы выстроить компанию, которая действует по принципам data-driven[2] и которой успешно управляют на основе данных, можно отнести:

- **концептуальный подход**
  - понимание основ работы с данными и основных подходов;
- **сбор данных**
  - обнаружение данных и их сбор;
  - оценка и подтверждение качества данных и источников данных;

1 Концепция, суть которой в предоставлении бизнес-пользователям удобного инструментария для самостоятельного и эффективного анализа данных, в зависимости от их функциональных потребностей и рабочих ролей: от простой фильтрации и группировки до модификации структуры отчетов и изменения семантических моделей

2 Принимает решения на основе данных

- **управление данными**
  - организация данных;
  - обработка данных;
  - конвертация данных;
  - создание и использование метаданных;
  - курирование данных, управление безопасностью;
  - хранение данных;
- **оценка данных**
  - знание инструментов;
  - базовый анализ данных;
  - интерпретирование и понимание;
  - распознавание проблем при работе с данными;
  - визуализация данных и устное представление данных;
  - принятие решений на основе данных;
- **применение данных**
  - критическое мышление;
  - культура данных;
  - этика данных;
  - цитирование данных;
  - совместное использование данных;
  - оценка решений, принятых на основе данных.

Данный свод знаний включает материал, который поможет понять основы управления данными и их применения в практике организаций. В основе – материалы издания «DAMA-DMBOK2: Свод знаний по управлению данными»[3]. Эта книга написана экспертами Международной ассоциации управления данными (DAMA) и представляет собой наиболее полное введение в дисциплину управления данными с обзором лучших практик.

Data Literacy Project в России запускает проект по популяризации «DAMA-DMBOK2: Свод знаний по управлению данными». В данном своде знаний также использованы материалы портала https://dataliteracy.ru/.

3 DAMA-DMBOK: свод знаний по управлению данными / Dama International; пер. с англ. Г. Агафонова. 2-е изд. М.: Олимп–Бизнес, 2020. с. 113.

П.4.1. Управление данными, архитектура и моделирование

4.1.1. Управление данными

Цели организации по управлению данными

Спектр видов деятельности в сфере управления данными широк и разнообразен — от принятия логически обоснованных решений, как извлечь стратегическую ценность из имеющихся или потенциально доступных 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, с тем чтобы обеспечить высококачественные данные, соответствующие стратегическим целям и задачам организации.

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.

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.

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[4].

Управление данными должно обеспечивать (цели управления данными):
- заинтересованных сторон
- сбор, хранение, защиту и обеспечение целостности данных
- обеспечение качества данных и информации
- обеспечение конфиденциальности и неразглашения данных
- предотвращение несанкционированного доступа к данным, их искажения и подтасовки
- обеспечение эффективного использования данных.

Данных в природе не существует, данные создаются. Данные — это стратегически значимый актив организации. Для извлечения максимальной пользы из данных необходимо

4 DAMA-DMBOK: свод знаний по управлению данными / Dama International; пер. с англ. Г. Агафонова. 2-е изд. М.: Олимп-Бизнес, 2020. с. 113.

осваивать новые способы их использования. В отрыве от контекста данные становятся бессмысленными, для понимания их смысла нужно знать контекст.

Американское общество качества (ASQ: Excellence Through Quality) 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»[5]. Последнее определение явным образом указывает на электронную форму данных и подразумевает (вполне справедливо), что данные требуют стандартизации, поскольку управляются посредством информационно-технологических систем. Однако в данном определении не упомянуты трудности непротиворечивой и однозначной формализации данных, передаваемых между несовместимыми системами, и понятие неструктурированных данных.

Оксфордский словарь определяет данные как факты и статистику, собранную вместе для анализа[6].

Ниже даны ключевые понятия в области управления данными[7].

Информация
данные, которым придается некоторый смысл (интерпретация) в конкретной ситуации в рамках некоторой системы понятий. Информация представляется посредством кодирования данных и извлекается путем их декодирования и интерпретации.

Данные
способ представления информации в определенной, фиксированной форме, пригодной для обработки, хранения и передачи.

Банк данных
система специальным образом организованных данных: баз данных, программных, технических, языковых, организационно-методических средств, предназначенных для обеспечения централизованного хранения и коллективного многоцелевого использования данных.

База данных (БД)
именованная совокупность данных, отражающая состояние объектов и их отношений в конкретной предметной области.

База знаний
это особого рода база данных, разработанная для оперирования знаниями (метаданными). Полноценные базы знаний содержат в себе не только фактическую информацию, но и правила вывода, допускающие автоматические умозаключения о вновь вводимых фактах и, как следствие, осмысленную обработку информации.

Система управления базами данных (СУБД)
совокупность языковых и программных средств, предназначенных для создания, ведения и совместного использования БД многими пользователями.

5 ГОСТ Р ИСО/МЭК 11179-1-2010 «Информационная технология. РЕГИСТРЫ МЕТАДАННЫХ (РМД). Часть 1. Основные положения»
6 New Oxford American Dictionary // URL: https://languages.oup.com
7 Рузаков А. А. Управление данными: учеб. пос. Челябинск: Изд-во Челяб. гос. пед. ун-та, 2015. с.132.

Принципы управления данными

К принципам управления данными можно отнести:

1. **Осознание полезности управления данными для организации**

Необходимы лидеры и приверженность руководства для эффективного управления данными.

2. **Данные имеют ценность**

Данные имеют уникальные свойства. Ценность данных может и должна оцениваться в экономических терминах.

3. **Управление данными включает комплекс навыков**

Одной команде не по силам управлять всеми данными организации. Управление данными — кросс-функциональный процесс, требующий широкого спектра знаний и навыков. Необходим комплексный взгляд на функционирование организации.

4. **Управление данными требует постоянного внимания и включение во все бизнес-процессы организации**

Необходимо управлять качеством данных. Для управления данными нужны метаданные. Необходимо планировать управление данными. ИТ должны обслуживать, а не определять стратегические потребности организации в данных.

5. **Управление данными – это управление жизненным циклом**

Необходимо понимать жизненный цикл данных и уметь управлять им. Управление рисками осуществляется на протяжении всего жизненного цикла данных. Управление данными должно непрерывно эволюционировать.

Проблемы управления данными

К проблемам управления данными можно отнести:

- данные должны быть качественными;
- компания накапливает огромный объем данных, о которых сама даже не догадывается;
- данные необходимо надежно хранить, а это стоит денег;
- данные легко скопировать и отправить куда угодно;
- одними и теми же данными могут одновременно пользоваться множество людей;
- многие способы использования данных порождают еще больше данных;
- если организация утратила свои уникальные данные, заменить их будет нечем, а восстановить невозможно или непомерно дорого;
- большинство информационных транзакций предусматривают обмен данными.

Ценность данных

Ценность — это разница между затратами на создание или приобретение вещи и полученной от нее выгоды. Стандартных методик по расчету ценности данных не существует.

Данные организации уникальны и присущи только ей, оценку их стоимости нужно начинать с определения общих категорий затрат и экономических выгод (рис. 1).

РИСУНОК 1
СЛАГАЕМЫЕ ЦЕННОСТИ ДАННЫХ

Качество данных

Необходимо гарантировать высокое качество данных. Без этого все затраты на сбор, хранение и защиту данных, окажутся напрасными.

Необходимо тесно работать с конечными потребителями данных, чтобы определить их потребности и совместно установить ключевые характеристики качества их данных.

Некачественные данные приведут к неверным решениям и негативным последствиям (издержки из-за низкого качества данных в 2016 году только в США составили около 3 трлн долл.).

Основные издержки, которые являются следствием низкого качества данных:

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

Выгоды высококачественных данных:

- повышение производительности;
- снижение рисков;
- повышение прибыли;
- новые конкурентные преимущества за счет глубокого понимания клиентов, продуктов, процессов и возможностей.

Жизненный цикл данных

Жизненный цикл данных включает процессы, которые обеспечивают 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.

На протяжении всего жизненного цикла данные можно очищать, преобразовывать, подвергать слиянию, улучшать или агрегировать.

Кроме жизненного цикла данные имеют определенную последовательность, цепочку – путь, по которому данные перемещаются от мест возникновения до места использования (data chain).

Важнейшие аспекты жизненного цикла данных:

1. Управлять данными нужно с четким пониманием, как они созданы и получены. Создание данных стоит денег. Затраты окупаются только в том случае, если полученные данные и имеют реальную ценность.

2. Управлять качеством данных на протяжении всего жизненного цикла. Поскольку качество данных может пострадать на любом этапе жизненного цикла, следует планировать меры по обеспечению качества данных в расчете на весь жизненный цикл данных.

3. Управлять качеством метаданных на протяжении всего жизненного цикла.

4. Обеспечить информационную безопасность данных.

5. Основные усилия необходимо фокусировать на критически важных данных, и минимизировать излишние, устаревшие и тривиальные данные.

Стратегия управления данными

Стратегия работы с данными должна предусматривать бизнес-планы использования информации для получения конкурентных преимуществ и достижения целей организации (рис. 2). Необходимо понимать, какие данные нужны организации, как она будет получать эти данные, управлять ими, обеспечивать их надежность и достоверность на протяжении всего жизненного цикла, каким образом будет использовать данные. Здесь важную роль играет директор по данным (CDO).

РИСУНОК 2
РЕЗУЛЬТАТЫ СТРАТЕГИЧЕСКОГО ПЛАНИРОВАНИЯ

Положение
об управлении данными

Описание содержания
программы управления данными

Дорожная карта
внедрения управления данными

ЧЕК-ЛИСТ: ЧТО ДОЛЖНА ВКЛЮЧАТЬ СТРАТЕГИЯ УПРАВЛЕНИЯ ДАННЫМИ

В стратегии управления данными должны быть:

- убедительно изложенное видение управления данными;
- краткое экономическое обоснование стратегии с примерами;
- руководящие принципы, ценности и перспективы с позиции управления;
- миссия и долгосрочные цели по основным направлениям управления данными
- показатели успешности управления данными;
- краткосрочные задачи на 1–2 года программы управления данными по принципу SMART;
- описание ролей и организационных систем, включая распределение обязанностей и прав принятия решений;
- описание компонентов и инициатив программы управления данными;
- приоритетную программу работ с объемами и сроками выполнения;
- вариант дорожной карты реализации с разбивкой по проектам и мероприятиям.

DAMA-DMBOK2 — фреймворк управления данными

DAMA – это Международная ассоциация управления данными (Data Management Association International), которая объединяет профессионалов как из сферы ИТ, так и бизнес и развивается абсолютно независимо. Уже 40 лет DAMA собирает, систематизирует и пропагандирует лучшие практики по управлению данными и информацией. Важная задача ассоциации — выработка единой терминологии и подхода к управлению данными[8].

Книга DAMA DMBOK2[9], как и сама ассоциация, абсолютно независима: она не ссылается на какие-то конкретные решения по работе с данными или вендоров, а определяет набор принципов и их применение в функциональных областях управления данными.

DAMA DMBOK ставит следующие задачи:

- 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;
- 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.

«Колесо DAMA» (рис. 3) отражает области знаний по управлению данными. В центре располагается руководство данными (Data Governance), поскольку оно требует консистентности и баланса функций управления данными. Остальные области знаний (архитектура, моделирование, хранение данных и т. д.) являются важными составляющими зрелого управления данными в организации, их можно внедрять в разные периоды, в зависимости от требований бизнеса.

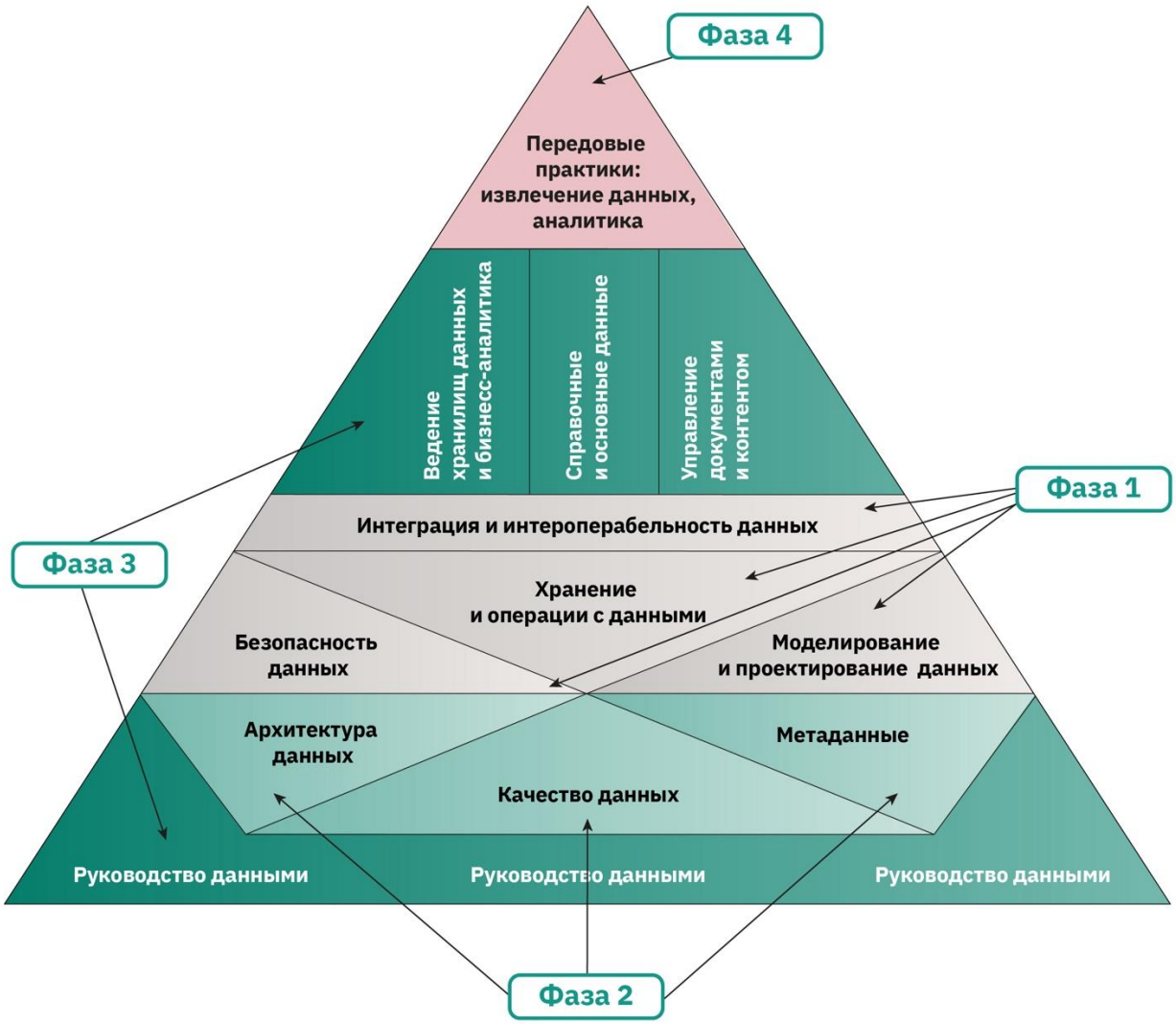
РИСУНОК 3
КОЛЕСО «DAMA»

8 DAMA International. URL: https://www.dama.org/cpages/home
9 Интервью с одним из ключевых авторов DAMA DMBOK: Кристофер Брэдли, CDMP Fellow // YouTube. URL: https://youtu.be/kxYdOhBE4Hw

ПИРАМИДА DMBOK2

Пирамида DMBOK2 (рис. 4) разработана Питером Айкеном. Схема включает все области знаний об управлении данными из DAMA DMBOK2, учитывает то, как в компаниях обычно начинают работать с данными (без стратегии управления данными). Пирамида помогает понять, на каком этапе компания находится и как двигаться дальше, чтобы выстроить надежные процессы управления данными.

РИСУНОК 4
ПИРАМИДА АЙКЕНА. ПРИОБРЕТЕННЫЕ ИЛИ СОЗДАННЫЕ ВОЗМОЖНОСТИ СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ



4.1.2. Архитектура данных

Понятие и цели архитектуры данных

Архитектура — основные понятия или свойства системы в окружающей среде, воплощенные в ее элементах, отношениях и конкретных принципах ее проекта и развития[10].

Архитектура предприятия (Enterprise Architecture) охватывает архитектуры нескольких предметных областей (или доменов, domains), включая бизнес, данные, приложения и технологии. Эффективное управление данными и системами хранения и использования данных — одна из общих целей всех разделов науки об архитектуре предприятия.

Архитектура данных
определение потребностей организации в данных (безотносительно к структуре), а также разработка и сопровождение основных рабочих описаний решений по обеспечению таких данных. Основные рабочие описания используются в качестве руководящих материалов при осуществлении интеграции данных и контроля информационных активов, а также при согласовании инвестиций в области данных с бизнес-стратегией[11].

Данные обычно являются одним из нескольких доменов архитектуры, которые составляют основу архитектуры предприятия или архитектуры решения.

В DAMA-DMBOK[12] архитектура данных рассматривается с трех позиций:

- выходные результаты архитектуры данных (или артефакты архитектуры данных, включая модели, определения и описания потоков данных на различных уровнях);
- работы, проводимые в области архитектуры данных (формирование, развертывание и внедрение целевых решений в области архитектуры данных);
- организационное поведение в рамках работ по архитектуре данных (формы сотрудничества, навыки, распределение по различным ролям, имеющим отношение к архитектуре данных).

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

ЦЕЛИ АРХИТЕКТУРЫ ДАННЫХ

Цель архитектуры данных: служить мостом между бизнес-стратегией и ее технологической реализацией.

---

10 Стандарт ISO/IEC/IEEE 42010:2011 Systems and software engineering. Architecture description («Системная и программная инженерия. Описание архитектуры»).

11 DAMA-DMBOK: свод знаний по управлению данными / Dama International; пер. с англ. Г. Агафонова. 2-е изд. М.: Олимп-Бизнес, 2020. С. 113.

12 DAMA-DMBOK: свод знаний по управлению данными / Dama International; пер. с англ. Г. Агафонова. 2-е изд. М.: Олимп-Бизнес, 2020. С. 110.

Будучи частью архитектуры предприятия, архитектура данных должна:

- стратегически подготавливать организации к быстрому развитию продуктов, услуг и данных с целью полностью использовать бизнес-возможности, 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.

Ценности архитектуры данных

Архитектур данных разрабатывают для того, чтобы она служила 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.

Архитектуру стремятся сделать такой, чтобы она приносила организации ценность, а это достигается за счет оптимизации требуемых ресурсов, операционной и проектной эффективности, расширения возможностей организации по использованию данных. Чтобы добиться этого, нужны качественные проектирование и планирование, а также 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.

В результате разработки архитектуры данных должны быть соблюдены:

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

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 на них;
- использование архитектуры данных для определения и закрепления семантики организации с помощью создания общего бизнес-словаря.

Корпоративная архитектура данных[13]

Корпоративная архитектура данных (enterprise data architecture) определяет стандартные термины и проектные решения, применяемые в отношении важных для организации элементов. Проект корпоративной архитектуры данных включает согласованные между собой спецификации:

- корпоративной модели данных (Enterprise data model, EDM);
- описания потоков данных в трех состояниях — текущем, целевом (архитектурная перспектива) и переходном (проектная перспектива).

Корпоративная модель данных 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, карты данных и бизнес-правила. EDM создается путем объединения моделей предметных областей.

Потоки данных отображают и документируют взаимосвязи между данными и другими объектами:

- приложениями, используемыми в рамках бизнес-процесса;
- хранилищами или базами данных в среде функционирования;
- сегментами сети (полезно для описания мер безопасности);
- бизнес-ролями (показывают, какие роли отвечают за создание, чтение, обновление, удаление данных (четыре базовые операции, выполняемые при работе с данными — создание (Create), чтение (Read), обновление (Update), удаление (Delete), CRUD).

Для схематичного представления потоков данных можно использовать двумерные матрицы или диаграммы потоков данных, и то и другое наглядно показывает, в каких процессах создаются и используются данные каждой категории.

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. Созданная команда 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.

13 Самая известная рамочная структура архитектуры предприятия, «модель Захмана», была разработана Джоном Захманом в 1980-х годах. Подробнее см.: The Concise Definition of The Zachman Framework by: John A. Zachman // Zachman International. URL: https://zachman.com/about-the-zachman-framework

Создание архитектуры данных сопряжено с необходимостью учета сложного комплекса вопросов, поскольку относительно архитектурных решений возможны:

- ориентация на качество
- ориентация на инновации.

Стратегия
выбор рамочных моделей, формулировка подходов, создание дорожных карт.

Признание и культура
информирование и мотивирование к изменениям поведения.

Организация
распределение обязанностей, механизмы подотчетности.

Рабочие методы
лучшие практики и работы по архитектуре данных.

Результаты
артефакты архитектуры по дорожной карте.

Корпоративная архитектура данных влияет на содержание и границы проектов, а также на разрабатываемые системы:

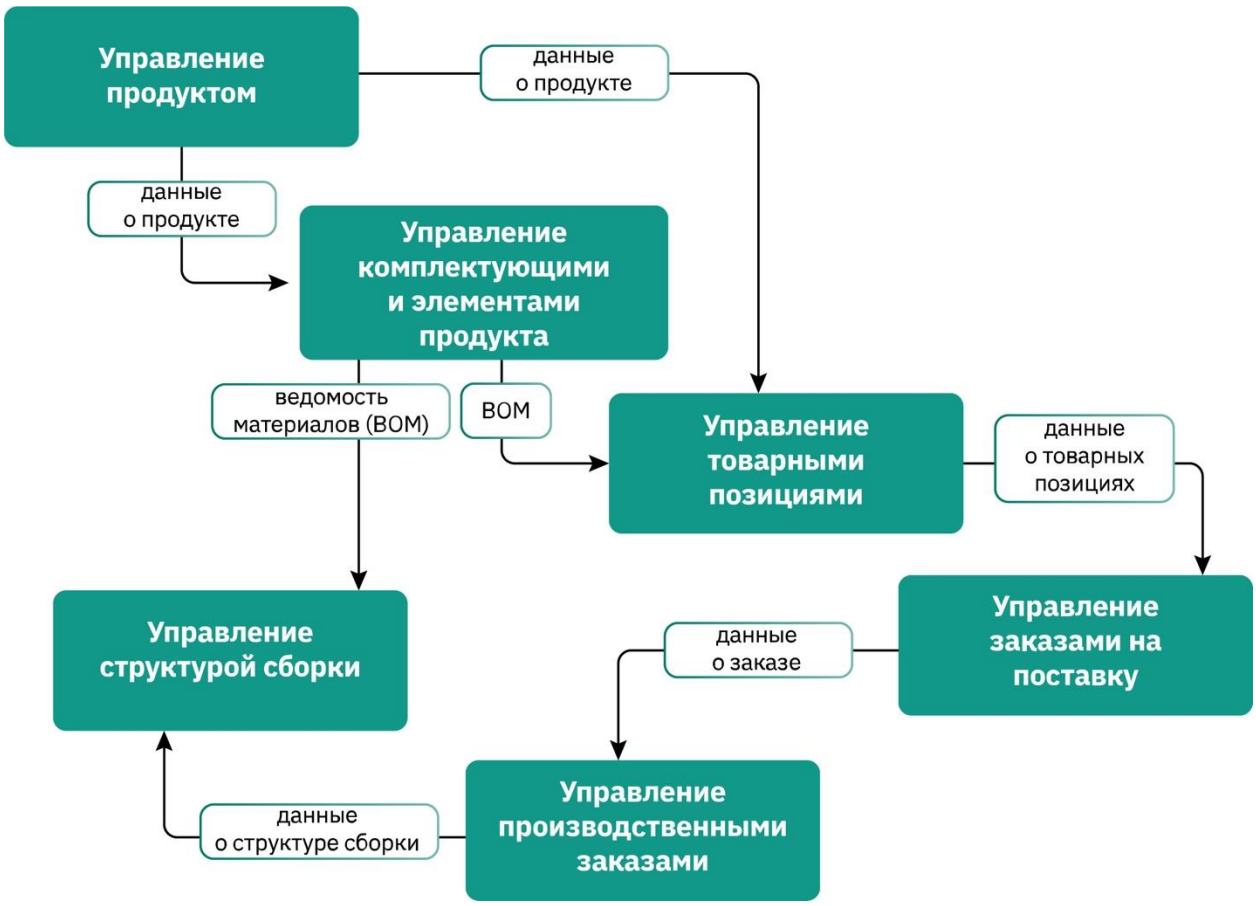
- определение проектных требований в области данных;
- проверка проектных решений в области данных;
- определение и учет факторов, оказывающих влияние на происхождение данных;
- контроль репликации данных;
- обеспечение соблюдения стандартов архитектуры данных;
- стимулирование использования новейших технологий работы с данными и решений по проведению обновлений.

Дорожная карта формирования корпоративной архитектуры данных:

- рассчитана на среднесрочную перспективу (3–5 лет);
- описывает, каким образом целевая архитектура превратится в реальность;
- строится с учетом оценки зрелости управления данными;
- 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.

Управляемая на основе данных (data-driven) дорожная карта начинается с мероприятий, относящихся к наиболее независимым бизнес-возможностям (не зависящих от результатов другой деятельности), и заканчивается зависимыми направлениями работ. Последовательность проработки бизнес-возможностей соответствует общему порядку возникновения бизнес-данных (рис. 5).

РИСУНОК 5
ЗАВИСИМОСТИ БИЗНЕС-ВОЗМОЖНОСТЕЙ В ОТНОШЕНИИ ДАННЫХ



УПРАВЛЕНИЕ КОРПОРАТИВНЫМИ ТРЕБОВАНИЯМИ В РАМКАХ ПРОЕКТОВ

Архитекторы данных должны воспринимать требования с учетом общей архитектуры. По завершении работы над проектной спецификацией архитекторы данных должны определить:

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

Работы в рамках проектов по корпоративной архитектуре данных включают:

- определение содержания и границ проекта;
- углубление понимания бизнес-требований;
- проектирование;
- реализацию (с учетом покупки готовых приложений и повторного использования данных).

ИНСТРУМЕНТЫ И МЕТОДЫ ПРОЕКТОВ ПО КОРПОРАТИВНОЙ АРХИТЕКТУРЕ

1. Инструменты моделирования данных и репозитории моделей

Необходимы для управления корпоративной моделью данных на всех уровнях. Большинство инструментов моделирования данных обеспечивают функции отслеживания происхождения данных и связей между ними (это позволяет архитекторам управлять комплексами взаимосвязанных моделей, созданных с различными целями и на разных уровнях абстракции.

2. Программное обеспечение для управления ИТ-активами

Используется для учета систем, описания их составных частей и отслеживания связей между системами, а также для контроля соблюдения требований лицензионных соглашений в отношении ПО и сбора данных, связанных с активами.

3. Приложения для графического проектирования

Используются для построения архитектурных диаграмм моделей данных, потоков данных, цепочек создания стоимости данных и прочих архитектурных артефактов.

Проекции на фазы жизненного цикла

Необходимо четко документировать, к каким фазам жизненного цикла относятся конкретные продукты архитектурного проектирования (например, продукты на текущий период, период развертывания, стратегический период, отменяемые, предпочтительные, органического использования, перспективные, на рассмотрении и т. д.).

Четкость и ясность графических представлений

В визуализации архитектуры, как и в визуализации данных, важны:

- четкая и последовательная легенда;
- соответствие между всеми объектами диаграммы и легендой;
- четкое и последовательное направление линий;
- единообразное отображение пересечений линий;
- единообразное отображение объектов каждого типа;
- выравнивание и симметрия.

ШАГИ ПО ВНЕДРЕНИЮ КОРПОРАТИВНОЙ АРХИТЕКТУРЫ ДАННЫХ

Внедрение корпоративной архитектуры данных требует, чтобы были последовательно выполнены:

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

Целесообразно приступать к внедрению корпоративной архитектуры данных сначала в какой-то одной бизнес-области или подразделении организации, по мере развития навыков и повышения уровня зрелости можно расширять сферу внедрения. Подход к внедрению корпоративной архитектуры данных должен быть гибким (agile). Он может предусматривать, например, верхнеуровневую концептуальную модель предметной области, гибко и оперативно дополняемую по мере необходимости детализациями на локальных уровнях в рамках спринтов Agile, тогда развитие происходит пошагово, за счет небольших приращений. Но 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

ОСНОВНЫЕ РИСКИ ПРИ ПОСТРОЕНИИ КОРПОРАТИВНОЙ АРХИТЕКТУРЫ ДАННЫХ

Построение корпоративной архитектуры данных сопряжено со следующими рисками:

- недостаточная поддержка со стороны руководства;
- отсутствие документально подтвержденных достижений;
- настороженность или обеспокоенность спонсора проекта;
- контрпродуктивные решения руководства;
- культурный шок;
- неопытный руководитель проекта;
- отсутствие комплексного подхода.

ОРГАНИЗАЦИОННЫЕ И КУЛЬТУРНЫЕ ИЗМЕНЕНИЯ

Скорость, с которой организация осваивает архитектурную практику, зависит от того, насколько адаптивна ее культура. На последнее влияют:

- восприимчивость организационной культуры к архитектурному подходу (выработка культуры, ориентированной на архитектурные решения);
- признание организацией данных как бизнес-актива, а не только как объекта заботы ИТ-специалистов;
- способность организации отойти от одностороннего 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.

РУКОВОДСТВО АРХИТЕКТУРОЙ ДАННЫХ

Архитекторы данных часто выполняют роль связующего звена между бизнесом и деятельностью по руководству данными, поэтому организационные системы по архитектуре и руководству данными должны действовать согласованно. В идеальной ситуации за каждой предметной областью должны быть закреплены ответственные архитектор данных и распорядитель данных

Работы в области руководства архитектурой данных включают:

- надзор за проектами;
- управление архитектурными решениями, жизненным циклом и инструментами;
- определение стандартов;
- определение артефактов, относящихся к данным.

Национальная система управления данными

Национальная система управления данными (НСУД)
совокупность взаимосвязанных элементов информационно-технологического, организационного, методологического, кадрового и нормативно-правового характера, обеспечивающая достижение целей и выполнение задач по эффективному управлению государственными данными[14].

На ее примере можно рассмотреть архитектуру управления данными. Вся имеющаяся у государства информация о гражданах, компаниях, товарах и объектах будет включена в единую информационную среду, которая объединит сотни государственных информационных систем (ГИС). Создание НСУД) предусмотрено федеральным проектом «Цифровое государственное управление» национальной программы «Цифровая экономика Российской Федерации». НСУД — ключевой элемент госуправления в рамках нацпроекта «Цифровая экономика». Она предполагает новый подход к госданным — единый порядок их сбора, хранения и использования.

ЦЕЛЬ И ЗАДАЧИ НСУД

Основной целью создания и обеспечения функционирования НСУД является повышение эффективности создания, сбора и использования государственных данных как для предоставления государственных и муниципальных услуг и осуществления государственных и муниципальных функций, так и для обеспечения потребности физических и юридических лиц в доступе к информации.

Для достижения указанной цели будут задействованы нормативные правовые, методологические, информационно-технологические, организационные и кадровые механизмы. Планируется обеспечить:

- повышение доступности государственных данных;
- полноту, актуальность, непротиворечивость и связанность государственных данных;
- информационную безопасность.

НСУД[15] предназначена для управления информацией, содержащейся в государственных информационных ресурсах и для осуществления информационного обмена между органами и организациями государственного сектора, иными организациями, уполномоченными на формирование и ведение информационных ресурсов, содержащих государственные данные (далее — поставщики данных) и / или обладающих правом доступа к таким государственным данным с учетом их полномочий (далее — потребители данных), в соответствии с законодательством Российской Федерации, законами и иными нормативными правовыми актами субъектов Российской Федерации, муниципальными правовыми актами.

Управление информационным обменом между поставщиками и потребителями данных осуществляется средствами федеральной государственной информационной системы «Единая информационная платформа Национальной системы управления данными» (ФГИС «ЕИП НСУД»).

Реализация витрин данных в составе ИТ-инфраструктуры поставщика и потребителя данных осуществляется с использованием типового программного обеспечения витрины

14 Распоряжение Правительства РФ от 03.06.2019 № 1189-р «Концепция создания и функционирования национальной системы управления данными» // Официальный интернет-портал правовой информации. URL: http://pravo.gov.ru/proxy/ips/?docbody=&nd=102559304&intelsearch=%25D0%25E0%25F1%25EF%25EE%25F0%25FF%25E6%25E5%25ED%25E8%25E5+%25CF%25F0%25E0%25E2%25E8%25F2%25E5%25EB%25FC%25F1%25F2%25E2%25E0+%25D0%25EE%25F1%25F1%25E8%25E9%25F1%25EA%25EE%25E9+%25D4%25E5%25E4%25E5%25F0%25E0%25F6%25E8%25E8+%25B91189-%25F0.

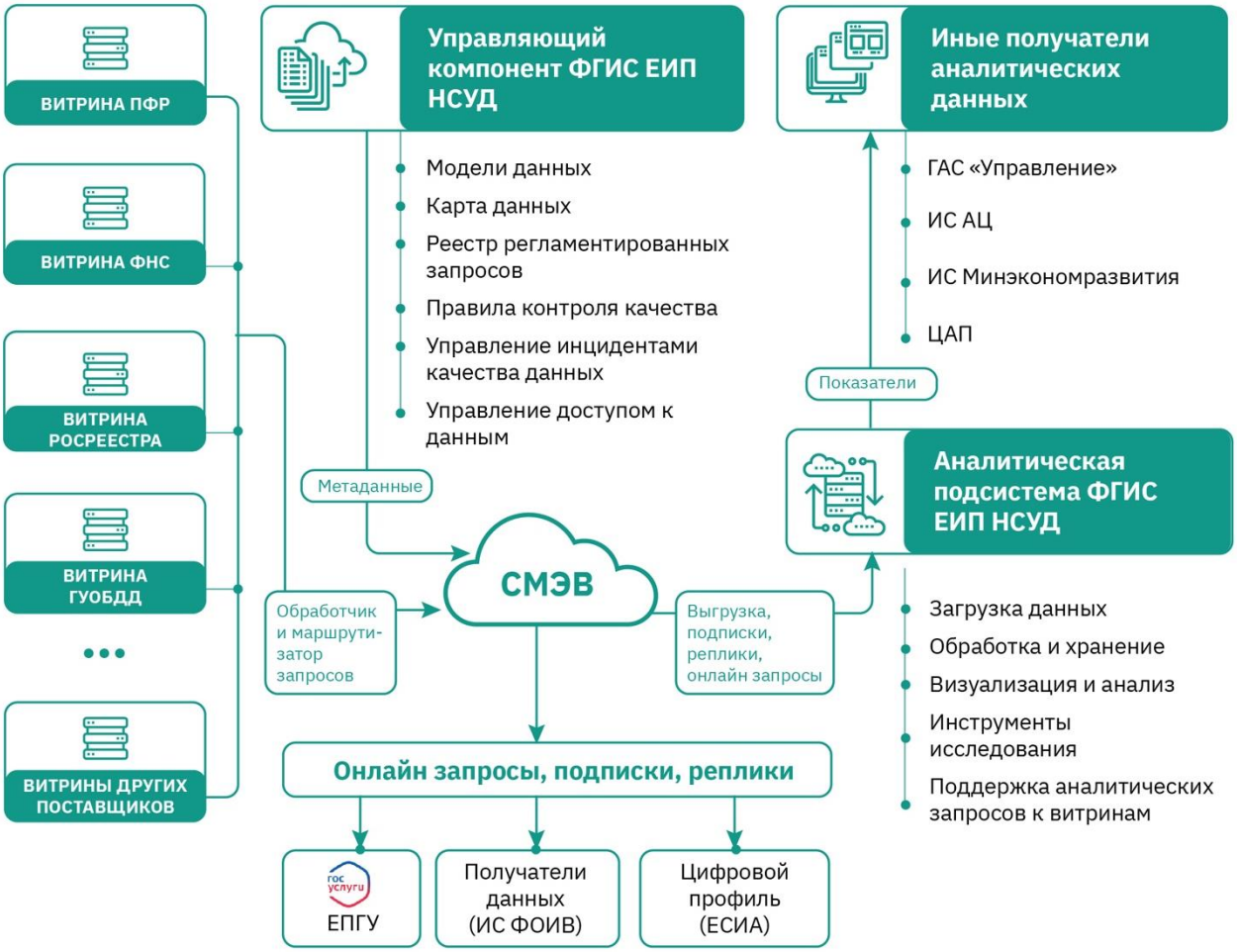
15 НСУД. URL: https://nsud.info.gov.ru/ifp/portals

данных. Типовую витрину данных можно получить в виде исходных кодов или через Государственную единую облачную платформу.

Типовое ПО витрины данных состоит из настраиваемого набора модулей. Конкретная конфигурация Витрины данных определяется поставщиком или потребителем данных на этапе реализации витрины данных в составе своей ИТ-инфраструктуры.

Информационное взаимодействие витрины данных и Подсистема обеспечения доступа к данным СМЭВ (ПОДД СМЭВ) в части передачи (получения) данных осуществляется через информационно-технологический компонент СМЭВ (адаптер СМЭВ), который устанавливается на стороне поставщика или потребителя данных, с использованием защищенных каналов передачи данных СМЭВ (рис. 6).

РИСУНОК 6
АРХИТЕКТУРА НСУД



4.1.3. Моделирование и проектирование данных

Определение моделирования данных

Моделирование данных
процесс выявления, анализа и формулирования требований к данным с последующим их представлением и распространением в точно определенной форме, называемой моделью данных. Этот процесс носит итерационный характер и может включать разработку концептуальной, логической и физической моделей[16].

Моделирование данных — это метод определения бизнес-требований к базе данных. Иногда это называют моделированием базы данных, потому что модель данных в конечном итоге реализуется в базе данных.

Виды моделируемых данных

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

1. Информация о категориях.

Это 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.

2. Информация о ресурсах.

Это 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».

3. Информация о бизнес-событиях.

Это 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».

4. Детальная информация о транзакциях.

Это 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.). Подобную детализированную информацию можно агрегировать и использовать с целью получить другие данные и провести анализ 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.

---

[16] DAMA-DMBOK: свод знаний по управлению данными / Dama International; пер. с англ. Г. Агафонова. 2-е изд. М.: Олимп-Бизнес, 2020. С. 142.

Концептуальные, логические и физические схемы

В 1975 году Американский национальный институт стандартов (англ. American national standards institute, ANSI) описал три типа схем модели данных:

1. **Концептуальная схема**

Описывает семантику предметной области (область действия модели). Например, это может быть модель области интересов организации или отрасли. Он состоит из классов сущностей, представляющих виды значимых в предметной области, и утверждений отношений об ассоциациях между парами классов сущностей. Концептуальная схема определяет виды фактов или предположений, которые могут быть выражены с помощью модели. В этом смысле она определяет разрешенные выражения на искусственном «языке» с областью действия, которая ограничена областью действия модели. Проще говоря, концептуальная схема — это первый шаг для организации требований к данным.

2. **Логическая схема**

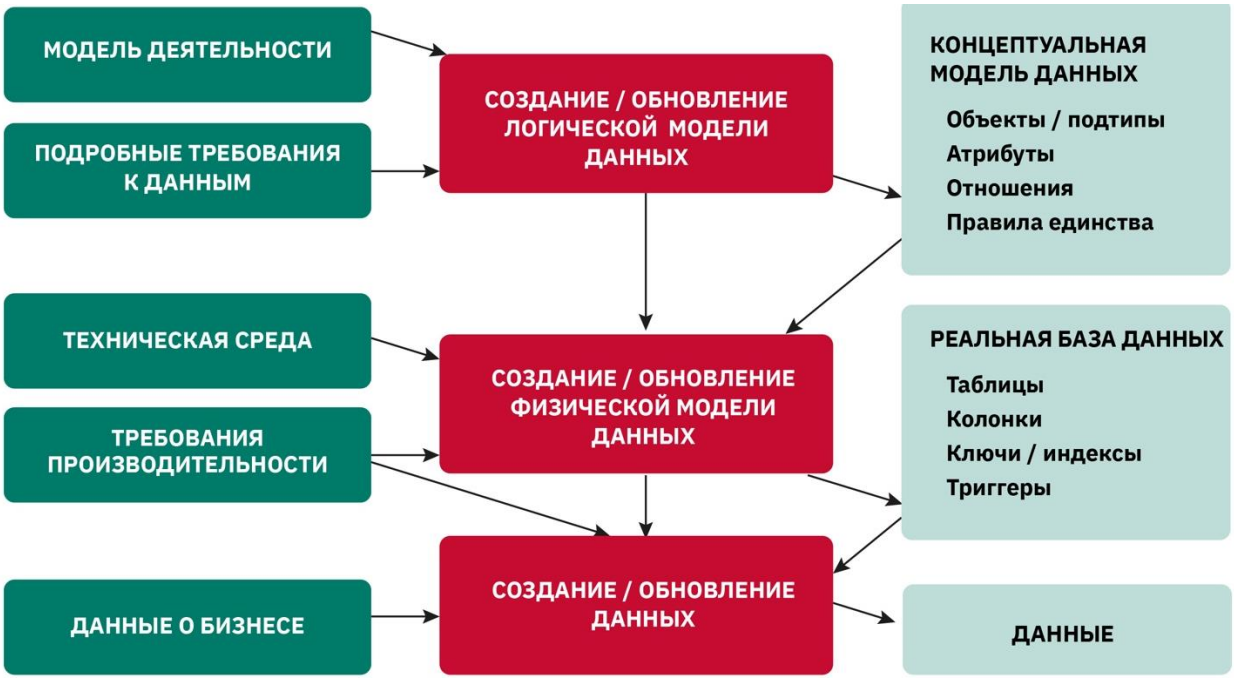
Описывает структуру некоторой области информации. Она состоит из описаний, например таблиц, столбцов, объектно ориентированных классов и тегов XML. Логическая схема и концептуальная схема иногда реализуются как одно и то же.

3. **Физическая схема**

Описывает физические средства, используемые для хранения данных. Это касается разделов, табличных пространств и т. д. Технология хранения может изменяться, не влияя ни на логическую, ни на концептуальную схему. Структуру таблицы / столбца можно изменять без влияния на концептуальную схему. В каждом случае, конечно, структуры должны оставаться согласованными для всех схем одной и той же модели данных.

На рис. 7 показано, как модели данных разрабатываются и используются сегодня.

РИСУНОК 7
АРХИТЕКТУРА НСУД



Концептуальная модель данных разрабатываются на основе данных требований для приложения, которое разрабатывается, возможно, в контексте модели деятельности. Модель данных обычно состоит из типов сущностей, атрибутов, отношений, правил

целостности и определений этих объектов. Затем модель используется в качестве отправной точки для проектирования интерфейса или базы данных.

Схемы представления данных при моделировании

Шесть наиболее распространенных схем представления данных:

- реляционная;
- многомерная;
- объектно- ориентированная;
- на основе фактов;
- хронологическая;
- NoSQL.

Для каждой схемы существуют собственные варианты формализованных систем условных обозначений (нотаций, notations) для построения диаграмм (табл. 1).

ТАБЛИЦА 1
ПРИМЕРЫ НОТАЦИЙ СХЕМ ПРЕДСТАВЛЕНИЯ ДАННЫХ

СХЕМА	ПРИМЕР 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 нотация; – колоночная нотация; – графовая нотация; – ключ-значение.

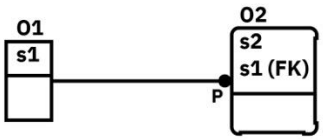
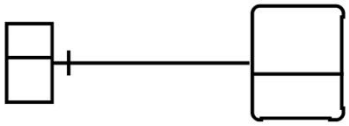
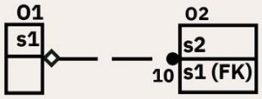
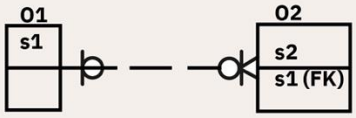
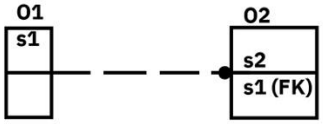
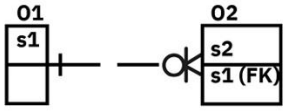
РЕЛЯЦИОННАЯ СХЕМА

Данная 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.

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).

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 (рис. 8).

РИСУНОК 8
РАЗЛИЧИЯ В ОТОБРАЖЕНИИ СВЯЗЕЙ В НОТАЦИЯХ IDEFIX И IE

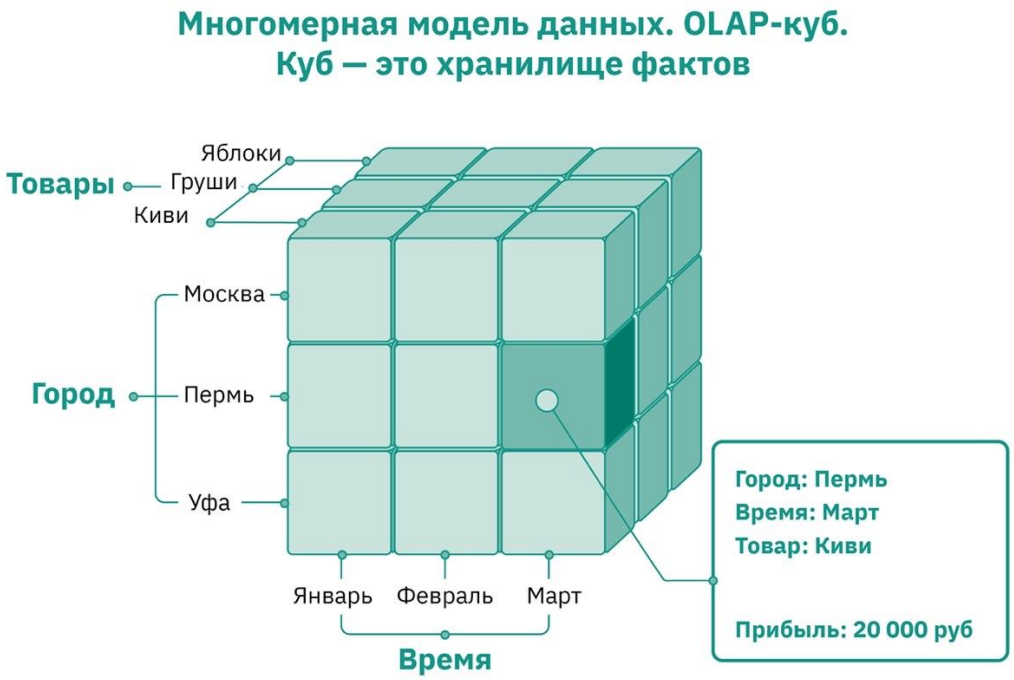
Вид связи	Изображение в IDEFIX	Изображение в IE
Идентифицирующая связь Exactly		
Неидентифицирующая связь Zero, One or Many, Nulls allowed		
Неидентифицирующая связьZero, One or Many, No nulls		

МНОГОМЕРНАЯ СХЕМА

Концепция многомерного моделирования была разработана в 1960-х годах в рамках совместного проекта корпорации General Mills и Дартмутского колледжа.

В многомерных моделях (рис. 9) 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.

РИСУНОК 9
МНОГОМЕРНАЯ МОДЕЛЬ ДАННЫХ



ОБЪЕКТНО- ОРИЕНТИРОВАННОЕ МОДЕЛИРОВАНИЕ И ЯЗЫК UML

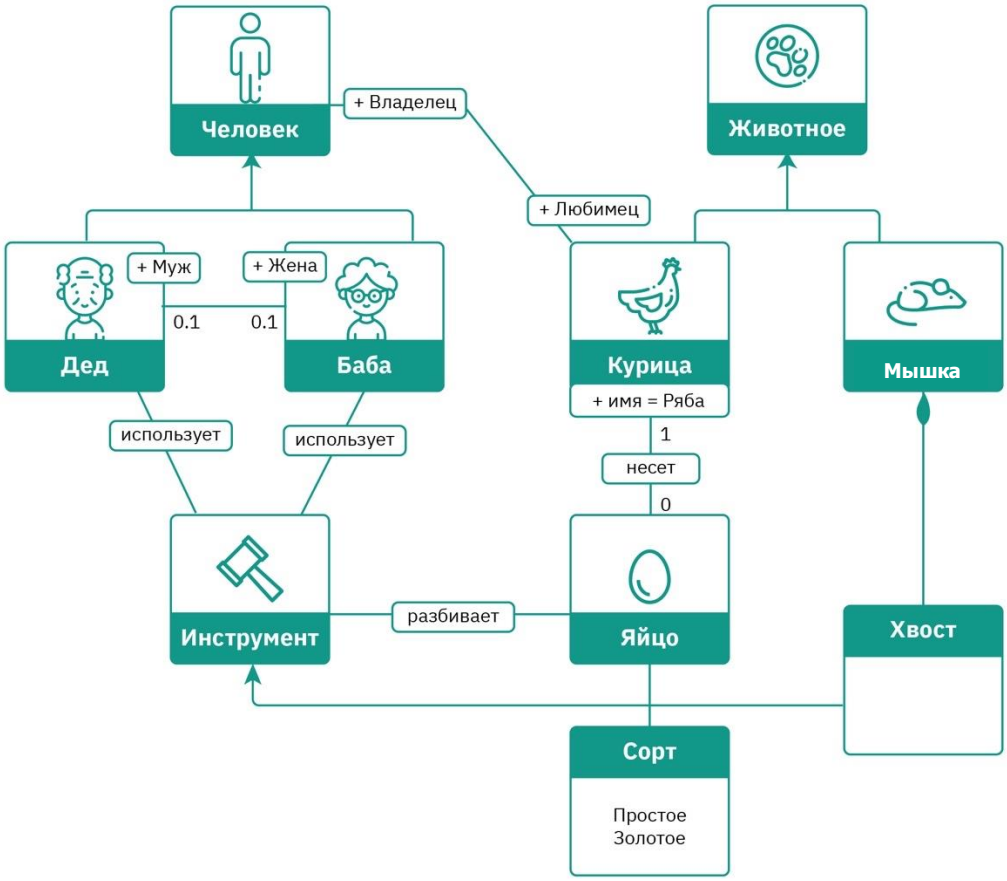
Унифицированный язык моделирования (Unified Modeling Language, UML) — графический язык, предназначенный для моделирования программного обеспечения. Одна из представленных в UML нотаций (модель классов) предназначена для моделирования баз данных. Модель определяет классы (типы сущностей) и взаимосвязи между ними, описывающие отношения различного типа. И наконец, доказательство того, что UML можно использовать для чего угодно, в том числе и для записи сказок: на рис. 10 приведена диаграмма, описывающая предметную область сказки о Курочке Рябе.

МОДЕЛИРОВАНИЕ НА ОСНОВЕ ФАКТОВ

Моделирование на основе фактов (Fact-Based Modeling, FBM) — общий подход, который зародился в конце 1970-х годов и охватывает семейство языков для концептуального моделирования. 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 характеризующих 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[17].

17 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.

РИСУНОК 10
ДИАГРАММА UML, ОПИСЫВАЮЩАЯ ПРЕДМЕТНУЮ ОБЛАСТЬ СКАЗКИ О КУРОЧКЕ РЯБЕ



ХРОНОЛОГИЧЕСКИЕ СХЕМЫ

Хронологические (time-based) схемы представления данных используются при необходимости их упорядочения в хронологическом порядке с привязкой к конкретным моментам времени, например 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 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.

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.

Свойства данных

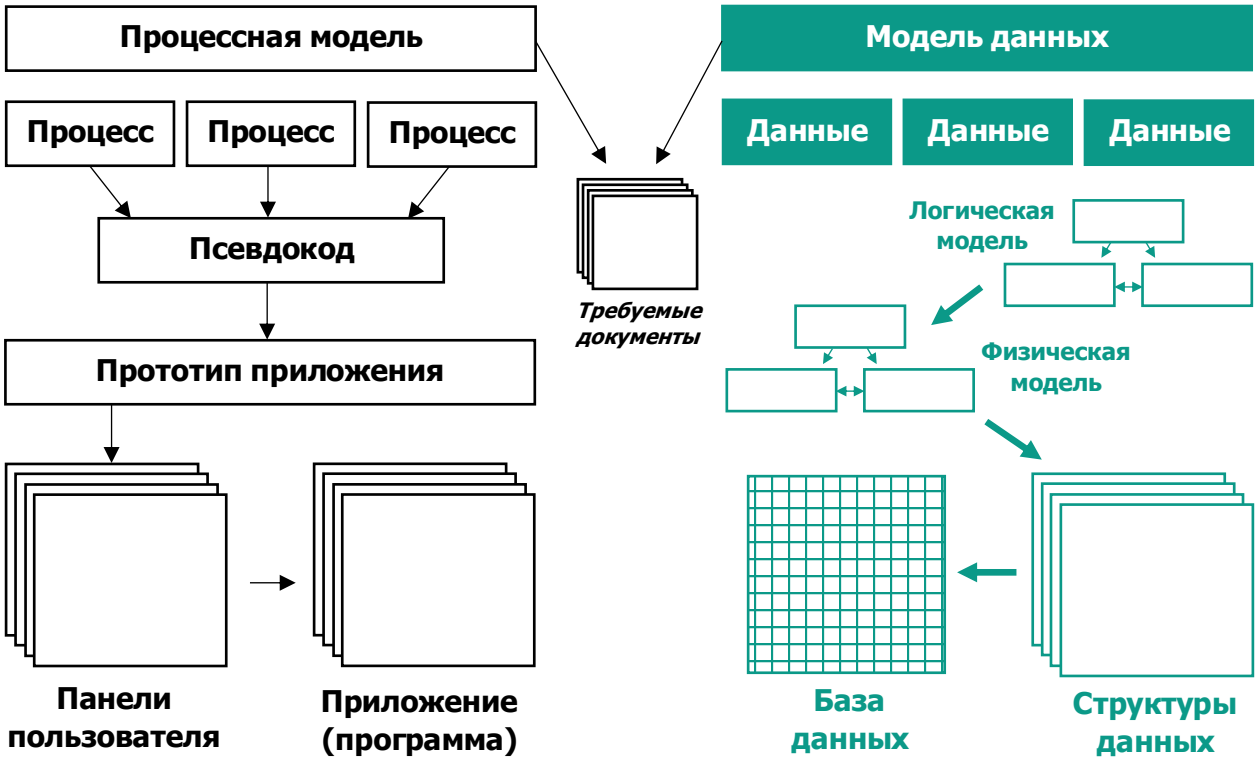
Некоторые важные свойства данных, для которых необходимо выполнить требования:

- свойства, связанные с определением:
  - *релевантность:* полезность данных в контексте вашего бизнеса;
  - *ясность:* наличие четкого и единого определения данных;
  - *согласованность:* совместимость одного и того же типа данных из разных источников;
- свойства, связанные с контентом:
  - *своевременность:* доступность данных в нужное время и актуальность этих данных;
  - *точность:* насколько близки к истине данные и их свойства, относящиеся как к определению, так и к содержанию;
  - *полнота:* сколько требуемых данных доступно;
  - *доступность:* где, как и кому данные доступны или недоступны (например, безопасность);
  - *стоимость:* затраты, понесенные при получении данных и предоставлении их для использования.

Процесс моделирования данных

Моделирование данных в контексте интеграции бизнес-процессов (рис. 11) дополняет моделирование бизнес-процессов и в конечном итоге приводит к созданию базы данных[18].

РИСУНОК 11
ИНТЕГРАЦИЯ БИЗНЕС-ПРОЦЕССОВ



Процесс проектирования базы данных включает создание описанных ранее трех типов схем — концептуальной, логической и физической. Проект базы данных, задокументированный в этих схемах, преобразуется с помощью языка определения данных, который затем можно использовать для создания базы данных. Полностью атрибутированная модель данных содержит подробные атрибуты (описания) для каждой сущности в ней. Термин «проект базы данных» может описывать множество различных частей дизайна всей системы базы данных. В принципе, это можно рассматривать как логический дизайн базовых структур данных, используемых для хранения данных. В реляционной модели это таблицы и представления. В объектной базе данных сущности и отношения отображаются непосредственно на классы объектов и именованные отношения. Однако термин «проектирование базы данных» можно также использовать для применения к общему процессу проектирования, к формам и запросам, используемым как часть общего приложения базы данных в системе управления базами данных или СУБД, а не только к базовым структурам данных. При этом на системные интерфейсы приходится от 25 до 70% затрат на разработку и поддержку существующих систем. Основная причина такой стоимости заключается в том, что эти системы не используют общую модель данных.

Если модели данных разрабатываются для каждой системы, то не только один и тот же анализ повторяется в перекрывающихся областях, но и должен выполняться дальнейший анализ для создания интерфейсов между ними. Большинство систем внутри

[18] Моделирование данных // Википедия. URL: https://ru.qaz.wiki/wiki/Data_modeling

организации содержат одни и те же базовые данные, переработанные для определенной цели. Следовательно, эффективно спроектированная базовая модель данных может минимизировать переделки с минимальными модификациями для различных систем в организации.

Методики моделирования данных

Модели данных представляют интересующие информационные области. Хотя существует множество способов создания моделей данных выделяются только две методологии моделирования — нисходящая и восходящая.

Восходящие модели, или модели интеграции представлений, часто являются результатом реинжиниринга. Обычно они начинаются с существующих форм структур данных, полей на экранах приложений или отчетов. Эти модели обычно являются физическими, ориентированными на приложения и неполными с точки зрения предприятия. Они могут не способствовать обмену данными, особенно если они созданы без привязки к другим частям организации.

Нисходящие логические модели данных создаются абстрактным способом, получая информацию от людей, знакомых с предметной областью. Система может не реализовывать все объекты в логической модели, но модель служит точкой отсчета или шаблоном. Иногда для создания модели используют два метода: с учетом потребностей в данных и структуры приложения и путем последовательной ссылки на модель предметной области. К сожалению, во многих средах различие между логической моделью данных и физической моделью данных нечетко. Кроме того, некоторые инструменты CASE не делают различий между логическими и физическими моделями данных.

Универсальное моделирование данных

Общие модели данных являются обобщением обычных моделей данных. Они определяют стандартизованные общие типы отношений вместе с видами вещей, которые могут быть связаны таким типом отношений. Определение общей модели данных аналогично определению естественного языка.

Например, общая модель данных может определять типы отношений, такие как классификационное отношение, являющееся бинарным отношением между отдельной вещью и видом вещи (классом), и отношение «части — целое», являющееся бинарным отношением между двумя вещами, одна — с ролью части, другая — с ролью целого, независимо от того, какие вещи связаны между собой. При наличии расширяемого списка классов это позволяет классифицировать любую отдельную вещь и определять отношения «часть —целое» для любого отдельного объекта.

Путем стандартизации расширяемого списка типов отношений универсальная модель данных позволяет выражать неограниченное количество видов фактов и приближается к возможностям естественных языков. Вместе с тем обычные модели данных имеют фиксированную и ограниченную область видимости, поскольку создание (использование) такой модели позволяет выражать только те виды фактов, которые предопределены в модели.

Семантическое моделирование данных

Логическая структура данных СУБД, будь то иерархическая, сетевая или реляционная, не может полностью удовлетворить требования к концептуальному определению данных, поскольку она ограничена по объему и смещена в сторону стратегии реализации, используемой СУБД. Если семантическая модель данных не реализована в базе данных специально, выбор может немного повлиять на производительность.

Следовательно, необходимость определения данных с концептуальной точки зрения привела к развитию методов семантического моделирования данных. То есть методы определения значения данных в контексте их взаимосвязей с другими данными. Как показано на рис. 12, реальный мир символически определяется с точки зрения ресурсов, идей, событий и т. д. в физических хранилищах данных.

РИСУНОК 12
ИЛЛЮСТРАЦИЯ СИМВОЛИЗМА ДАННЫХ

Семантическая модель данных — это абстракция, которая определяет, как хранимые символы относятся к реальному миру. Таким образом, модель должна достоверно отражать реальный мир.

Семантическая модель данных может использоваться для многих целей, например:

- планирование ресурсов данных;
- создание общих баз данных;
- оценка программного обеспечения поставщика;
- интеграция существующих баз данных.

Общая цель семантических моделей данных — уловить больше смысла данных путем интеграции реляционных концепций с более мощными концепциями абстракции, известными в области искусственного интеллекта. Идея состоит в том, чтобы предоставить примитивы моделирования высокого уровня как неотъемлемую часть модели данных, чтобы облегчить представление ситуаций реального мира.

Инструменты работы с данными

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.

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 при перемещении сущностей линии связей 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 такие функции, как проверка соблюдения стандартов именования, проверка правописания, хранение метаданных (например, определений и происхождения), а также поддержка совместного использования (в частности, веб-публикаций).

2. Инструменты для отслеживания происхождения данных

Инструменты для отслеживания происхождения — это программное обеспечение, позволяющее собирать и вести сведения об источниках данных для каждого атрибута, включенного в модель, в структурированном виде. 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» можно рассчитать по исходным данным, поступающим из нескольких приложений, — и инструменты для отслеживания происхождения данных фиксируют и сохраняют всю эту информацию.

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.

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 должностях и более, этот факт будет зафиксирован как аномалия данных.

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 ограничены.

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, 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.

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 ее важнейших 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.

Лучшие практики проектирования баз данных[19]

В процессе проектирования и построения базы данных ее администратору следует постоянно иметь в виду следующие принципы PRISM:

1. 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.

2. Возможность повторного использования (Reusability)

Структура базы данных должна обеспечивать возможность использования данных по мере необходимости различными приложениями и в разных 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.

3. 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 следует постоянно контролировать, чтобы гарантировать незамедлительное выявление любых нарушений целостности и уведомление о них.

4. Безопасность (Security)

Достоверные и точные данные всегда должны быть 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 незамедлительно выявлять нарушения требований информационной безопасности и уведомлять о них.

5. 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.

19DAMA-DMBOK: свод знаний по управлению данными / Dama International; пер. с англ. Г. Агафонова. 2-е изд. М.: Олимп-Бизнес, 2020. 187 с.

КЛЮЧЕВАЯ РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Звонарев, С. В. Основы математического моделирования: учебное пособие / С. В. Звонарев. – Екатеринбург: Изд-во Уральского ун-та, 2019. – 112 с.

2. ГОСТ Р ИСО/МЭК ТО 10032-2007. Эталонная модель управления данными = Reference model of datamanagement: национальный стандарт Российской Федерации: издание официальное: утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 27 декабря 2007 г. № 573-ст.: введен впервые: дата введения 2008-09-01 / подготовлен Открытым акционерным обществом «Научно-исследовательский центр контроля и диагностики технических систем». – Москва: Стандартинформ, 2009. – 40 с.

3. Рузаков, А. А. Управление данными: учебное пособие / А. А. Рузаков. – Челябинск: Изд-во Челябинского гос. пед. ун-та, 2015. – 132 с.

4. Управление данными [извлечение из интерактивной справки для программы Qlik Sense] / QlikTechInternational AB. [S.l.,], 2019. – URL: https://help.qlik.com/ru-RU/sense/June2019/pdf/Управление%20данными.pdf (дата обращения 18.10.2021)

5. Управление данными // YouTube. URL: https://youtu.be/pIepQd4YSEQ (дата обращения 18.10.2021)

П.4.2. Обеспечение хранения, безопасности и этичного обращения с данными

4.2.1. Хранение данных

База данных: понятие, назначение

База данных (БД) – любая совокупность хранимых данных, вне зависимости от их структуры или содержания[20]. БД также понимается как «совокупность взаимосвязанных данных, организованных в соответствии со схемой базы данных таким образом, чтобы с ними мог работать пользователь»[21]. Однако, на наш взгляд, более полным является определение Д. Л. Осипова:

База данных (database)
это организованная совокупность совместно используемых логически связанных данных и описаний этих данных, относящаяся к определенной предметной области, предназначенная для удовлетворения информационных потребностей организации[22].

Данное определение базируется на следующих выводах:

- БД представляет собой долговременное хранилище данных, структура которого определяется на этапе проектирования БД.
- Находящиеся в БД данные являются общекорпоративным ресурсом, в котором не отдельные файлов, разбросанные по отделам и службам предприятия, а данные, собранные с минимальной избыточностью.
- В БД хранятся не только рабочие данные, описывающие какую-то предметную область, но и метаданные, предназначенные для описания структуры самих данных. Такую важную способность БД часто называют самодокументированием.
- Потребителем данных может выступать не только человек, но и программно-аппаратные комплексы (станки с числовым программным управлением, автоматизированные системы управления, робототехнические системы и т. д.), именно поэтому в определении говорится об удовлетворении информационных потребностей не просто пользователей, а организации в целом.

Классификации баз данных:

- по модели данных;
- по степени распределенности.

Существует огромное количество разновидностей БД, отличающихся по различным критериям. Рассмотрим основные из них.

---

20 DAMA-DMBOK: свод знаний по управлению данными. Второе издание / Dama International; пер. с англ. Г. Агафонова. 2-е изд. М.: Олимп-Бизнес, 2020. С. 201.
21 ГОСТ 34.321–96 Информационные технологии. Система стандартов по базам данных. Эталонная модель управления данными // Электронный фонд правовых и нормативно-технических документов. URL: https://docs.cntd.ru/document/1200017662
22 Осипов Д. Л. Технологии проектирования баз данных. М.: ДМК Пресс, 2019. 498 с.

Классификация по модели данных

Центральным понятием в области БД является понятие модели.

Модель данных
это некоторая абстракция, которая, будучи приложима к конкретным данным, позволяет пользователям и разработчикам трактовать их уже как информацию, то есть сведения, содержащие не только данные, но и взаимосвязь между ними.

Виды:
- иерархическая;
- объектная и объектно ориентированная;
- объектно-реляционная;
- реляционная;
- сетевая;
- функциональная.

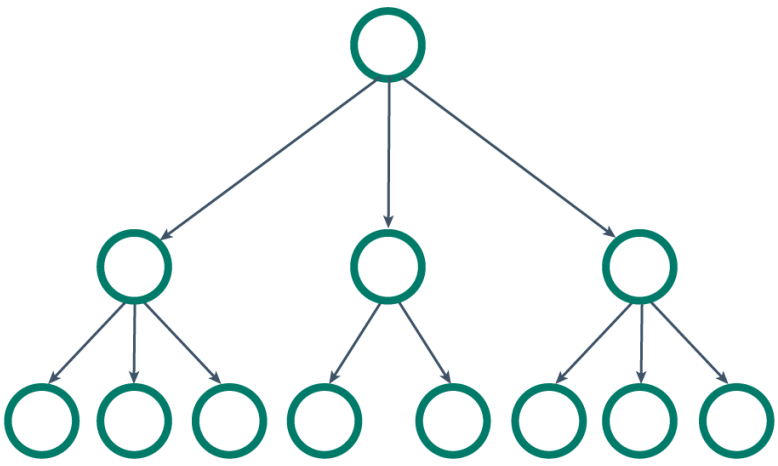
ИЕРАРХИЧЕСКАЯ БАЗА ДАННЫХ

Это совокупность, где каждый объект представляется в виде определенной сущности, то есть у этой сущности могут быть дочерние элементы, родительские элементы, а у тех дочерних могут быть свои дочерние элементы, но есть один объект, с которого все начинается. Получается своеобразное дерево. Примером иерархической базы данных может быть документ в формате XML или файловая система компьютера.

Следует сказать, что базы данных подобного вида оптимизированы под чтение информации, то есть БД, имеющие иерархическую структуру, умеют очень быстро выбирать запрашиваемую информацию и выдавать ее пользователям. Но такая структура не позволяет столь же быстро перебирать информацию. Здесь можно привести первый пример: компьютер может легко работать с каким-либо конкретным файлом или папкой (которые, по сути, являются объектами иерархической структуры), но проверка компьютера антивирусом осуществляется очень долго. Второй пример – реестр Windows.

На рис. 13 вы можете увидеть структуру иерархической базы данных. В самом верху находится родитель, или корневой элемент, ниже находятся дочерние элементы. Элементы, находящиеся на одном уровне, называются братьями, или соседними элементами. Соответственно, чем ниже уровень элемента, тем вложенность этого элемента больше.

РИСУНОК 13
СТРУКТУРА ИЕРАРХИЧЕСКОЙ БАЗЫ ДАННЫХ



ОБЪЕКТНЫЕ БАЗЫ ДАННЫХ

Это модель работы с объектными данными. Такая модель баз данных считается новой, хотя она существует уже много лет. Ее создание открывает большие перспективы, в связи с тем, что использование объектной модели баз данных легко воспринимается пользователем, так как создается высокий уровень абстракции. Объектная модель идеально подходит для трактовки такого рода объектных данных, как изображение, музыка, видео, тексты разных видов.

ОБЪЕКТНО-ОРИЕНТИРОВАННАЯ БАЗА ДАННЫХ

Это БД, в которой данные моделируются в виде объектов, их атрибутов, методов и классов. Объектно-ориентированные БД обычно рекомендованы для тех случаев, когда требуется высокопроизводительная обработка данных, имеющих сложную структуру.

ОБЪЕКТНО-РЕЛЯЦИОННЫЕ СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ (СУБД)

Эти базы данных объединяют в себе черты реляционной и объектной моделей. Их возникновение объясняется тем, что реляционные БД хорошо работают со встроенными типами данных и гораздо хуже — с пользовательскими, нестандартными. Когда появляется новый важный тип данных, приходится либо включать его поддержку в СУБД, либо заставлять программиста самостоятельно управлять данными в приложении.

Не всякую информацию имеет смысл интерпретировать в виде цепочек символов или цифр. Представим себе музыкальную базу данных. Песню, закодированную в виде аудиофайла, можно поместить в текстовое поле большого размера, но как в таком случае будет ли осуществляться текстовый поиск?

РЕЛЯЦИОННАЯ (ИЛИ ТАБЛИЧНАЯ) БАЗА ДАННЫХ

Содержит перечень объектов одного типа, то есть объектов с одинаковым набором свойств. Такую базу удобно представлять в виде двумерной таблицы (или, чаще всего, нескольких связанных между собой таблиц). Примером такой таблицы может служить БД «Учащиеся», представляющая собой перечень объектов (учеников), каждый из которых имеет фамилию, имя, отчество, дату рождения, класс, номер личного дела и др.

Столбцы такой таблицы называют полями; каждое поле характеризуется своим именем (названием соответствующего свойства объекта) и типом данных, которые это поле может хранить. Каждое поле обладает определенным набором свойств (размер, формат и т. п.). Таким образом, поле БД — это столбец таблицы, содержащий значения определенного свойства объектов.

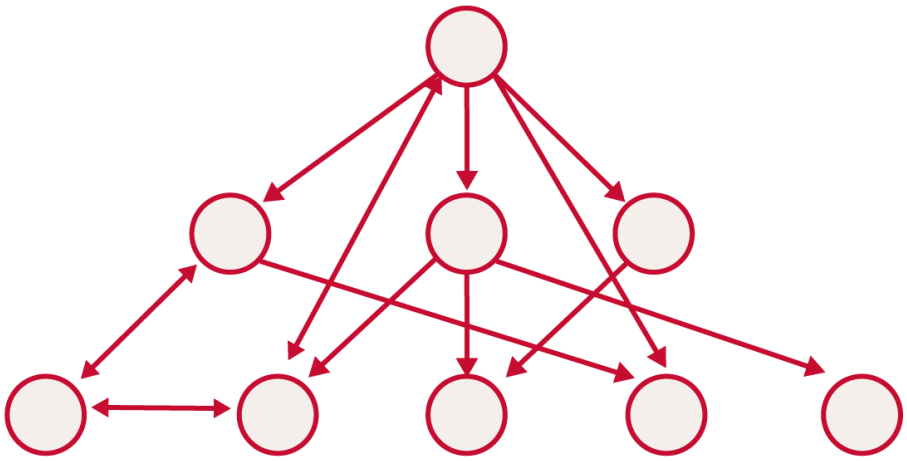
Строки таблицы являются записями. Записи разбиты на поля. Каждая строка таблицы содержит запись об одном единственном объекте, включая все его свойства.

В каждой таблице должно быть хотя бы одно ключевое поле, содержимое которого уникально для любой записи в этой таблице. Значения ключевого поля однозначно определяют каждую запись в таблице. В приведенном выше примере ключевым полем может являться поле «Номер личного дела». Очень часто в качестве ключевого поля используется поле, содержащее данные типа «счетчик».

СЕТЕВЫЕ БД

Являются своеобразной модификацией иерархических БД. Если вы внимательно смотрели на изображение выше, то, скорее всего, обратили внимание, что к каждому нижнему элементу идет только одна стрелочка от верхнего элемента. В иерархических БД у каждого дочернего элемента может быть только один потомок. Сетевые БД отличаются от иерархических тем, что у дочернего элемента может быть несколько предков, то есть элементов, стоящих выше него (рис. 14).

РИСУНОК 13
СТРУКТУРА СЕТЕВОЙ БАЗЫ ДАННЫХ



Стоит заметить, что сетевые БД обладают примерно теми же характеристиками, что и иерархические БД. Но сейчас нас не особо интересуют иерархические и сетевые БД, данная тема больше относится к формату XML.

ФУНКЦИОНАЛЬНЫЕ БД

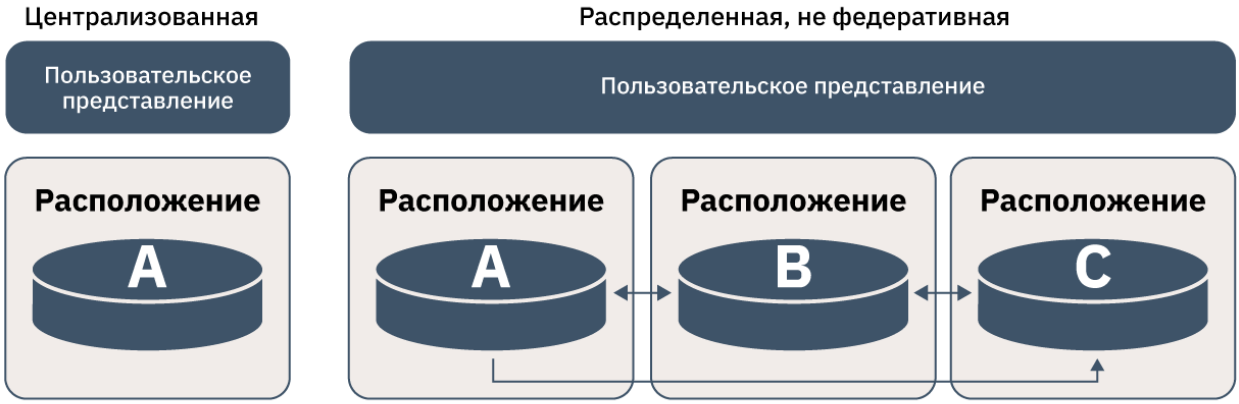
Используются для решения аналитических задач: финансовое моделирование и управление производительностью. Функциональная БД, или функциональная модель, отличается от реляционной модели. Функциональная модель также отличается от других аналогично названных концепций, включая модель функциональной БД DAPLEX и бБД функциональных языков.

Функциональная модель является частью категории оперативной аналитической обработки, поскольку включает многомерное иерархическое объединение. Но такая модель выходит за рамки OLAP, требуя ориентирования ячейки, подобно тому, где ячейки могут быть введены или рассчитаны как функции других ячеек. Как и в электронных таблицах, данная модель поддерживает интерактивные вычисления, в которых значения всех зависимых ячеек автоматически обновляются каждый раз, когда изменяется значение ячейки.

Классификация по степени распределенности

Прежде всего, БД подразделяются на централизованные и распределенные. Централизованная система управляет одной БД, в то время как распределенная система управляет множеством БД, реализованных во множестве систем. Распределенные системы (рис. 14) можно разделить на два класса по степени автономности входящих в них компонентов: федеративные (автономные компоненты) и нефедеративные (неавтономные компоненты).

РИСУНОК 14
ЦЕНТРАЛИЗОВАННЯ И РАСПРЕДЕЛЕННАЯ АРХИТЕКТУРЫ БД



ЦЕНТРАЛИЗОВАННАЯ, ИЛИ СОСРЕДОТОЧЕННАЯ, БД (CENTRALIZED DATABASE)

Находится, хранится и поддерживается в одном месте. Чаще всего это центральный компьютер или система БД, например настольный или серверный центральный процессор (ЦП), или мэйнфрейм. В большинстве случаев централизованная БД будет использоваться организацией.

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.

РАСПРЕДЕЛЕННАЯ БД (DISTRIBUTED DATABASE)

БД такова, что ее составные части размещаются в различных узлах компьютерной сети в соответствии с каким-либо критерием.

Распределенные БД обеспечивают возможность быстрого доступа к данным через множество узлов. Популярные технологии распределенных БД 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. При разработке таких программ большое внимание 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 БД реализована модель распределенных вычислений MapReduce, способствующая еще большему повышению производительности за счет дробления запросов к БД 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.

НЕОДНОРОДНАЯ БД (HETEROGENEOUS DISTRIBUTED DATABASE)

Отличается тем, что фрагменты распределенной БД в разных узлах сети поддерживаются средствами более чем одной СУБД.

ОДНОРОДНАЯ БД (HOMOGENEOUS DISTRIBUTED DATABASE)

Состоит в том, что фрагменты распределенной БД в разных узлах сети поддерживаются средствами одной и той же СУБД.

ФРАГМЕНТИРОВАННАЯ ИЛИ СЕКЦИОНИРОВАННАЯ БД (PARTITIONED DATABASE)

Работает по принципу, в соответствии с которым методом распределения данных является фрагментирование (партиционирование, секционирование), вертикальное или горизонтальное.

ТИРАЖИРОВАННАЯ БД (REPLICATED DATABASE)

Отличается тем, что методом распределения данных является тиражирование.

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), центры обработки данных (ЦОД) и частные, публичные и гибридные облачные хранилища.

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.

БАЗЫ ДАННЫХ В ОПЕРАТИВНОЙ ПАМЯТИ

БД 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.

БД 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.

РЕШЕНИЯ СО СЖАТИЕМ ПО КОЛОНКАМ

Колоночные БД 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. Сжатие заключается в том, что в каждом столбце сохраняется не само значение, а указатель на одно из допустимых значений, содержащихся в отдельной таблице значений. За счет этого объем данных в основной таблице сжимается многократно.

ФЛЭШ-ПАМЯТЬ

Последние достижения в области технологий хранения данных отражает флэш-память, или твердотельные накопители (Solid State Drives, SSDs), ставшие привлекательной альтернативой жестким дискам. По скорости доступа они теперь практически не уступают решениям для оперативной памяти, а по надежности и долговечности — дисковым хранилищам.

Среды баз данных

На протяжении жизненного цикла разработки систем БД используются в различных средах. Для обеспечения тестирований изменений администраторы базы данных (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.

Большинство технологий управления данными основаны на использовании программного обеспечения, установленного на оборудовании общего применения, но 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.

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, которые потеряют доступ к поддерживаемым системой сервисам. Применительно к экстренным службам или системам 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, разработку и тестирование недопустимо проводить в эксплуатационной среде, поскольку это поставило бы под угрозу производственные процессы и сохранность данных.

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.

Основными типами предэксплуатационных сред являются среда разработки, среда тестирования и среды специального применения.

Общие процессы обработки данных в базах данных

В базах данных любого типа в том или ином виде реализованы следующие процессы:

1. Архивирование

Архивирование
это процесс перемещения данных из операционной среды на носители, не поддерживающие прямой доступ 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 носителях информации. При этом процедура восстановления данных из архива должна быть предельно простой и сводиться к копированию данных из архива обратно в систему.

2. Прогнозирование роста требуемой емкости БД

Представьте себе базу данных в образе коробки для фруктов, данные — в образе фруктов, а надстройку над данными (индексы и т.п.) — в роли обертки для фруктов. Коробка внутри разделена перегородками на ячейки, куда укладываются фрукты в обертке. А теперь задайте себе следующие вопросы:

- 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 определите, нужно или не нужно со временем увеличивать размеры и вместимость коробки. Соответственно, 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.

3. Регистрация изменений данных

Регистрация изменений данных (Change Data Capture, CDC)
процесс выявления факта изменений и сохранения исчерпывающей информации о них.

Процедура CDC, 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).

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 БД, 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.

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, распределение рабочей нагрузки между многими серверами или даже вычислительными центрами с идентичными БД 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 все большем географическом удалении от первичной реплики.

6. Отказоустойчивость и восстановление работоспособности

Применительно к БД 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 явно недостает отказоустойчивости. 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 обойти (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.

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 больше 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.

8. Сегментирование (шардинг)

Сегментирование (или шардинг, sharding)
это разбиение БД 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.

Функции системы управления базами данных

Программное обеспечение включает, прежде всего, систему управления базами данных (СУБД).

Система управления базами данных (СУБД)
комплекс программных средств, которые обслуживают базу данных и позволяют использовать содержащуюся в ней информацию многим пользователям[23].

Функции СУБД:

1. Управление данными во внешней памяти (позволяет пользователям выполнять основные операции с данными: сохранять, извлекать и обновлять информацию).
2. Управление транзакциями. К транзакциям относится совокупность операций над БД, рассматриваемых СУБД как единое целое. В то же время транзакция — это набор действий, которые могут быть выполнены с целью получить доступ или изменить содержимое БД. Когда необходимо внести в БД сразу несколько изменений, создается сложная транзакция. При использовании транзакций не нарушается целостное состояние БД.
3. Восстановление БД. СУБД должна быть в состоянии восстановить после любого аппаратного или программного сбоя последнее согласованное состояние БД. Это требует избыточности хранения данных. Наиболее распространенным методом поддержания такой избыточности является ведение журнала изменения БД.
4. Поддержка языков БД.
5. Ведение словаря данных. Словарь данных — системный каталог с данными о схемах, пользователях, приложениях и т.д.
6. Управление параллельным доступом. При одновременном доступе к обрабатываемым данным СУБД должна гарантировать, что пользователи не будут мешать друг другу, то есть не произойдут конфликты с нежелательными последствиями для хранимых данных.
7. Управление буферами оперативной памяти. СУБД зачастую функционирует с БД значительного объема. Поэтому при обращении к каждому элементу данных необходимо осуществлять обмен с внешней памятью, что вызывает значительное замедление работы системы. Реальное увеличение скорости возможно за счет буферизации данных в оперативной памяти, но возможности общесистемной буферизации при этом оказываются недостаточными. Поэтому СУБД поддерживает свой собственный набор буферов оперативной памяти с собственной дисциплиной замены буферов.
8. Контроль доступа к данным. СУБД должна поддерживать различный уровень доступа к данным разных категорий пользователей.
9. Поддержка обмена данными. СУБД поддерживает работу в локальных сетях и обладает возможностью объединения с коммуникационным программным обеспечением и с различными существующими менеджерами по обмену данными.
10. Поддержка целостности данных, которая предполагает, что СУБД содержит сведения обо всех правилах, которые ни в коем случае нельзя нарушать при

23 Костюкова А.П. Базы данных: общий курс: учеб. пос. М.: Издательский дом Академии естествознания, 2018. Ч. 1. 82 с.

работе с данными, и обладает инструментами контроля над данными и их изменения при условии соответствия этим правилам.

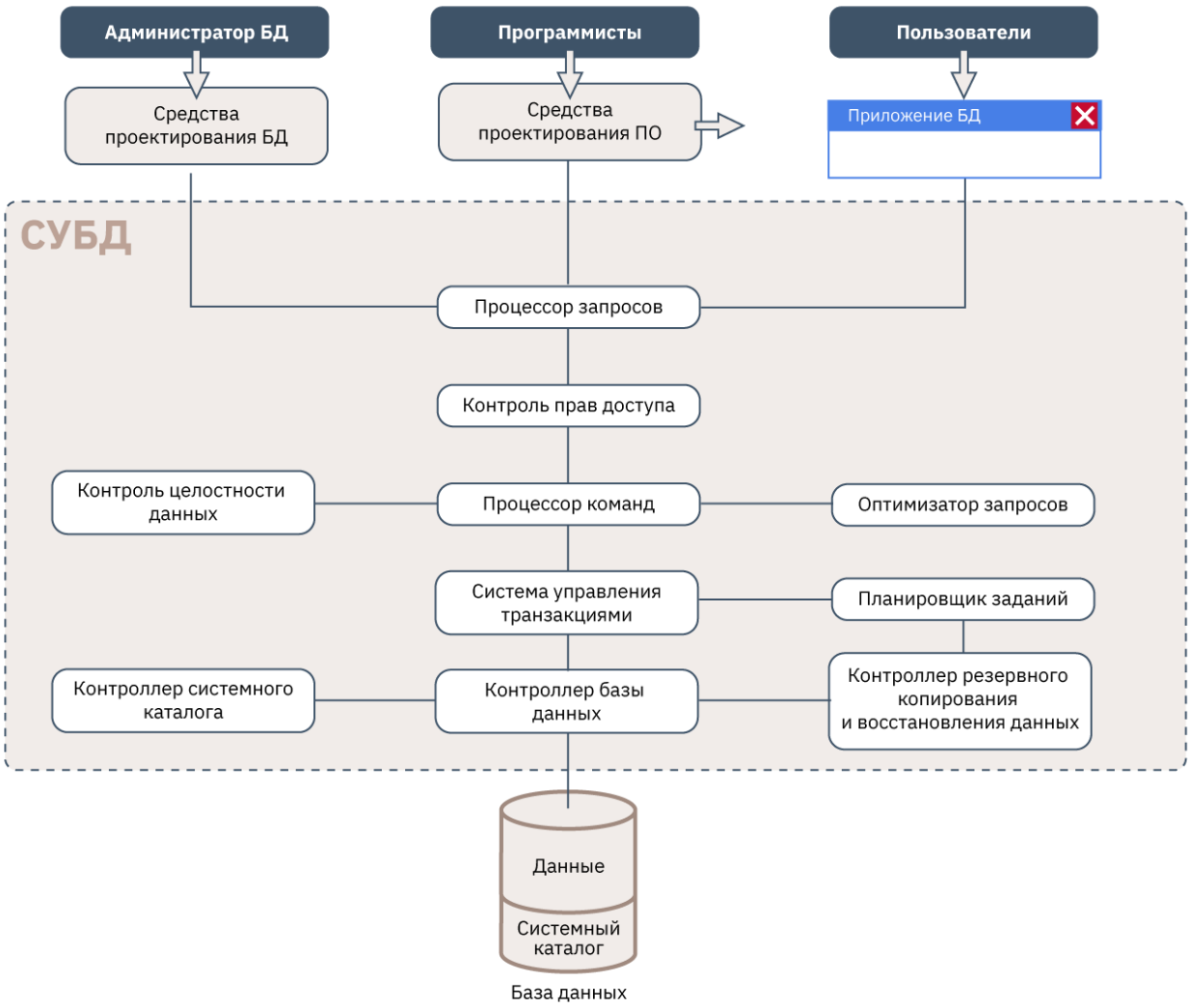
11. Поддержка независимости от данных, то есть независимость программ от структур данных. Основными средствами СУБД являются:
    - инструменты структурирования данных (языки описания данных);
    - инструменты обработки данных, позволяющие заносить данные в БД, удалять, модифицировать или выбирать существенные данные (языки манипулирования данными);
    - инструменты конструирования экранных форм, которые предназначены для ввода данных, просмотра и обработки в диалоговом режиме (генераторы экранных форм);
    - инструменты создания запросов для поиска данных при заданных условиях (язык запросов);
    - инструменты создания отчетов, вывода на печать результатов обработки в удобном для пользователя виде (генераторы отчетных форм);
    - графические языки выборки данных по различным критериям;
    - инструменты создания пользовательских приложений, которые позволяют объединять разные операции работы с БД в единый технологический процесс (генераторы приложений);
    - инструменты, позволяющие специфицировать приложения различных типов (языки спецификаций – документаторы приложений). Большинство современных СУБД включают в себя несколько языковых средств.

Компоненты СУБД

СУБД – это сложный вид программного обеспечения, над созданием которого работают большие коллективы высококвалифицированных программистов. На современном рынке программного обеспечения конкурирует около трех десятков коммерческих СУБД. Среди малых систем, рассчитанных на одного пользователя, сегодня наибольшей популярностью пользуются Microsoft Access и SQLite. В перечень получивших широкое признание многопользовательских реляционных СУБД входят: Oracle, SQL Server компании Microsoft, InterBase, Firebird, MySQL, PostgreSQL, Informix и т. д. Несмотря на то что все перечисленные программные продукты предназначены для решения практически одних и тех же задач, входящие в перечисленные СУБД программные компоненты и взаимосвязи между ними далеко не идентичны. Поэтому любая попытка обобщить все известные технические решения построения СУБД и на основе этого обобщения построить структурную схему типичной СУБД, пусть даже высокой степени абстракции, несколько наивна. Тем не менее мы попытаемся это сделать.

Для того чтобы СУБД оказалась в состоянии предоставлять минимальный набор из рассмотренных выше восьми сервисов, она должна состоять из набора компонент, представленных на рис. 15.

РИСУНОК 15
ОБОБЩЕННАЯ СТРУКТУРА СУБД



Над верхним уровнем системы расположены потребители услуг СУБД:

- администратор БД;
- программисты;
- конечные пользователи.

Администратор БД отвечает за планирование и физическую реализацию проекта. Он создает основные объекты БД, определяет правила поддержания целостности и непротиворечивости данных, управляет политикой безопасности, анализирует процесс эксплуатации проекта, контролирует производительность системы — одним словом, поддерживает жизнедеятельность БД. В распоряжении администратора имеются разнообразные средства проектирования БД, эти средства могут входить в состав СУБД в виде дополнительных программных модулей, но могут быть поставлены и сторонними разработчиками.

Программисты совместно с администратором трудятся над физическим созданием проекта БД. Прикладных программистов в большей степени интересует не сама концепция проекта БД, а способы донесения этой концепции до конечного пользователя. Поэтому в сфере интересов программистов находится скорее разработка клиентских приложений и отчетов. Основным инструментом прикладного программиста выступают многочисленные среды проектирования, как правило, не ниже четвертого поколения.

Основным потребителем услуг СУБД выступает обычный пользователь: в его интересах создаются БД и прикладное программное обеспечение. Среди пользователей есть и неплохие специалисты, чьи советы способны помочь разработчикам БД, а на другом конце «спектра» находятся пользователи, с трудом обнаруживающие нужные кнопки на клавиатуре. Последние могут даже и не понимать, что они работают с БД, расположенной на другом компьютере.

Основным средством общения между пользователями и БД выступает структурированный язык запросов SQL. Поэтому средства проектирования БД, ПО и клиентские приложения БД отправляют в адрес СУБД инструкции на языке SQL. Эти команды поступают в процессор запросов, который преобразует их в набор низкоуровневых команд, понятных ядру СУБД.

SQL
язык структурированных запросов, основной задачей которого является предоставление простого способа считывания и записи информации в базу данных.

Функции языка SQL:

- организация данных – создание и изменение структуры баз данных;
- чтение данных;
- обработка данных – удаление, добавление и корректировка данных;
- управление доступом к данным – предоставление привилегий (ограничение возможностей) пользователю для чтения и изменения данных;
- совместное использование данных – координация общего пользования данных многими пользователями;
- целостность данных – защита данных от разрушения при сбое системы или других обстоятельствах.

Подходы к обработке данных

Известны два основных типа подходов к обработке данных:

- ACID или BASE;
- теорема CAP.

Они являются двумя полюсами спектра возможных промежуточных вариантов. Для того чтобы оценить степень соответствия распределенной системы подходу ACID или BASE, используется теорема CAP.

ПОДХОД ACID

Акроним ACID появился в начале 1980-х годов[24] 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.

- **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.

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 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.

[24] Саму концепцию сформулировал еще в 1970-х годах американский теоретик вычислительных систем Джим Грей (James Nicholas «Jim» Gray), а термин ACID впервые был введен в работе Haerder and Rueter (1983) // URL: https://sites.fas.harvard.edu/~cs265/papers/haerder-1983.pdf.

Сравнение подходов ACID и BASE представлено в табл. 2.

ТАБЛИЦА 2
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

ТЕОРЕМА CAP

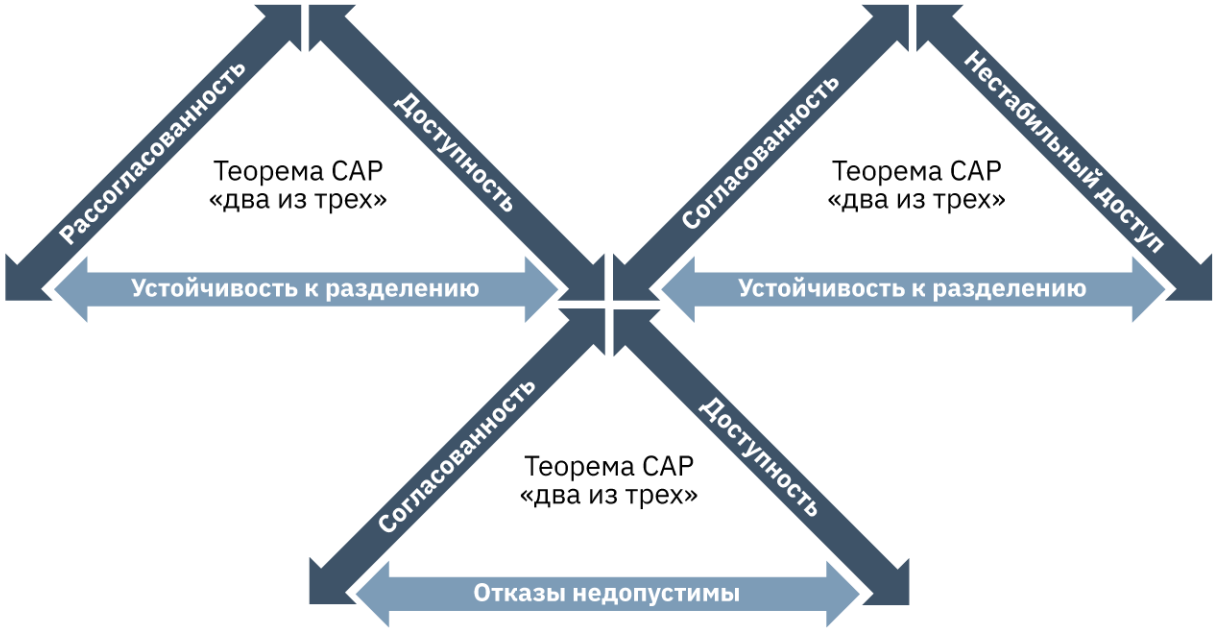
Теорема CAP (теорема Брюера[25]) стала ответом на наметившийся переход ко все более распределенным вычислительным системам. Положения теоремы говорят о том, что для распределенной системы не может быть обеспечено одновременное выполнение всех требований ACID в любой момент времени. Более того, чем крупнее система, тем хуже эти требования соблюдаются. Поэтому 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.

25 Эрик Брюер (англ. Eric Allen Brewer, p. 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.

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» (рис. 16).

РИСУНОК 16
ГРАФИЧЕСКОЕ ИЗОБРАЖЕНИЕ ТЕОРЕМЫ CAP



Жизненный цикл базы данных

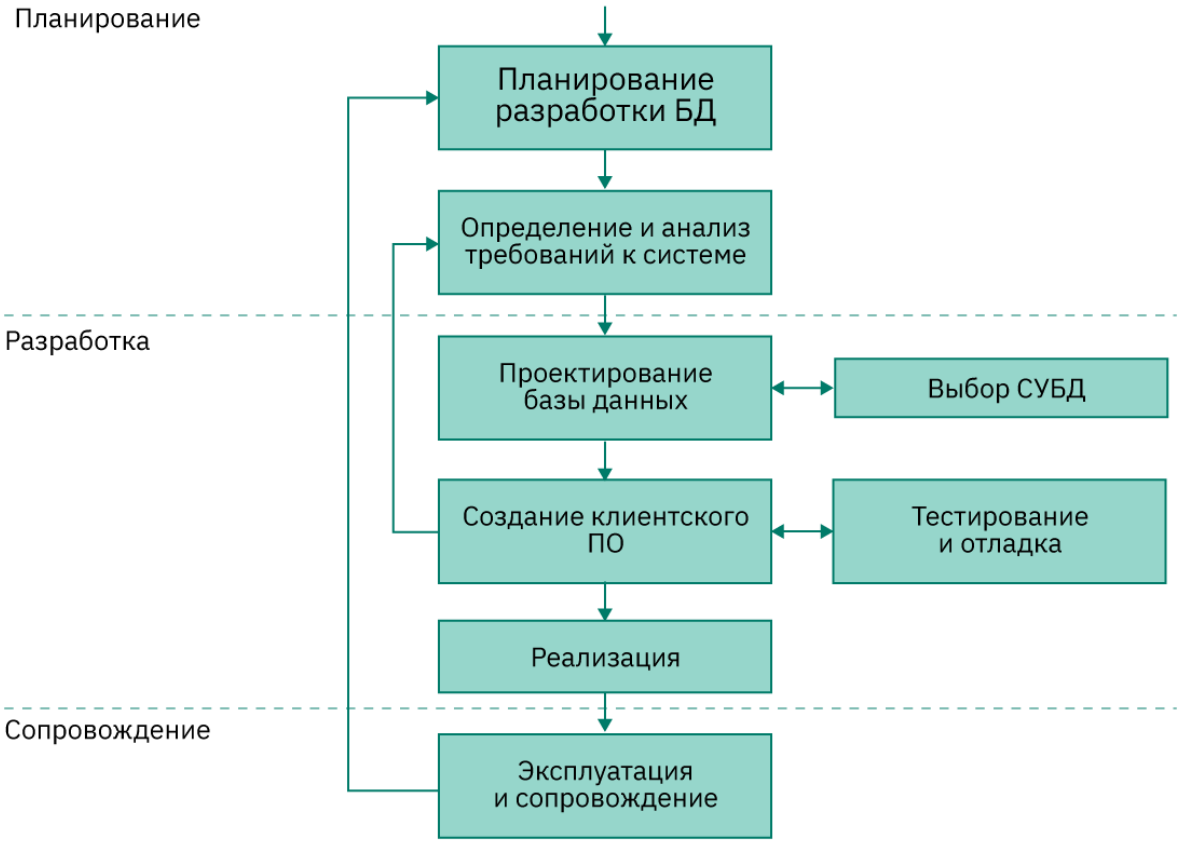
БД не только функционирует внутри информационной системы, но и развивается вместе с этой системой, поэтому в начале 1980-х годов был введен термин «жизненный цикл базы данных» (Database Live Cycle, DBLC).

Жизненный цикл БД
период от момента появления идеи создания БД до момента завершения его поддержки разработчиком или организацией, выполнявшей сопровождение.

Состав процессов жизненного цикла БД, как, впрочем, и всего остального программного обеспечения, регламентируется международным стандартом ISO/IEC/IEEE 12207:2017 Systems and software engineering — Software life cycle processes (Системная и программная инженерия – Процессы жизненного цикла)[26]. У нас с вами нет возможности подробно изложить концепцию этого стандарта, однако постараемся выделить важнейшие его элементы, относящиеся к жизненному циклу БД.

На рис. 17 представлены наиболее существенные этапы жизненного цикла БД, эта схема отчасти станет путеводителем по излагаемому в этой книге материалу. Вне всякого сомнения, специалист по проектированию БД способен расширить представленную схему, добавив в нее дополнительные элементы и связи, но вряд ли у него поднимется рука удалить хотя бы один из предложенных на рисунке блоков, ведь здесь собран тот золотой минимум, без которого нельзя обойтись.

РИСУНОК 17
ЖИЗНЕННЫЙ ЦИКЛ ПРОЕКТА БАЗЫ ДАННЫХ



26 URL: https://www.iso.org/ru/standard/63712.html

Процесс проектирования и реализации жизнеспособной БД содержит серию этапов, включающих планирование разработки, определение и анализ требований к системе, проектирование БД, выбор целевой СУБД, создание прикладного ПО, тестирование, реализацию, эксплуатацию и сопровождение. Перечисленные этапы обладают итеративными чертами и предусматривают возможность возврата к предыдущему или даже к начальному этапу проектирования в случае невозможности соблюсти пожелания заказчика к проекту в заданном объеме и с требуемым качеством. Примерные критерии для оценки СУБД представлены в табл. 3.

ТАБЛИЦА 3
ПРИМЕРЫ КРИТЕРИЕВ ДЛЯ ОЦЕНКИ СУБД

КРИТЕРИИ	ЭЛЕМЕНТЫ КРИТЕРИЕВ
Возможности по определению данных	– возможность определения доменов; – простота реструктуризации БД; – стандартные и расширенные типы данных; – механизм представлений; – производительные индексы;
Требования по определению данных	– аппаратные требования; – требования к операционной системе и другому ПО; – требования к сетевой инфраструктуре;
Защита данных	– авторизация; – аутентификация; – шифрование;
Доступность	– степень поддержки стандарта SQL; – многопользовательский доступ; – интеграция с ПО других производителей;
Особенности управления транзакциями	– число одновременных подключений; – модель обработки транзакций; – поддержка точек восстановления; – особенности ведения системного журнала;
Современные средства разработки	– удобный интерфейс разработчика; – модель обработки транзакций; – поддержка точек восстановления; – особенности ведения системного журнала;
Инструментарий администратора	– резервное копирование и восстановление; – контроль производительности; – ведение протокола работы СУБД; – импорт/экспорт данных;
Поддержка производителем	– возможность модернизации; – регулярный выпуск исправлений; – полнота документации;
Экономические параметры	– стоимость приобретения; – стоимость эксплуатации.

Состав и особенности функционирования любой сложной системы, в том числе и СУБД, определяются стоящими перед ней задачами. Еще на заре разработки реляционной модели доктор Эдгар Фрэнк Кодд сформулировал функциональные обязанности классической СУБД, чем определил вектор развития этой области информационных технологий. За более чем 40 лет существования реляционных БД идеи Кодда совершенствовались учеными, инженерами и программистами, в результате к сегодняшнему дню разработан широкий перечень эффективных программных и аппаратных решений, применяемых в СУБД.

К настоящему времени существует более трех сотен разнообразных СУБД, способных удовлетворить самого привередливого разработчика. Среди множества архитектурных решений доступа к БД на сегодняшний день доминирующее положение занимают клиент-серверные системы с централизованным хранением данных. Вместе с тем развитие сетевых технологий и возрастающие требования предприятий и организаций к быстрому и надежному доступу к данным привели к появлению распределенных БД, основанных на идее децентрализации.

Интеграция и интероперабельность данных

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 данных
процесс объединения данных в согласованные физические или виртуальные формы.

Интероперабельность данных
способность двух информационных систем или компонентов или более к обмену информацией и к использованию информации, полученной в результате обмена.

Интегрированная система
система, в которой все входящие в нее подсистемы работают по единому алгоритму, то есть имеет единую точку управления.

Интероперабельная система
система, в которой входящие в нее подсистемы работают по независимым алгоритмам, не имеют единой точки управления, все управление определяется единым набором стандартов — профилем интероперабельности[27]

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 ЦОД;
- 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.

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, в

27 Национальный стандарт РФ ГОСТ Р 55062-2012 «Информационные технологии. Системы промышленной автоматизации и их интеграция. Интероперабельность. Основные положения» // Электронный фонд правовых и нормативно-технических документов. URL: https://docs.cntd.ru/document/1200102958

дальнейшем консолидированные данные передаются в целевые системы, которые предоставляют их потребителям.

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 данные. Большие данные могут быть объектом интеллектуального анализа, материалом для построения предиктивных моделей и средством получения оперативной информации.

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 Business intelligence (BI), аналитики, управления основными данными и обеспечение операционной эффективности.

Внедряя решения в области DII, организация должна руководствоваться следующими принципами:

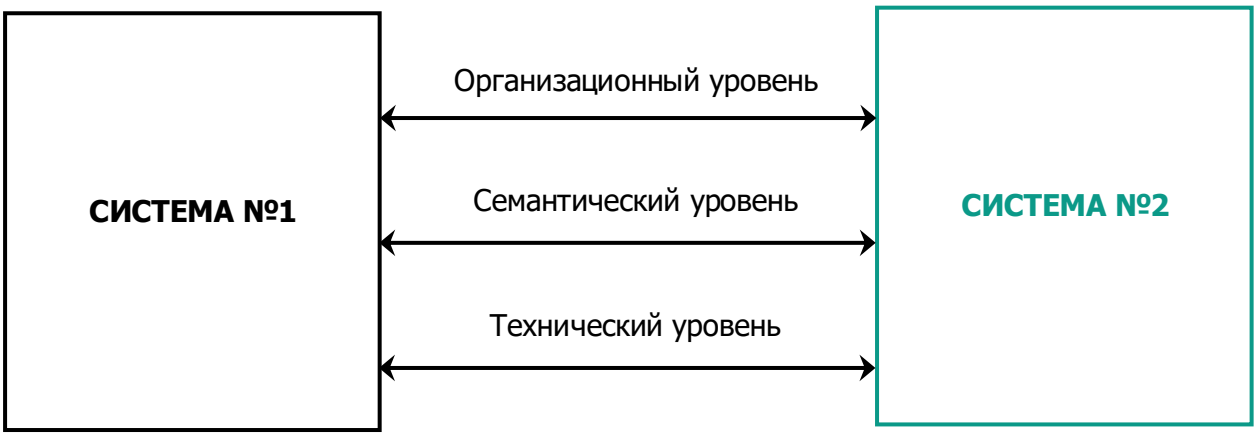
- при проектировании надлежит придерживаться корпоративного подхода, обеспечивающего возможность последующего расширения и масштабирования, но реализацию проводить итерационно, методом пошагового ввода новых решений в эксплуатацию;
- должны сбалансированно учитываться локальные и корпоративные потребности в данных, а также в поддержке и сопровождении;
- следует обеспечивать ответственность бизнеса за проектирование решений и проведение других работ в области DII. Экспертов со стороны бизнеса следует привлекать к разработке и модификации правил преобразования данных, как постоянно хранимых, так и виртуальных.

ЭТАЛОННАЯ МОДЕЛЬ ИНТЕРОПЕРАБЕЛЬНОСТИ

Эталонная модель интероперабельности представляет собой развитие семиуровневой базовой эталонной модели ВОС согласно ГОСТ Р ИСО/МЭК 7498-1-99 «Информационная технология. Взаимосвязь открытых систем. Базовая эталонная модель. Часть 1. Базовая модель» (рис. 18).

Для систем конкретных классов на базе эталонной модели интероперабельности должны создаваться проблемно ориентированные модели интероперабельности, которые могут иметь большее число уровней.

РИСУНОК 18
ЭТАЛОННАЯ МОДЕЛЬ ИНТЕРОПЕРАБЕЛЬНОСТИ



ТЕХНИЧЕСКИЙ УРОВЕНЬ

Технический уровень описывает синтаксис или форматы передаваемой информации, заостряя внимание на том, как представлена информация в коммуникационной среде. Технический уровень включает такие ключевые аспекты, как открытые интерфейсы, службы связи, интеграция данных и промежуточный слой программного обеспечения (Middleware), представление и обмен данными, службы доступности и защиты информации. Техническая интероперабельность достигается главным образом за счет использования стандартных протоколов связи типа TCP/IP.

СЕМАНТИЧЕСКИЙ УРОВЕНЬ

Данный уровень описывает семантические аспекты взаимодействия, то есть содержательную сторону обмениваемой информации. Семантическая интероперабельность позволяет системам комбинировать полученную информацию с другими информационными ресурсами и обрабатывать ее смысловое содержание. Семантическая интероперабельность достигается за счет применения стандартов типа XML.

ОРГАНИЗАЦИОННЫЙ УРОВЕНЬ

Организационный уровень акцентирует внимание на прагматических аспектах взаимодействия (деловых или политических). На этом уровне согласуются бизнес-цели и достигаются соглашения о сотрудничестве между административными органами, которые хотят обмениваться информацией, хотя имеют отличающиеся внутреннюю структуру и процессы. Организационная интероперабельность нацелена на то, чтобы удовлетворить требования сообщества пользователей: службы должны стать доступными, легко идентифицироваться и быть ориентированными на пользователя. Организационная интероперабельность достигается за счет применения не стандартов (нормативно-технических документов), а нормативно-правовых документов (соглашений, конвенций, договоров о сотрудничестве).

ОСНОВНЫЕ ЭТАПЫ ОБЕСПЕЧЕНИЯ ИНТЕГРАЦИИ И ИНТЕРОПЕРАБЕЛЬНОСТИ

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.

ЭТАП 1. ПЛАНИРОВАНИЕ И АНАЛИЗ

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 хранения данных (data retention) и других аспектов их жизненного цикла. Требования в отношении сроков хранения часто сильно зависят от предметной области и вида данных.

Обычно требования к интеграции и жизненному циклу данных определяют бизнес-аналитики, распорядители данных и архитекторы, которые работают в различных областях деятельности организации, включая ИТ, и 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 выполнить требования.

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.

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 определить их пригодность 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.

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 полагается вести 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.

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 допустимо 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 могут оказать существенную помощь в успешной реализации проекта и повышают способность организации эффективно использовать свои данные.

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 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 удастся выявить, 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 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.

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.

Требование профилирования данных не должно вступать в противоречие с установленными в организации правилами информационной безопасности и внешними требованиями по защите конфиденциальных данных.

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 в целях:

- 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

распорядителями 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.

ЭТАП 2. ПРОЕКТИРОВАНИЕ РЕШЕНИЙ ПО ИНТЕГРАЦИИ ДАННЫХ

2.1 Разработка архитектуры интеграционного решения

Решения по интеграции данных следует принимать как на корпоративном уровне, так и на уровне отдельных приложений. 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 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), 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.

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 и 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 можно производить по расписанию в пакетном режиме или запускать автоматически по мере регистрации определенных событий в режиме реального времени. Результаты преобразований можно либо физически сохранять в целевом формате, либо отображать в этом формате в виде виртуального представления данных.

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 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, касающихся события, например добавления или обновления данных. Такая оркестровка обычно сложнее, чем в пакетном режиме, и реализуется посредством применения многих инструментов. Она по своей природе не может быть линейной.

ЭТАП 3. РАЗРАБОТКА РЕШЕНИЙ ПО ИНТЕГРАЦИИ ДАННЫХ

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 единых в рамках всей организации взаимосогласующихся инструментов и стандартных пакетов во всех указанных решениях значительно упрощает техническую поддержку и снижает операционные расходы.

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 (enterprise service bus, 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, которые запускают 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.

3.3 Выработка подхода к миграции данных

При внедрении новых приложений, при выводе из эксплуатации или объединении существующих возникает необходимость в переносе данных. Этот процесс включает преобразование данных 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, не учитывают из-за того, что программистов просят просто перенести данные в новую среду; программистов не привлекают 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.

3.4 Выработка подхода к публикации

Системы, в которых создаются или ведутся критически важные данные, должны сделать их доступными другим системам организации. Публикацию новых или изменившихся данных, которые требуются другим системам (прежде всего, хабам данных или корпоративным шинам данных), следует осуществлять 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 уведомления потребителей (приложений и/или пользователей), подписавшихся на интересующие их данные, о внесенных изменениях.

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.

Подготовку и предварительную обработку исторических данных, используемых в предиктивной модели, можно производить либо в ночное время в пакетном режиме, либо в режиме, близком к реальному времени. Обычно некоторое информационное наполнение предиктивной модели возможно до начала обработки событий (например, на основании имеющихся сведений о том, какие продукты или услуги покупатели практически всегда заказывают 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).

3.6 Ведение метаданных, необходимых для обеспечения DII

Как отмечалось ранее, в процессе разработки решений по обеспечению интеграции и интероперабельности данных организации создают и выявляют полезные метаданные. Метаданные требуют тщательного учета и управления с целью обеспечить правильное понимание данных в информационных системах, 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 (структуру, 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 сервис-ориентированная архитектура (service-oriented architecture, SOA), который предоставляет контролируемый доступ к постоянно обновляющемуся каталогу имеющихся сервисов для работы с данными и приложениями.

ЭТАП 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.

Инструменты интеграции и интероперабельности данных

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 проектирование, так и непосредственное выполнение операций по преобразованию данных.

Существуют чрезвычайно сложные инструменты для разработки и выполнения ETL-операций (физически или виртуально) как в пакетном режиме, так и в режиме реального времени. 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.

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. Корпоративная шина данных (ESB)

Корпоративной шиной данных (enterprise service bus, 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, есть ли исходящие сообщения, 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.

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.

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.

6. Инструменты профилирования данных

Профилирование данных предполагает статистический анализ содержимого информационных массивов с целью получить более полное представление 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 определить пригодность 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.

7. Репозиторий метаданных

Репозиторий метаданных содержит информацию о данных организации с описанием их структуры, содержимого и бизнес-правил управления данными. В ходе интеграционного проекта (для 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 общий репозиторий для хранения метаданных. Для консолидации всех метаданных, создаваемых и поддерживаемых с помощью различных операционных инструментов, какой-либо из имеющихся репозиториев метаданных может быть выделен в качестве центрального.

Руководство интеграцией и интероперабельностью данных

Проектные решения по структуре сообщений, моделям данных и правилам их преобразования оказывают прямое влияние на способность организации использовать свои информационные массивы. 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 полагается фиксировать в качестве метаданных и консолидировать с целью провести сквозной анализ 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 учтен 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 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.

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, требования должны быть согласованы 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.

DLL И ПРОИСХОЖДЕНИЕ ДАННЫХ

При разработке DLL-решений (решений на основе динамически подключаемых библиотек - Dynamic Link Library, DLL) 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 гарантированно осуществлялось документирование информации о первоисточниках и перемещении данных, требуется контроль со стороны 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. Возможность 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.

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.

4.2.2. Безопасность данных

Ключевые понятия

Исторически развитие систем безопасности БД происходило как реакция на действия злоумышленников в соответствии с этапами эволюции самих хранилищ (БД) и изменениями типа и вида возрастающих угроз.

Угроза (безопасности информации)
совокупность условий и факторов, создающих потенциальную или реально существующую опасность нарушения безопасности информации[28].

Эти изменения были обусловлены общим развитием БД от решений на мейнфреймах до облачных хранилищ. Однако цифровизация значительно повысила уровень угроз:

- Массовая доступность технологий взлома и доступов к корпоративным системам: 58% предложений в даркнете по продаже доступов приходится на компании по промышленности, финансов, образовании и ИТ.
- Смешение ИТ и ОТ контуров создает новые способы атак через уязвимости операционных систем: 25% кибергруппировок ищут способы обойти защитный компонент UAC в Windows.
- Увеличение размера площади атаки: 85% кибергруппировок рассылают фишинговые письма, получают доступ в корпоративную сеть, а после — в технологическую.
- Развитие 5G, Интернета вещей и облачных сервисов повышает уровень угроз: >50% промышленных компьютеров подключаются к интернету регулярно. Самые влиятельные источники заражения промышленных систем по-прежнему: через USB-порт, почтовые клиенты, веб-сайты и приложение.
- Целевые атаки сложнее обнаружить, чем вредоносное ПО: в среднем разведка в Сети у киберпреступников занимает 3–6 месяцев.

В национальном стандарте РФ ГОСТ Р 50922-2006 безопасность информации (данных) определяется как состояние защищенности информации (данных), при котором обеспечены ее (их) конфиденциальность, доступность и целостность[29]. Есть и более широкая трактовка понятия с позиции процессов, которые необходимо организовать для обеспечения безопасности данных.

Безопасность данных
планирование, разработка и осуществление политик и процедур, обеспечивающих надлежащую аутентификацию, авторизацию и доступ пользователей, а также аудит данных и информационных ресурсов организации[30].

28 Национальный стандарт РФ ГОСТ Р 50922–2006 «Защита информации. Основные термины и определения» // Электронный фонд правовых и нормативно-технических документов. URL: http://docs.cntd.ru/document/1200058320
29 Национальный стандарт РФ ГОСТ Р 50922-2006 «Защита информации. Основные термины и определения» // Электронный фонд правовых и нормативно-технических документов. URL: http://docs.cntd.ru/document/1200058320
30 DAMA-DMBOK: свод знаний по управлению данными / Dama International; пер. с англ. Г. Агафонова. 2-е изд. М: Олимп-Бизнес, 2020.

Ключевые цели:

- 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 заинтересованных сторон.

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.

Факторы обеспечения информационной безопасности

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.** Регулированию подлежат различные аспекты 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 и прозрачность.
- **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.
- **Потребности в санкционированном доступе к данным.** Защита данных не должна ограничивать законный доступ к данным тех лиц, которые имеют на это право согласно действующему законодательству, равно как и санкционированный доступ к ним для 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 (рис. 19).

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 исключительно на доверии, ключевой составляющей которого является ответственное обращение с данными.

РИСУНОК 19
ИСТОЧНИКИ ТРЕБОВАНИЙ ПО ЗАЩИТЕ ДАННЫХ

Политика безопасности: актуальная документация

На сегодня политика безопасности ключевых информационных систем как никогда требует усиленного внимания и постоянной актуализации требований. Экспертно-аналитический центр ГК InfoWatch[31] подготовил обновленный перечень документации, необходимой для приведения информационных систем в соответствие требованиям, которые диктует политика безопасности объектов критической информационной инфраструктуры. В документе систематизированы нормативно-правовые акты, организационно-методические документы для приведения информационных систем в соответствие требованиям на июнь 2021 года. В частности, по сравнению с предыдущей версией, опубликованной в июне 2020 г., в документ добавлена информация о Федеральном законе от 26.05.2021 №141-ФЗ «О внесении изменений в Кодекс Российской Федерации об административных правонарушениях».

Документ «Нормативно-правовые акты, организационные и методические документы по обеспечению безопасности критической информационной инфраструктуры Российской Федерации»[32] включает актуальные данные о федеральных законах, указах Президента РФ, постановлениях Правительства РФ, приказах и сообщениях ФСТЭК России, приказах ФСБ, Минцифры в сфере БД.

31 InfoWatch – российский разработчик решений для обеспечения информационной безопасности организаций: URL: https://www.infowatch.ru/

32 Нормативно-правовые акты, организационные и методические документы по обеспечению безопасности критической информационной инфраструктуры Российской Федерации. 2021. URL: https://www.infowatch.ru/analytics/digest/politika-bezopasnosti-kii-aktualnaya-dokumentatsiya

Риски информационной безопасности

Риски информационной безопасности
это потенциальные возможности использования уязвимостей актива или группы активов конкретной угрозой для причинения ущерба организации[33].

Другими словами, риски информационной безопасности — это вероятные потери организации в результате инцидентов.

Риски информационной безопасности связаны с нарушениями конфиденциальности, целостности и доступности информационных активов компании. Они являются, как правило, результатом промышленного шпионажа, саботажа, кибератак на информационные ресурсы, а также целевых атак на компьютеры организации.

Выделяют несколько видов рисков информационной безопасности:

- стратегические риски:
  - более быстрый выход конкурентов на рынок быстрее;
  - сокращение рыночной доли компании, отток клиентов к конкурентам;
  - потеря бизнеса;
  - нарушение непрерывности производственных и бизнес-процессов;
  - финансовые потери (недополученная прибыль, расходы на нейтрализацию последствий и т.д.);
- репутационные риски:
  - распространение негатива как внутри компании, так и за ее пределами
  - информационные вбросы негативного или компрометирующего характера (черный PR);
  - клевета в отношении компании и должностных лиц;
  - разглашение секретной информации в социальных сетях, интернет-медиа и т.п.;
  - потеря доверия к веб-ресурсам компании (вследствие DDoS-атак);
- операционные риски:
  - «слив» конфиденциальной информации нелояльными сотрудниками;
  - проигранные конкурсы и тендеры;
  - мошенничество с использованием корпоративных ресурсов компании;
  - халатность персонала при работе с конфиденциальной информацией;
  - переход сотрудников к конкурентам с наработками и базами данных;
  - отток клиентов на ресурсы конкурирующих компаний;
  - недоступность веб-ресурсов компании (нарушение доступности);
- юридические риски:
  - несоответствие требованиям регуляторов, влекущее за собой санкции;
  - утечки персональных и конфиденциальных данных, приводящие к судебным искам;
  - использование сотрудниками корпоративных ресурсов в противозаконных целях;
  - промышленный шпионаж и следующие за ним длительные и дорогостоящие судебные тяжбы;
  - невозможность доказать в судебном порядке факт совершения сотрудником преступления.

33 ГОСТ Р ИСО/МЭК 27005-2010 «Информационная технология. Методы и средства обеспечения информации. Менеджмент риска информационной безопасности» // URL: https://docs.cntd.ru/document/1200084141

Для оценки рисков необходимо определить, какие бизнес-процессы в организации наиболее критичны, а какие — менее (так называемый Business Impact Analysis). Какие угрозы и уязвимости могут повлиять на непрерывность бизнес-процессов и как минимизировать вероятность этих угроз?

В рамках комплекса мер в области управления рисками информационной безопасности необходимо обеспечить:

- оценку рисков утечки конфиденциальной информации;
- разработку модели угроз утечки информационных активов;
- разработку процедур реагирования на инциденты;
- подготовку рекомендаций по развитию процесса защиты от утечек;
- внедрение технических средств защиты.

Утечка данных в России: факты

Утечка данных — это массовое явление по всему миру, от него не застрахован ни один оператор данных: ни государственные органы и их информационные системы, ни коммерческие организации. Прямой целенаправленный взлом происходит редко. Причинами утечек, как правило, являются:

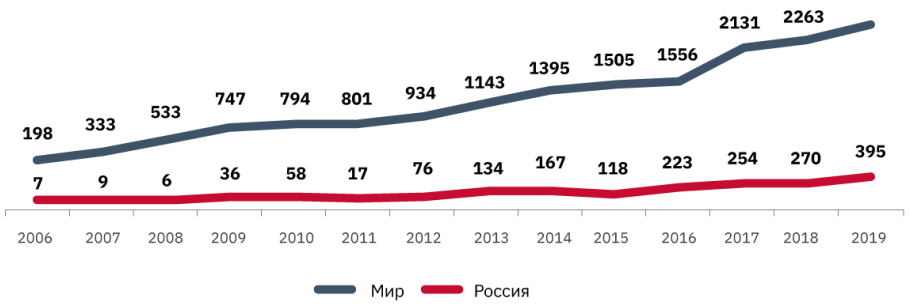
- ошибки в законодательстве;
- недостаточно продуманная работа регулирующих и контролирующих органов;
- ошибки непосредственных разработчиков баз данных (например, неправильно настроенные серверы, из-за чего сторонние специалисты по анализу могут получить данные, доступ к которым должен быть закрыт для них по умолчанию);
- действия недобросовестных сотрудников, которые копируют данные для последующей продажи[34].

Исследование утечек информации ограниченного доступа, зафиксированных в российских (работающих в России) коммерческих и некоммерческих компаниях, государственных органах и организациях в 2019 году экспертно-аналитическим центром ГК InfoWatch, показало, что:

- зафиксировано 395 случаев утечки данных, что составляет 15,7% от числа утечек данных по всему миру;
- в результате утечек оказались скомпрометированы более 172 млн записей персональных данных и платежной информации;
- по сравнению с данными 2018 года число утечек увеличилось на 46%, а объем скомпрометированной пользовательской информации вырос более чем в 6 раз;
- по числу утечек Россия седьмой год подряд занимает второе место в мировом распределении (после США). И чаще всего в России «утекают» персональные данные и платежная информация, на эти типы данных приходится 87,3% утечек, случившихся в 2019 году (рисунок №);
- в 72,1% случаев виновными в утечке информации оказались рядовые сотрудники компаний, в 4,6% случаев — топ-менеджмент организаций, в 18,4% — хакеры и неизвестные лица;
- доля утечек, случившихся под воздействием внутреннего нарушителя, составила 88,2%, по вине или неосторожности внутренних нарушителей было скомпрометировано 117 млн записей персональных данных и платежной информации — 68,1% от совокупного количества записей, скомпрометированных в 2019 году[35].

РИСУНОК 20
ЧИСЛО ЗАФИКСИРОВАННХ УТЕЧЕК, РОССИЯ – МИР, 2006-2019 гг.

34 Этика и «цифра»: Этические проблемы цифровых технологий: в 2 т. М.: РАНХиГС, 2020. Т. 1. URL: https://ethics.cdto.center/2_1/#2.1.2
35 Утечки данных. Россия. 2019 год // Infowatch. URL: https://www.infowatch.ru/analytics/reports/27614



Защита конфиденциальной информации

КАКИЕ ДАННЫЕ НЕОБХОДИМО ЗАЩИЩАТЬ?

Коммерческая тайна, производственные секреты, ноу-хау, базы данных клиентов, персональные данные — все это критически важная для бизнеса информация, доступ и использование которой должны быть тщательно регламентированы.

Нарушение конфиденциальности информации может быть связано как с мошеннической деятельностью сотрудников компании, так и с действиями вредоносного программного обеспечения или внешних злоумышленников. Независимо от этого, задачу защиты конфиденциальных данных следует решать комплексно: от разработки документации, процедур реагирования и оценки рисков до внедрения технических средств защиты.

Для защиты конфиденциальной информации необходимо обеспечить устранение бизнес-рисков:

- формирование полного перечня информации ограниченного доступа;
- введение режима коммерческой тайны, который полностью отвечает соответствующим нормативно-правовым актам РФ, принятие юридически грамотно составленного «Положения о коммерческой тайне» и сопутствующих документов, которые позволят минимизировать правовые риски компании в случае утечки коммерческой тайны и обеспечить возможность привлечения к ответственности нарушителей;
- контроль за движением конфиденциальной информации и ее защиту:
- защита от утечки коммерческой тайны;
- защита интеллектуальной собственности;
- предотвращение утечек персональных данных и баз данных;
- контроль доступа сотрудников к конфиденциальной информации;
- выявление злоумышленников, лиц, занимающихся промышленным шпионажем, а также привлечение их к ответственности за нарушения.

Для защиты конфиденциальной информации показаны следующие меры:

- категоризация информационных активов и создание перечня информации ограниченного доступа;
- оценка критичности информационных активов;
- создание модели угроз и оценка рисков утечек конфиденциальной информации;
- разработка процедур реагирования на инциденты;
- подготовка рекомендаций по внедрению технических средств защиты конфиденциальных данных от утечек.

Такой подход обеспечивает точное выявление конфиденциальных данных в большом объеме передаваемой информации, а также позволяет настроить технические средства, в частности DLP-систему, в соответствии с регламентами, принятыми в компании,

предотвратить утечку путем блокирования процесса передачи в случае нарушения политики безопасности.

СОВРЕМЕННЫЕ ПРОБЛЕМЫ ОБЕСПЕЧЕНИЯ БЕЗОПАСНОСТИ БАЗ ДАННЫХ[36]

Список основных уязвимостей хранилищ данных, актуальный на сегодняшний день, не претерпел существенных изменений за последние более чем пять лет. Проанализировав средства обеспечения безопасности СУБД, архитектуру БД, интерфейсы, известные уязвимости и инциденты безопасности, можно выделить следующие причины возникновения такой ситуации:

- проблемами безопасности серьезно занимаются только крупные производители, причем прежде всего относительно ведущих продуктов линеек для хранения данных;
- программисты БД, прикладные программисты и администраторы БД не уделяют должного внимания вопросам безопасности;
- разные масштабы и виды хранимых данных требуют разных подходов к безопасности;
- различные СУБД используют разные языковые диалекты для доступа к данным, организованным на основе одной и той же модели;
- появляются новые виды и модели хранения данных.

Многие уязвимости остаются актуальными из-за невнимания или незнания администраторами систем БД вопросов безопасности. Например, известные техники простой SQL-инъекции широко эксплуатируются сегодня в отношении различных веб- и иных приложений, в которых не уделяется внимание контролю входных данных запроса. Причинами этого являются как недостаточная информированность или внимание администраторов СУБД и прикладных программистов, так и отсутствие встроенных средств контроля известных уязвимостей в большинстве СУБД. Хорошим решением были бы автоматизация и перенос контроля таких угроз на уровень сервера, однако многообразие языковых диалектов не позволяет это сделать.

Также нужно отметить, что применение различных средств обеспечения информационной безопасности является для организации компромиссом в финансовом плане: внедрение более защищенных продуктов и подбор более квалифицированного персонала требуют б*о*льших затрат. К тому же компоненты безопасности могут отрицательно влиять на производительность систем управления БД. Полное соответствие модели ACID – самый медленный способ обеспечения целостности при многопользовательской работе.

Еще одна причина такой ситуации – разрозненность диалектов языка запросов к СУБД. Если рассматривать только известные реляционные СУБД, несмотря на наличие развивающегося стандарта SQL (SQL-92, SQL-99, SQL-2003, SQL-2008), даже крупные производители не только используют собственные расширения языка, но и не поддерживают до конца операции принятой версии стандарта. Этот факт осложняет разработку единых механизмов защиты БД на уровне сервера.

Приведенные выше проблемы усугубляются с появлением и широким распространением нереляционных хранилищ данных и СУБД, использующих другую модель данных, однако построенных по тем же принципам и имеющих аналогичное назначение, что и традиционные, реляционные серверы. По сути, многообразие современных NoSQL (нереляционных) решений приводит к разнообразию применяемых моделей данных и размывает границу понятия БД.

---

36 Полтавцева М. А., Хабаров А. Р. Безопасность баз данных: проблемы и перспективы // Программные продукты и системы. 2016. №3 (115). URL: https://cyberleninka.ru/article/n/bezopasnost-baz-dannyh-problemy-i-perspektivy

Следствием этих проблем и отсутствия единых методик является нынешняя ситуация с безопасностью NoSQL-систем. Эти решения появились на рынке недавно и еще не успели пройти «путь ошибок и уязвимостей», характерный для их более зрелых реляционных аналогов. В большинстве NoSQL-систем отсутствуют не только общепринятые механизмы безопасности вроде шифрования, поддержки целостности и аудита данных, но даже развитые средства аутентификации пользователей.

ОСОБЕННОСТИ СИСТЕМ БАЗ ДАННЫХ КАК ОБЪЕКТА ЗАЩИТЫ

В связи с появлением новых решений в области нереляционных хранилищ, размывающих границу традиционного представления о СУБД (например, система кэширования данных в памяти MemcasheDB или Hadoop HDFS), определим функции, отличающие СУБД от файлового хранилища и других типов программных продуктов.

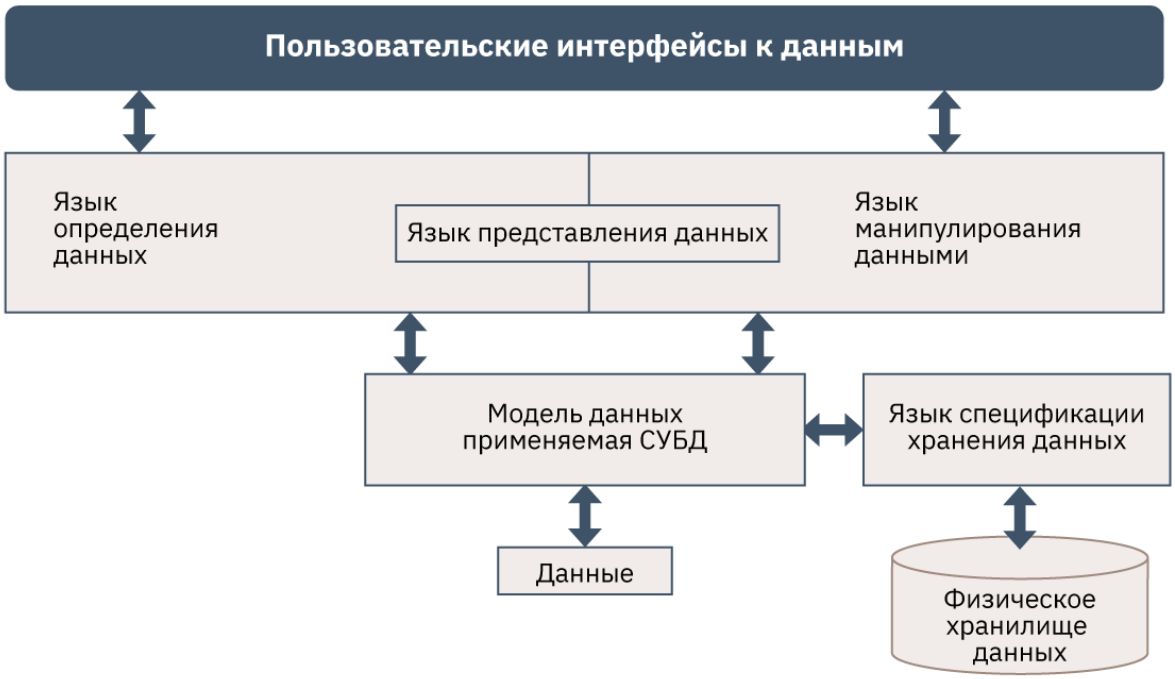
От остальных видов прикладного ПО (относительно информационной безопасности и не только) системы БД отличаются своей двойственной природой. С этой точки зрения СУБД включает в себя два компонента: хранимые данные (собственно БД) и программы управления (СУБД).

Обеспечение безопасности хранимой информации, в частности, невозможно без обеспечения безопасного управления данными. Исходя из этой концепции, все уязвимости и вопросы безопасности СУБД можно разделить на две категории: зависимые от данных и независимые от данных.

Отметим, что уязвимости, независимые от данных (их структуры, организации и т.д.), являются характерными для всех прочих видов ПО. К этой группе можно отнести несвоевременное обновление ПО или наличие неиспользуемых функций.

Зависимыми от данных (в той или иной степени) являются многие аспекты безопасности. В частности, зависимыми напрямую можно назвать механизмы логического вывода и агрегирования данных, называемые специфичными проблемами СУБД. В то же время многие уязвимости являются косвенно зависимыми от данных. Например, современные СУБД (и реляционные, и нереляционные решения) поддерживают запросы к данным с использованием некоторого языка запросов. В свою очередь, в этом качестве используются специализированные языки запросов (SQL, CQL, OQL и др.), наборы доступных пользователю функций (которые, в свою очередь, тоже можно считать операторами запросного языка) или произвольные функции на языке программирования (чаще всего Java). Обобщенные интерфейсы работы с данными представлены на рис. 21.

РИСУНОК 21
ВЛИЯНИЕ МОДЕЛИ ДАННЫХ СУБД НА ЯЗЫК ЗАПРОСОВ



Архитектура применяемых языков, по крайней мере то, что касается специализированных языков (запросов) и наборов функций, напрямую связана с моделью данных, применяемой для хранения информации. Таким образом, модель определяет особенности языка, а особенности языка — наличие в нем тех или иных уязвимостей. Причем уязвимости общего типа, например инъекция (под инъекцией будем понимать атаку на БД путем модификации входных запросов, заставляющую сервер СУБД выполнить нелегитимный набор действий), выполняются по-разному (SQL-инъекция, JAVA-инъекция), в зависимости от синтаксиса и семантики языка, которые, как уже сказано выше, отчасти определяются моделью данных и, следовательно, являются зависимым от данных компонентом.

ТРЕБОВАНИЯ К БЕЗОПАСНОСТИ БАЗ ДАННЫХ

Таким образом, на основании разделения уязвимостей можно выделить зависимые и независимые от данных меры обеспечения безопасности хранилищ информации. Независимыми от данных можно назвать следующие требования к безопасной системе БД.

- **Функционирование в доверенной среде.** Под доверенной понимается информационная среда, интегрирующая совокупность защитных механизмов, которые обеспечивают обработку информации без нарушения политики безопасности. В данном случае СУБД должна функционировать в доверенной информационной системе с соответствующими методами обмена данными.
- **Организация физической безопасности файлов данных.** Данный вопрос требует более детального изучения, так как применяемые структуры данных в различных моделях данных СУБД могут иметь значение при шифровании и защите файлов данных. Однако в первом приближении вопрос физической безопасности файлов данных сходен с вопросом физической безопасности любых других файлов пользователей и приложений.
- **Организация безопасной и актуальной настройки СУБД.** Имеются в виду такие общие вопросы обеспечения безопасности, как своевременная установка обновлений, отключение неиспользуемых модулей или применение эффективной политики паролей.

Следующие требования можно назвать зависимыми от данных:

- **Безопасность пользовательского слоя ПО.** С этой целью ставятся задачи построения безопасных интерфейсов и вызовов (в том числе с учетом интерфейса СУБД и механизма доступа к данным).
- **Безопасная организация данных и манипулирование ими.** Вопрос организации данных и управления ими является ключевым в системах хранения информации. Несмотря на то что в приведенном перечне он указан последним, именно в эту область входят задачи организации данных с контролем целостности, обеспечение защиты от логического вывода и другие, специфичные для СУБД проблемы безопасности. Фактически эта задача включает в себя основной пул зависимых от данных уязвимостей и защиты от них.

ПУТИ СОЗДАНИЯ ЗАЩИЩЕННЫХ БАЗ ДАННЫХ

Для преодоления названных проблем обеспечения информационной безопасности СУБД необходимо перейти от метода закрытия уязвимостей к комплексному подходу обеспечения безопасности хранилищ информации. Основными этапами этого перехода, по мнению авторов, должны стать следующие положения.

1. Разработка комплексных методик обеспечения безопасности хранилищ данных на текущем этапе.

Создание комплексных методик позволит применять их (или их соответствующие версии) при разработке хранилищ данных и пользовательского ПО. Следование комплексной методике позволит избежать многих ошибок управления СУБД и защититься от наиболее распространенных на сегодняшний день уязвимостей.

2. Оценка и классификация угроз и уязвимостей СУБД.

Специализированная классификация угроз и уязвимостей СУБД позволит упорядочить их для последующего анализа и защиты, даст возможность установить зависимость между уязвимостями и причинами (источниками) их возникновения. В результате при введении конкретного механизма в СУБД появится возможность установить и спрогнозировать связанные с ним угрозы и заранее подготовить соответствующие средства обеспечения безопасности.

3. Разработка стандартных (применимых к различным СУБД без внесения изменений или с минимальными изменениями) механизмов обеспечения безопасности.

Стандартизация подходов и языков работы с данными позволит создать мультиплатформенные средства обеспечения безопасности, применимые к разным СУБД. С одной стороны, это методические и теоретические подходы, применимые в рамках модели данных. На сегодняшний день есть разработки таких механизмов по реляционной модели, однако они не решают всех насущных вопросов безопасности. С другой стороны, это разработка теоретического базиса для новых СУБД, в частности конкретизация и формализация агрегатных моделей данных. Появление готовых программных средств во многом зависит от производителей и разработчиков СУБД и их следования стандартам, а также достаточности определенных в стандарте средств для построения развитых механизмов безопасности.

4. Разработка теоретической базы информационной защиты систем хранения и манипулирования данными.

Выше отмечен ряд характерных особенностей, специфичных для хранилищ данных: двойственная природа СУБД, зависимость уязвимостей и механизмов управления от данных и описывающей их организацию модели, угрозы логического вывода, различная значимость сочетаний данных. Все это определяет специфический характер безопасности СУБД и требует новых теоретических подходов к обеспечению защиты данных в программных системах такого рода. Отдельные большие вопросы – развитие теоретического базиса в контексте формализации модели данных, разработка подходов обеспечения целостности информации для новых NoSQL-хранилищ.

На основании изложенного сделаем следующие выводы. В результате рассмотрения российских и зарубежных работ ситуации на рынке СУБД выделены существующий подход к обеспечению безопасности, принцип и основные эволюционные этапы систем БД. Сформулированы проблемы информационной безопасности современных СУБД: разнообразие языковых средств, появление новых моделей данных, не подкрепленных теоретическим базисом, необходимость поиска баланса между безопасностью и ее стоимостью, развитие систем защиты как реакции на потерю средств и престижа, а также общее невнимание к вопросам безопасности.

Сформулированы критерии, выделяющие СУБД среди сходных программных продуктов, с учетом новых кластерных и in memory решений, особенности этого класса ПО с точки зрения информационной безопасности, и предложено базовое деление уязвимостей на зависимые и независимые от данных и их организации.

В результате сформулированы общие требования к безопасности БД, перспективные пути исследования и развития систем защиты для построения надежных и защищенных серверов по обработке информации. В них вошли как систематизация и развитие существующих подходов в виде выработки методик и стандартизации механизмов защиты, так и направления новых исследований, например классификация уязвимостей СУБД и формализация новых моделей данных, построение прогнозирующих средств защиты[37].

37 Полтавцева М. А., Хабаров А. Р. Безопасность баз данных: проблемы и перспективы // Программные продукты и системы. 2016. №3 (115). URL: https://cyberleninka.ru/article/n/bezopasnost-baz-dannyh-problemy-i-perspektivy.

4.2.3. Этика обращения с данными

Цели организации по этике данных

Компании обеспечивают защиту данных, руководствуясь правовыми актами и требованиями регулирующих органов. Поскольку данные касаются людей — клиентов, сотрудников, пациентов, поставщиков, при работе с данными нужно отдавать себе отчет, что имеются и этические причины не допускать нарушений. Данные, на первый взгляд не отражающие никаких сведений о физических лицах, все равно могут быть использованы для принятия решений и в конечном счете такие решения навредят людям.

Этика обращения с данными
комплекс мер по обеспечению соответствия практик получения, хранения, управления, интерпретации, анализа, применения и ликвидации данных общечеловеческим этическим принципам, включая ответственность перед обществом.

К целям организации по этике данных можно отнести:

- определение и понятие принципов этичного обращения с данными в рамках организации;
- разъяснение сотрудникам рисков, обусловленных ненадлежащим обращением с данными;
- привитие желаемой культуры и моделей поведения в сфере обращения с данными;
- мониторинг нормативно-правовой среды: отслеживание, оценка и корректировка организационных подходов к обеспечению этичного обращения с данными.

Работу с данными следует выполнять с минимально возможным риском искажения, неверного представления, нецелевого использования или ошибочной интерпретации данных на протяжении всего жизненного цикла данных.

Бизнес-драйверы развития этичной работы с данными

Соблюдение норм этики при работе с данными повышает доверие к организации, к ее данным и к результатам деятельности. Это способствует улучшению отношений между организацией и третьими лицами.

В DAMA DMBOK2 достаточно подробно рассматриваются различные подходы к соблюдению этических принципов обращения с данными и регламенты Организации экономического сотрудничества и развития (ОЭСР), ЕС, США, Канады.

Законодательное регулирование этических вопросов

Глобальная тенденция такова, что повсеместно усиливается законодательное регулирование защиты персональных данных, а за основу модели в целом берутся стандарты ЕС, преимущественно «Общий регламент по защите данных» ЕС (General Data Protection Regulation, GDPR). Федеральный закон РФ от 27.07.2006 № 152-ФЗ «О персональных данных» в целом основан на тех же принципах защиты данных, унаследованных от ОЭСР, что и европейский регламент GDPR.

Принципы защиты данных в рамках этических вопросов

Принципы «Общего регламента по защите данных» ЕС:

1. Честность, законность, прозрачность

Персональные данные должны обрабатываться согласно букве закона, честно и прозрачно в отношении субъектов данных.

2. 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.

3. Минимизация объема данных

Персональные данные должны быть адекватными и релевантными, а объем собираемых данных не должен превышать минимума, необходимого для их целевого использования.

4. Точность данных

Персональные данные должны быть точными и при необходимости обновляться. Должны предприниматься все разумные меры по оперативному выявлению и незамедлительному удалению неточных данных.

5. Ограничение сроков хранения

Персональные данные должны храниться в форме, допускающей идентификацию субъектов данных (физических лиц), не дольше, чем это необходимо для их обработки.

6. 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.

7. 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.

8. Проектируемая конфиденциальность (privacy by design)

Концепция интеграции мер защиты конфиденциальных данных готовится на этапе проектирования системы их обработки. При этом защита персональных данных рассматривается как одна из основных функций системы.

Примеры неэтичного обращения с данными и риски

Большинство людей, работающих с данными, понимают, что данные вполне можно использовать для формирования неверных представлений.

Манипуляции с хронологией временем. Пример: на фондовых рынках публикуют котировки акций «по состоянию на момент закрытия биржевых торгов», когда они, как правило, дорожают по сравнению с усредненными котировками за прошедший день

Нечеткие определения или некорректные сравнения. Слепое прочесывание данных (Data mining snooping) – поиск выборки данных, дающая формально «статистические значимые» результаты корреляций, которые в реальности не выходят за пределы статистической ошибки и являются чисто случайными.

Вводящие в заблуждения визуальные представления. Пример: играя масштабом шкалы, можно представить тенденцию более выигрышной или неприглядной, чем она есть на самом деле.

Предвзятость, систематические ошибки и искажения. Пример: Сбор данных для подтверждения предопределенного результата. Вместо объективного сбора и анализа данных происходит подгонка выборки исходных данных под заказанный результат.

Этические трудности могут возникнуть в процессе интеграции и преобразования данных, и связаны с фундаментальными процессами управления данными, включая:

- ограниченность знаний о первоисточнике и происхождении данных;
- некачественные данные;
- ненадежные метаданные;
- отсутствие документированной истории исправления данных.

Формирование культуры этичного обращения с данными

Требования этичного обращения с данными, безусловно, включают соблюдение норм действующего законодательства, кроме того, они должны распространяться и на процедуры сбора, анализа и интерпретации данных, и на вопросы, связанные с повышением эффективности их использования.

Шаги по формированию культуры этичного обращения данными:

1. Обзорный анализ текущего состояния практик обращения с данными.
2. Выявление и определение принципов, практик и факторов риска, механизмов контроля соблюдения установленных практик.
3. Разработка стратегии обеспечения этичного обращения с данными и дорожной карты ее реализации. Стратегия должна включать:
   - положение о ценностях;
   - принципы этичного обращения с данными;
   - рамочную структуру обеспечения соответствия;
   - оценки рисков;
   - обучение и коммуникации: сотрудники должны дать расписку в том, что они изучили этический кодекс и знают о последствиях неэтичного обращения с данными, обучение должно быть непрерывным;
   - дорожную карту;
   - аудит и мониторинг.
4. Непрерывное обучение и коммуникация в области этичного обращения с данными:
   - Методы:
     - чек-листы по плану коммуникаций;
     - ежегодные подтверждения корпоративной декларации этических принципов.
   - Инструменты:
     - вики-сайты, базы знаний, корпоративная сеть;
     - микроблоги и другие средства внутренней коммуникации.
   - Метрики:
     - количество сотрудников, прошедших обучение;
     - количество инцидентов с соблюдением/несоблюдением сроков и качества;
     - вовлеченность корпоративного руководства.

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 компании. Совместно они отвечают за стандарты и политики обращения с данными, за механизмы надзора за их соблюдением на практике.

15 января 2020 года на Гайдаровском форуме был анонсирован аналитический доклад Центра подготовки руководителей и команд цифровой трансформации ВШГУ РАНХиГС «Этика и "цифра": этические проблемы цифровых технологий»[38], в котором этические проблемы рассматриваются преимущественно с точки зрения их применения в государственном управлении. Особое внимание обращено на три ключевые технологии: сбор и обработку цифровых данных, искусственный интеллект, интернет вещей. Их масштабное применение, с одной стороны, необходимо для повышения эффективности госуправления и развития экономики в целом, а с другой — способно вызвать целый ряд

38 Этика и «цифра»: Этические проблемы цифровых технологий: в 2 т. М.: РАНХиГС, 2020. Т. 1. URL: https://ethics.cdto.center/

этических коллизий, включая нарушение приватности граждан, дискриминацию отдельных категорий людей, цифровое неравенство и т. д.

Как потребители государственные органы и службы нуждаются в большом количестве актуальных и качественных данных в пригодном для анализа виде, чтобы использовать их для повышения эффективности госуправления:

- государственные данные муниципальных и региональных органов — для реализации актуальных, полезных проектов на местах благодаря более полному отображению действительного положения дел;
- качественная статистика, показывающая реальное положение дел в стране, регионах, муниципалитетах, – для принятия своевременных и точных управленческих решений;
- данные о социальном статусе и материальном положении граждан — для перехода от заявительной системы социального обеспечения к созданию и развитию механизмов оказания адресной социальной помощи;
- данные о передвижениях, проживании, пребывании горожан в течение дня — для модернизации инфраструктуры города, регулирования транспортных потоков, создания комфортной городской среды и формирования экосистемы «умных» городов в будущем;
- данные систем видеонаблюдения, мобильных операторов и интернет-провайдеров – для обеспечения национальной безопасности и безопасности граждан.

Некорректная работа с данными государства может создать проблемы этического плана:

- непреднамеренное создание алгоритмов и систем принятия решений, дискриминирующих отдельные группы населения, из-за недостаточно тщательного отбора данных для обучения алгоритмов;
- конфликт интересов общества и отдельных групп граждан и общественное неприятие инициатив, в том числе из-за отсутствия информационной поддержки и предварительной работы с общественным мнением;
- принятие на государственном уровне неверных решений с отдаленными последствиями в виде финансовых, репутационных и иных потерь для государства;
- непреднамеренное нарушение законодательства госслужащими по причине неэтичного использования данных, повлекшего за собой значительный ущерб.

Все эти проблемы требуют своевременного решения на уровнях регулирования, правоприменения, создания соответствующей инфраструктуры, внедрения культуры работы с данными в органах власти и т. д.[39]

39 Рекомендуется к просмотру видео // URL: https://youtu.be/aA1RHEy7A2c

КЛЮЧЕВАЯ РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. ГОСТ 34.321–96. Информационные технологии. Система стандартов по базам данных. Эталонная модель управления данными = Information technology. Database standards system. Reference model of dat a management: национальный стандарт Российской Федерации: издание официальное: утвержден и введен в действие Постановлением Государственного комитета Российской Федерации по стандартизации и метрологии от 22 февраля 2001 г. № 88-ст.: введен впервые: дата введения 2001-07-01 / разработан Институтом программных систем НАН Украины. – Москва: ИПК Издательство стандартов, 2001. – 24 с.

2. Осипов, Д. Л. Технологии проектирования баз данных / Д. Л. Осипов. – Москва: ДМК Пресс, 2019. – 498 с.

3. Костюкова, А.П. Базы данных: общий курс: учебное пособие / А.П. Костюкова. – Москва: Издательский дом Академии Естествознания, 2018. – 82 с.

4. ГОСТ Р 55062-2012 Информационные технологии. Системы промышленной автоматизации и их интеграция. Интероперабельность. Основные положения = Information technologies. Industrial automation systems and integration. Interoperability. Basic principles: национальный стандарт Российской Федерации: издание официальное: утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 13 ноября 2012 г. № 751-ст: введен впервые: дата введения 2013-09-01 / разработан Федеральным государственным бюджетным учреждением науки Институтом радиотехники и электроники им. В.А.Котельникова РАН. – Москва: Стандартинформ, 2018. – 8 с.

5. ГОСТ Р 50922–2006 Защита информации. Основные термины и определения = Protection of information. Basic terms and definitions: национальный стандарт Российской Федерации: издание официальное: утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 27 декабря 2006 г. № 373-ст: введен впервые: дата введения 2008-02-01 / разработан Федеральным государственным учреждением «Государственный научно-исследовательский испытательный институт проблем технической защиты информации Федеральной службы по техническому и экспортному контролю». – Москва: Стандартинформ, 2008. – 8 с.

6. Стырин, Е.М., Государственные цифровые платформы: от концепта к реализации / Е.М. Стырин, Н.Е. Дмитриева, Л.Х. Синятуллина // Вопросы государственного и муниципального управления. – 2019. – № 4. – URL: https://cyberleninka.ru/article/n/bezopasnost-baz-dannyh-problemy-i-perspektivy (дата обращения 19.10.2021)

7. О персональных данных: федеральный закон от 27.07.2006 № 152. // Собр. законодательства РФ. – 2006. – №31. – Ст. 3451

8. Этика и «цифра»: этические проблемы цифровых технологий. – Москва: РАНХиГС, 2020 – 207 с.
9. DAMA-DMBOK: свод знаний по управлению данными / [Dama International]; [перевод с английского Г. Агафонова]. – 2-е изд. – Москва: Олимп-Бизнес, 2020. – 828 с.

ПК 4.3. Управление большими данными и качеством данных

4.3.1. Ключевые понятия управления большими данными

Понятие больших данных

Согласно отчету McKinsey Institute, термин «большие данные» относится к наборам данных, размер которых превосходит возможности типичных баз данных (БД) по занесению, хранению, анализу информации и управления ею[40]. Тем не менее большие данные предполагают нечто большее, чем просто анализ огромных объемов информации. Проблема не в том, что организации создают огромные объемы данных, а в том, что б*о*льшая их часть представлена в формате, плохо соответствующем традиционному структурированному формату БД: это веб-журналы, видеозаписи, текстовые документы, машинный код или, например, геопространственные данные. Все это хранится во множестве разнообразных хранилищ, иногда даже за пределами организации.

В результате корпорации могут иметь доступ к огромному объему своих данных и не иметь необходимых инструментов, чтобы установить взаимосвязи между этими данными и сделать на их основе значимые выводы. Добавьте сюда то обстоятельство, что данные сейчас обновляются все чаще и чаще, и вы получите ситуацию, в которой традиционные методы анализа информации не способны обрабатывать огромные объемы постоянно обновляемых данных, что в итоге и открывает дорогу технологиям больших данных[41]. Сравнение между традиционными базами данных и базами больших данных представлено в табл. 4.

ТАБЛИЦА 4
ТРАДИЦИОННАЯ БАЗА ДАННЫХ И БАЗА БОЛЬШИХ ДАННЫХ

ХАРАКТЕРИСТИКА	ТРАДИЦИОННАЯ БАЗА ДАННЫ	БАЗА БОЛЬШИХ ДАННЫХ
Объем информации	От гигабайт до терабайт	От петабайт до эксабайт
Способ хранения	Централизованный	Децентрализованный
Структурированность данных	Структурированная	Полуструктурированная или неструктурированная
Модель хранения и обработки данных	Вертикальная модель	Горизонтальная модель
Взаимосвязь данных	Сильная	Слабая

Термин «большие данные» указывает не только на объем данных, но и на их разнообразие (структурированные и неструктурированные, документы, файлы, аудио- и

40 Manyika J., Chui M., Brown B. et al. Big Data: The Next Frontier for Innovation, Competition, and Productivity / McKinsey Global Institute. [S.l.,] 2011. / URL: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation#
41 Большие данные (Big Data) // TAdviser. URL: https://www.tadviser.ru/index.php/Статья:Большие_данные_(Big_Data)

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 «дата саентистами» (data scientists).

Большие данные
объединение техник и технологий, которые извлекают смысл из данных на экстремальном пределе практичности (краткая формулировка консалтинговой компании Forrester[42]).

Big Data — это:
- ETL[43]/ELT[44];
- технологии хранения больших объемов структурированных и не структурированных данных;
- технологии обработки таких данных;
- управление качеством данных;
- технологии предоставления данных потребителю.

Data Science — это:
- распознавание видео;
- распознавание текстов;
- распознавание речи;
- построение рекомендательных моделей;
- сегментация;
- кластеризация и т.д.[45]

На самом 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, 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,

42 Исследовательская компания Forrester Research предоставляет данные о рынке новых технологий, профильные консультации.

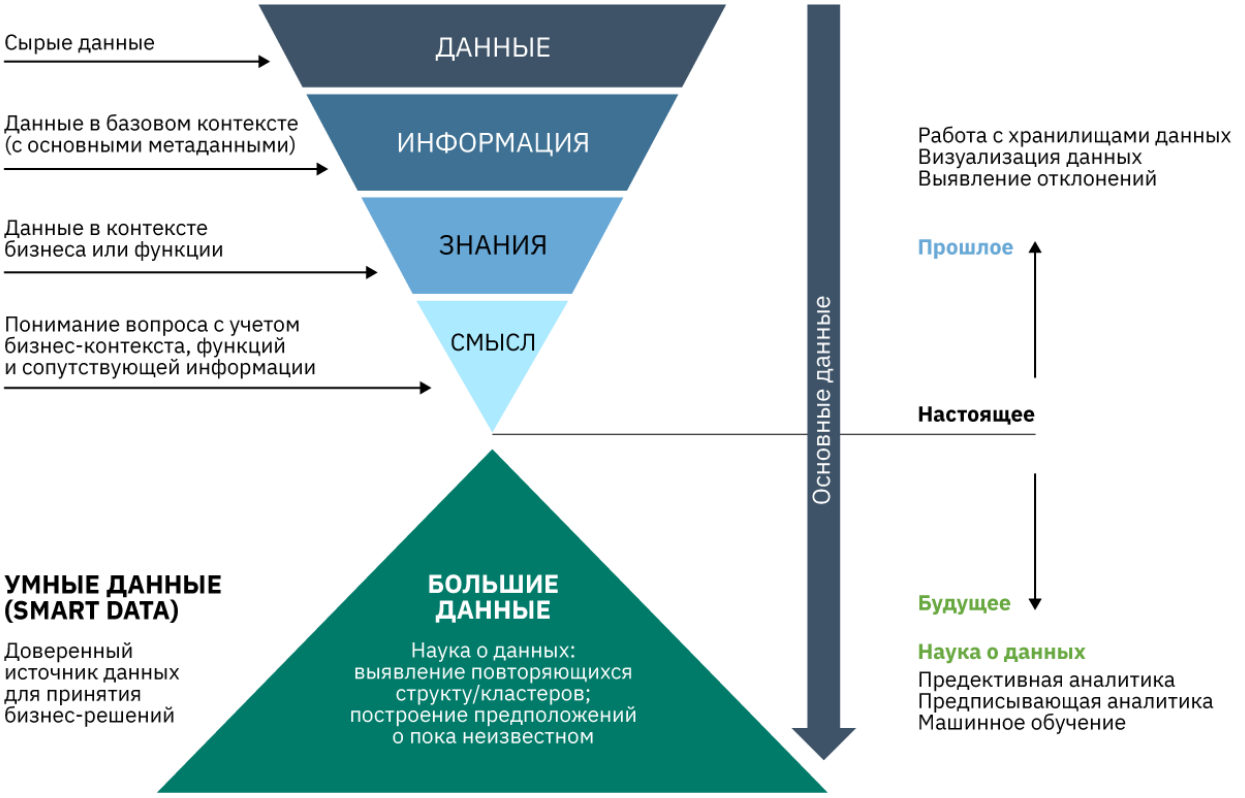
43 ETL – аббревиатура от Extract, Transform, Load (дословно - дословно «извлечение, преобразование, загрузка») – один из основных процессов в управлении хранилищами данных, который включает в себя: извлечение данных из внешних источников; их трансформация и очистка, чтобы они соответствовали потребностям бизнес-модели; и загрузка их в хранилище данных.

44 ELT – аббревиатура от Extract, Load, Transform (дословно «извлечение, загрузка, преобразование») – это процесс, с помощью которого данные извлекаются из исходной системы, загружаются в выделенный пул SQL, а затем преобразовываются.

45 Большие данные (Big Data) // TAdviser. URL: https://www.tadviser.ru/index.php/Статья:Большие_данные_(Big_Data)

машинного обучения, моделирования неопределенности, визуализации данных и других прикладных областей знания в целях 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. С появлением больших данных методы науки о данных стали использоваться для обеспечения возможности смотреть вперед. Возможность прогнозирования на основе моделей, в том числе в режиме, близком к реальному времени, с использованием разнородных данных из множества различных источников помогает организациям лучше понимать направления своего развития (рис. 22).

РИСУНОК 22
ИНФОРМАЦИОННЫЙ ТРЕУГОЛЬНИК АБАТЕ



Однако для использования преимуществ больших данных требуется изменить методы управления данными. Большинство хранилищ данных применяют традиционную реляционную модель. Большие данные, как правило, в виде такой модели не представлены. В большинстве хранилищ данных обработка тесно связана с процедурами ETL.

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.

Сбор (больших данных), анализ и визуализация (наука о данных) множества разнородных данных различных видов проводятся с целью получения ответов на те вопросы, которые будут сформулированы лишь в процессе анализа.

Цели:

- раскрытие связей между данными и бизнесом;
- итеративное включение источников данных в среду организации;
- выявление и анализ новых факторов, которые могут оказывать влияние на бизнес;
- публикация (и визуализация) достоверных данных в подходящей и этичной форме.

Принципы:

Большие данные сулят заманчивую перспективу глубокого осмысления реальности под новыми и неожиданными углами зрения, но для этого нужно для начала уметь управлять большими данными. 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 пока не сформированы исчерпывающим образом, 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.

Наука о данных (data science)

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. Это 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 категории (пример: приобретая жилье через какое-то время приобретается мебель). Затем аналитик исследует большие объемы исторических данных с целью проверить справедливость этой гипотезы и определить статистическую корреляцию 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 все новых и новых источников данных по мере наработки статистически значимых результатов для углубления и детализации полученных выводов. На эффективность практического применения методологии науки о данных влияют следующие факторы:

- богатство исходных данных как признак потенциала выявления в них скрытых закономерностей и тенденций в поведении организаций или потребителей;
- сопоставление и анализ информации: технические приемы используются для понимания смыслового наполнения данных и правильного сочетания их наборов для 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 для их распространения.

Табл. 5 сравнивает роль традиционной модели хранилища данных / бизнес-аналитики с моделями прогнозной и предписывающей аналитики, которые можно реализовать в рамках методологии науки о данных.

ТАБЛИЦА 5
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 произойдёт?	Сценарный анализ: какая последовательность действий даст желаемый результат?

Рис. 23 иллюстрирует последовательность итераций в рамках осуществления деятельности в области науки о данных. Результаты на выходе предыдущего этапа этого циклического процесса служат исходными данными для следующего.

РИСУНОК 23
ПРОЦЕСС ОСУЩЕСТВЛЕНИЯ ДЕЯТЕЛЬНОСТИ В ОБЛАСТИ НАУКИ О ДАННЫХ

Наука о данных следует общепринятой методологии познания посредством последовательного приближения к объективной истине посредством циклов наблюдений, выдвижения и опытной проверки гипотез, накопления результатов экспериментов в рамках предложенной модели и формулирования 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:

1. **Определение стратегии и потребностей бизнеса в области изучения больших данных.** Сформулируйте требования к желаемым результатам с указанием измеримых материальных выгод от их выполнения.
2. **Выбор источников данных.** Идентифицируйте пробелы в имеющейся базе информационных ресурсов и изыщите источники данных, которые позволят заполнить эти пробелы.
3. **Получение и освоение источников данных.** Получите все необходимые наборы данных или доступ к их источникам.

4. **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 уточнить требования. 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.).
5. **Интеграция и согласование данных для анализа.** Пригодность модели зависит еще и от качества источников данных. Используйте данные из надежных и достоверных источников. По мере необходимости привлекайте средства интеграции, очистки и доработки данных с целью повысить качество и полезность вводимых наборов.
6. **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, по мере накопления статистики дорабатывайте и саму модель. Машинное обучение подразумевает многократные прогоны больших массивов реальных данных через модель выявления закономерностей для проверки гипотез и внесения корректив в настройки алгоритмов (например, для выявления выпадающих из общего статистического ряда значений). В процессе такой проработки окончательно уточняются 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.
7. **Внедрение и мониторинг.** Модели, которые выдают полезную информацию, можно переносить в производственную среду и использовать для текущего мониторинга ситуации с целью получать данные или выявлять нежелательные тенденции, ставящие 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, основными данными и т.д.).

Большие данные

На ранней стадии формирования понятия большие данные определялись по признаку соответствия трем V-характеристикам: Volume — объем, Velocity — скорость, Variety — разнообразие. Вместе с широким распространением концепции больших данных 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:

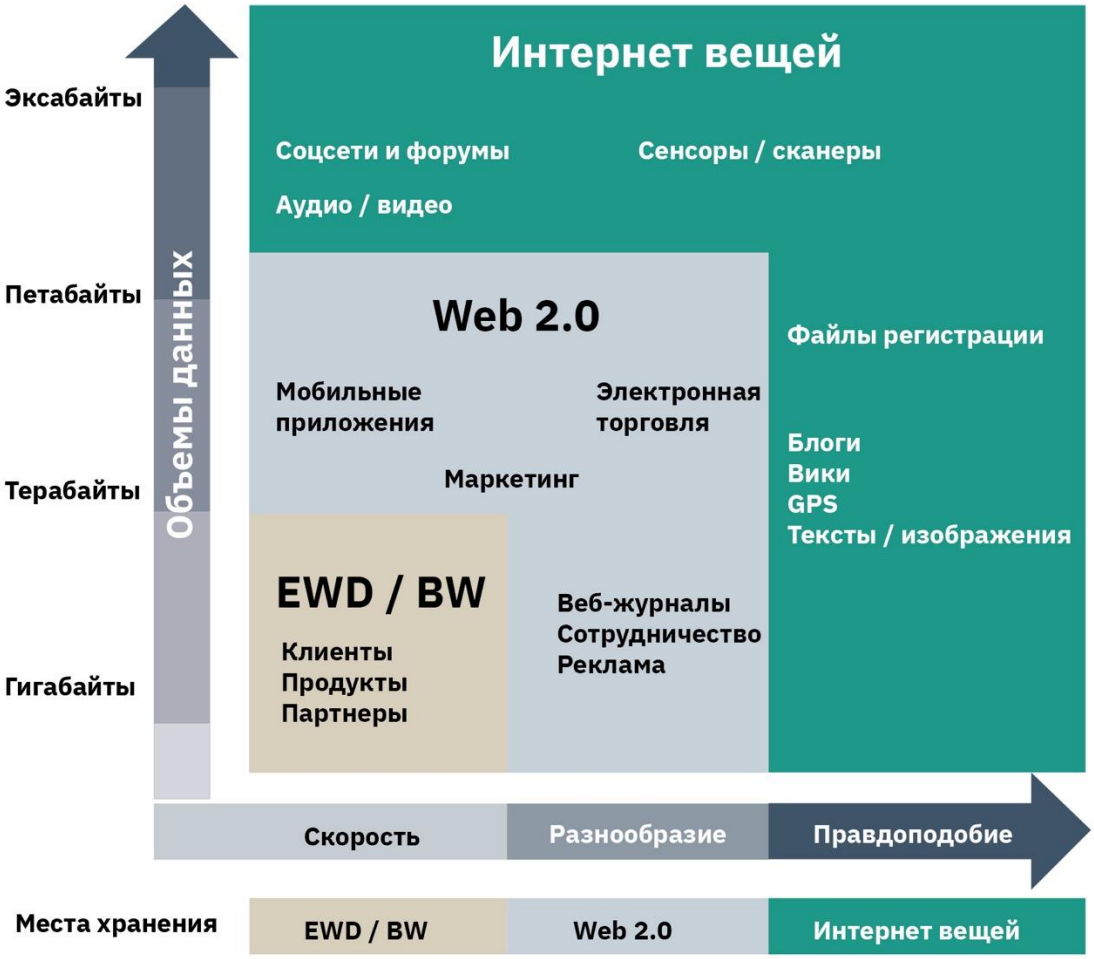
- **Volume** — объем как мера количества данных: большие данные включают миллиарды полей или записей, описывающих тысячи сущностей или элементов.
- **Velocity** — скорость регистрации/генерирования, обработки или распространения: большие данные зачастую не только создаются, но и распространяются и даже 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 для практического использования.
- **Volatility** — волатильность как мера непостоянства: большие данные крайне переменчивы, что весьма ограничивает сроки годности полученных результатов.
- **Veracity** — правдоподобие по критериям проверки подлинности источника.

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[46]/BI обработка подобных объемов становится весьма проблематичной, поскольку для этого нужен ЦОД с серьезными серверными мощностями и пропускной способностью каналов связи для загрузки, моделирования, очистки и анализа данных. Проблему часто решают за счет массивно параллельной архитектуры обработки данных или сочетания параллельной обработки с распределенными вычислениями и облачными хранилищами. Однако все это не более чем локальные и временные решения, поскольку проблемы, обусловленные нарастанием объемов и потоков больших данных, имеют далеко идущие последствия. Колоссальные размеры наборов данных требуют изменений общего подхода к хранению данных, доступу к данным, концептуального представления о данных (в частности, отказа от традиционного мышления категориями структур данных, описываемых реляционными моделями), а также методов управления данными.

Рис. 24 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.

46 Data Warehousing – хранилища данных

РИСУНОК 24
МАСШТАБЫ ЗАДАЧ В ОБЛАСТИ ХРАНЕНИЯ ДАННЫХ



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, согласовать его с существующими средствами сбора и анализа данных.

Рис. 25 описывает концептуальную архитектуру рабочей среды для областей DW/BI и больших данных. Ключевое различие между средой работы с большими данных и традиционным хранилищем 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 раньше, чем будут приведены к совместимому с имеющейся структурой данных виду, что необходимо для интеграции. Во многих случаях большим данным и не требуется интеграция в традиционном смысле приведения к общей модели. Вместо их подготовки к использованию в составе общего комплекса интегрированных данных применяется метод выборочного включения больших 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).

РИСУНОК 25
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 без моделирования данных предприятия. Однако подобное упрощение рискует обернуться утерей большей части информативного содержания обрабатываемых данных в процессе беспорядочного освоения. Отсюда вытекает необходимость управляемого сбора метаданных о накапливаемых данных, чтобы со временем не утрачивалось понимание их смысла и назначения.

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 социальных сетях, 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).

Озера данных

Озеро данных
среда накопления массы разнородных по типу и структуре данных, откуда их можно брать для сохранения, оценки или анализа.

Озера данных создают 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 как сложная конфигурация с использованием продвинутых средств обработки данных, включая системы управления хранилищами (например, Hadoop), службы кластеризации, преобразования и интеграции данных. Все подобные обработчики озер данных специально ориентированы на работу на базе распределенной инфраструктуры хранения данных и имеют инструменты, позволяющие собирать данные согласно заданной структурной конфигурации.

Главный риск заключается в том, что озеро данных 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[47].

47 DAMA-DMBOK: Свод знаний по управлению данными / Dama International; пер. с англ. Г. Агафонова.2-е изд. М.: Олимп-Бизнес, 2020. С. 623–668.

4.3.2. Техники и методы анализа больших данных

Техники и методы анализа, применимые к Big data по McKinsey:

- Data Mining;
- Краудсорсинг;
- смешение и интеграция данных;
- машинное обучение;
- искусственные нейронные сети;
- распознавание образов;
- прогнозная аналитика;
- имитационное моделирование;
- пространственный анализ;
- статистический анализ;
- визуализация аналитических данных.

Далее в документе рассматриваются некоторых их основных техник и методов анализа больших данных.

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 в статистике, комбинаторике 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.

Машинное обучение занимается структурным построением алгоритмов познания и усвоения знаний. Выделяют три типа таких алгоритмов:

- **обучение с учителем** основано на применении обобщенных правил (например, настраиваемый фильтр спама в почтовом приложении);
- **обучение без учителя** основано на выявлении скрытых паттернов, связей, закономерностей (например, интеллектуальный анализ данных);
- **обучение с подкреплением** основано на достижении цели (например, выигрыша шахматной партии).

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 стимулирует организации к инициативам по поиску глубинных закономерностей посредством многократного повторения процессов.

Хотя машинное обучение использует весьма новые способы получения данных, в этой новой области знания должны соблюдаться все традиционные принципы этичного обращения с данными, 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 прозрачны такие алгоритмы 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 выходят за пределы возможностей человека понимать и интерпретировать логику машинного мышления. Необходимость обеспечения прозрачности принятия решений по мере дальнейшего совершенствования функциональности неконтролируемого самообучения и его применения во все более широком спектре ситуаций, вероятно, будет только возрастать.

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, социальные сети, 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 к 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.

Проиллюстрируем этот подход простейшим примером 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 наличие слов «плохой», «дрянь», «гадость» может служить признаком негативного отношения. Распределив отзывы по категориям, можно выяснить отношение, преобладающее в целевом сообществе (например, в данной соцсети, блоге и т. п.). К слову, реальные чувства и эмоции, вызываемые предметом обсуждения, не так легко 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.

Интеллектуальный анализ данных и текстов

Интеллектуальный анализ данных (извлечение информации, data mining)
применение к массивам разнородных данных разнообразных алгоритмов выявления скрытых структурных закономерностей.

Интеллектуальный анализ данных постепенно отделился от машинного обучения и сложился как подобласть 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.

Извлечение данных и интеллектуальный анализ текстов основаны на использовании ряда стандартных технических приемов, включая описанные ниже.

1. Профилирование

Заключается в описании характерных типов поведения людей, групп или организаций и используется для определения признаков нормального поведения с целью выявить серьезные отклонения 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.

2. 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.

3. Ассоциирование

Определение ассоциаций 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.

4. Кластеризация

Группировка элементов в кластеры по признаку близкого сродства или общности неких характеристик, она упрощает и ускоряет статистический анализ типичных схем и стереотипов поведения. Классический пример кластеризации — сегментация потребительского рынка.

5. Самоорганизующиеся карты

Метод кластерного анализа нейронных сетей, известный также под названием самоорганизующихся карт Кохонена или топологически упорядоченных карт. Их использование позволяет снизить размерность пространства оценки без ущерба для результатов аппроксимации. Устранение избыточных пространственных измерений по эффективности не уступает изъятию вырожденных переменных из алгебраических уравнений — и решать проще, и результат нагляднее.

Предиктивная аналитика

Предиктивной аналитикой называют разновидность обучения с учителем, где пользователи пытаются смоделировать элементы данных и предсказать будущие исходы по оцениваемым вероятностям событий. В теории вероятностей и математической статистике 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.

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, например 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 фильтрами экспоненциального сглаживания и более, позволяющие учитывать, например, недельные колебания на фоне сезонных.

Предписывающая аналитика

Предписывающим анализом называют прогнозный анализ, дополненный определениями корректирующих воздействий на ситуацию с целью изменить конечные результаты, а не ограничивающийся простым их прогнозированием. Таким образом, предписывающая аналитика позволяет предсказывать, что случится, когда это случится и — главное — при совокупности каких факторов это случится. Будучи способным демонстрировать последствия различных сочетаний решений, предписывающий анализ позволяет моделировать их комбинации в целях максимизации выигрыша или минимизации риска. Методы предписывающего анализа удобны тем, что предусматривают возможность непрерывной подачи скорректированных вводных на вход и перерасчета прогнозов с выдачей скорректированных предписаний. Это повышает и точность прогноза, и результативность предписаний.

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 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 неструктурированных данных в аналитическую модель. Однако анализ неструктурированных данных осложняется необходимостью предварительного отделения интересующих исследователей данных от лишних элементов.

Сканирование и тегирование — два способа «выуживания» полезных неструктурированных данных из озера, позволяющий отфильтровать их от лишних данных 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.

Операционная аналитика

Концепция операционной аналитики (она же операционная BI, бизнес-аналитика, потоковая аналитика данных и т. п.) появилась в результате интеграции функций анализа данных в режиме реального времени в операционную деятельность. 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.

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.

Визуализация данных

Визуализация данных
процесс интерпретации концепций, идей и фактов посредством наглядных представлений, 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 подвести зрителей к полезным выводам о скрытых возможностях, рисках или смыслах.

Визуализация данных, по мнению экспертов Forrester[48], необходима компании в следующих случаях:

- Невозможно увидеть паттерн, не прибегая к визуализации данных. Традиционные отчеты, строки, колонки и таблицы не дают понимание общей картины в целом, что может привести к неверной интерпретации данных.
- Значительные массивы данных невозможно разместить на одном экране. Как бы ни был мелок шрифт, сотни и тысячи строк данных на одном экране будут абсолютно не информативны в сравнении с визуальным представлением этой же самой информации. Эдвард Тафт приводит пример отображения 21 тыс. значений данных на карте США, поместившейся на одном экране[49].
- Невозможно передать глубину данных посредством информации, представленной на одном экране. Например, в фармацевтике при анализе свойств тех или иных препаратов используются сотни различных психофизических параметров человека, охватить которые в линейном виде не представляется возможным. Однако медицинские эксперты присваивают каждому параметру разноцветные пиксели и помещают их в динамические модели, где любой участок при необходимости может быть увеличен и проанализирован детально[50].

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 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 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.

---

48 The Forrester Wave: Advanced Data Visualization (ADV) Platforms, Q3 2012 // Forrester. URL: https://www.forrester.com/report/The+Forrester+Wave+Advanced+Data+Visualization+ADV+Platforms+Q3+2012/-/E-RES71903
49 Tufte E. The Visual Display of Quantitative Information. 2nd ed. Cheshire, 2007.
50 Визуализация данных // TAdviser. URL: https://www.tadviser.ru/index.php/Статья:Визуализация_данных

По мере дальнейшего развития и совершенствования средств визуализации организациям придется взращивать такие команды бизнес-аналитиков, которые смогут обеспечивать им конкурентоспособность в мире, который все больше управляется через компьютер в потоковом режиме. И вот тогда в бизнес-аналитических отделах будут крайне востребованы эксперты с навыками визуализации: знатоки данных, художники данных, визионеры данных, в дополнение к традиционно ценным архитекторам и разработчикам моделей данных. Это будет более чем оправданно, если помнить о рисках, проистекающих от искажающих восприятие обманчивых визуальных представлений.

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 за счет функциональности 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 с помощью внешнего ключа. Техническая возможность создания различных данных, например, весьма полезна для получения пользовательских представлений и идеально подходит для реализации задач, которые возникают в 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 сегментами, 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.

Динамические, сетевые бизнес-процессы, построенные с высокой степенью детализации, могут генерировать миллиарды и триллионы байтов информации каждый месяц. Учитывая это, следует понимать, что требования больших объемов данных традиционно опережали любые улучшения цены/производительности технологий. К счастью, новые архитектуры и подходы, развиваемые в последнее десятилетие, смогли значительно упростить управление этими чудовищными массивами данных и были, в конце концов, интегрированы в корпоративную архитектуру многих крупных компаний. К таким архитектурам и подходам в том числе относятся:

1. **Программно-аппаратные комплексы и ускорители для баз данных.**

Значительные усовершенствования технологии реляционных баз данных на протяжении последнего десятилетия позволяют быстро и эффективно загружать и обрабатывать на одной платформе терабайты и даже петабайты данных. Специализированные устройства объединяют хранение, обработку, подключение и быстрое выполнение запросов на выделенной программно-аппаратной платформе, оптимизированной для работы и управления базой данных. Ускорители баз данных используют самые последние достижения хранения данных и оптимизации запросов для снижения размеров базы данных и повышения скорости выполнения сложных запросов. Если просто обновление аппаратной платформы может повысить производительность традиционной реляционной базы данных в два раза, то применение специализированных устройств и ускорителей способно улучшить показатель «цена/производительность» почти в 100 раз. Наиболее важно, что эти технологии упрощают управление и администрирование, устраняя потребность в квалифицированной настройке и конфигурировании.

2. **NOSQL хранилища данных.**

Технология Not-Only-SQL, буквально рожденная в среде Интернета, с самого начала была спроектирована для управления огромными распределенными наборами данных, запрос к которым должен был выполняться за миллисекунды. Вместо нормализации данных по реляционным таблицам, которые затем должны объединяться для ответов на запросы, сверхбольшие массивы данных распределяются по сотням или тысячам процессоров, организованных так, чтобы связанные данные располагались рядом. Запросы выполняются параллельно на всех процессорах; каждый возвращает ответы на основе своих локальных данных. Этот невероятно простой и масштабируемый подход оказался очень эффективным

и гибким, позволяя совместно хранить данные самых разных типов, а также выполнять сложные запросы.

3. Автоматизированная аналитика.

Чтобы получить преимущество от обработки сверхбольших массивов данных, требуется хорошая аналитика, и в большинстве компаний это удел небольшой группы высококвалифицированных специалистов. Выборка, очистка и обработка терабайтов данных часто представляют собой скорее искусство, чем науку, и большинство аналитиков скажет вам, что выполняемые ими ручные операции не могут быть автоматизированы. Однако прогресс самообучающихся алгоритмов, генетических алгоритмов и автоматизированного тестирования, достигнутый в последнее десятилетие, привел к появлению программ, способных распознавать образцы, делать заключения и улучшаться с течением времени, — другими словами, они самообучаются. Эти системы могут не всегда превосходить своих «коллег»-людей, но их автоматизированные процессы могут оказаться единственным способом масштабирования в соответствии с требованиями сверхбольших массивов.

Велико значение технологии обработки больших данных для снижения объемов и сложности хранимой информации. Например, большинство крупных компаний используют процессы управления важными для бизнеса событиями, такими как реакция на потенциальную попытку мошенничества или на обращение за поддержкой. Технология работы с большими данными позволяет сканировать миллиарды транзакций для определения или предугадывания важных бизнес-событий; после чего в качестве реакции могут запускаться уже традиционные технологии и процессы. В этом качестве технология сверхбольших данных выступает как преобразователь, превращающий какофонию данных в полезную, управляемую информацию[51].

Большие данные — важный шаг человечества в постоянном стремлении количественно измерить и постичь окружающий мир. То, что прежде невозможно было измерять, хранить, анализировать и распространять, находит свое выражение в виде данных. Использование огромных массивов данных вместо их малой доли и выбор количества в ущерб точности открывают путь к новым способам понимания мира. Это подталкивает общество отказаться от освященного веками поиска причинности и в большинстве случаев пользоваться преимуществами корреляций[52].

51 Беркана А. Что такое Big data: собрали все самое важное о больших данных // Rusbase. URL: https://rb.ru/howto/chto-takoe-big-data/

52 Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / пер. с англ. И. Гайдюк. М.: Манн, Иванов и Фербер, 2014. 240 с.

4.3.3. Управление качеством данных

Основные понятия

Данные — ключевой фактор цифровой экономики, однако на деле успешных проектов оказывается много меньше, чем ожидается. Виной тому низкое качество данных, беспорядок в которых не позволяет получить достоверные результаты эксплуатации сложных и дорогих систем интеллектуальной автоматизации.

Переизбыток различных данных и изобилие разнообразных средств работы с ними могут ввести в заблуждение: может показаться, что для монетизации данных и повышения производительности сотрудников достаточно лишь инвестировать в передовые инструменты (машинное обучение, средства бизнес-аналитики и пр.), позволяющие, например, разработать индивидуальные привлекательные предложения за счет глубокого понимания рынка и потребителей. Но большие данные (3V: Variety, Velocity, Volume), которыми обладают компании, ничего не стоят без полноты и достоверности (Veracity).

Количество, скорость сбора и разнообразие сами по себе не гарантируют получения массива качественных данных, пригодных для работы. Мало того, как свидетельствуют многочисленные опросы, избыток данных вызывает у сотрудников стресс, а разнообразие информации, разрозненность ее источников и отсутствие стандартизации — это ключевые факторы, мешающие компаниям получать новые знания из данных.

Безусловно, инструменты необходимы: ожидать, что сегодня сотрудники будут без них работать с данными, — то же самое, что предлагать им идти на охоту без оружия и снаряжения. Сотрудники уже «скрылись в чаще данных», но слепо блуждают в ней и даже если случайно набредут на нужный пул полезных сведений, могут оказаться перед ним безоружны. Тем не менее инструменты доступа к данным — это необходимое, но недостаточное условие. Требуется создать условия для уверенного самостоятельного использования данных сотрудниками. Нужна корпоративная стратегия управления как данными в целом, так и их качеством, а главное, необходима культура работы с данными.

Конвейер на производстве позволил в свое время на порядок уменьшить время сборки автомобилей. Аналогичный конвейер (рис. 26) нужен и при работе с данными: он должен предусматривать автоматическую проверку качества всех «узлов» для их доставки на линию «сборки». Вполне логично предположить, что на этом конвейере работают множество специалистов: эксперты по качеству данных знают, где взять данные и как оценить их качество; инженеры по данным понимают, как получить данные, часто разбросанные по разным источникам, хранящиеся в разных форматах, поступающие с разной скоростью и в разных объемах; аналитики способны извлечь из данных полезную информацию; дата саентисты (data scientist) могут обнаружить скрытые закономерности, подозрительные аномалии или логические связи в данных. Здесь нужны и другие специалисты по организации и поддержке (DataOps): коммуникаторы, юристы, сотрудники службы информационной безопасности[53].

Термин «качество данных» (Data Quality, DQ) распространяется как на характеристики, связанные с высоким качеством данных, так и на процессы измерения или повышения качества данных.

53 Волков Д., Незнанов А. Качество данных: от стратегии к практике // Lobanov Logist. URL: https://www.lobanov-logist.ru/library/all_articles/64180/

РИСУНОК 26
КОНВЕЙЕР МОНЕТИЗАЦИИ ДАННЫХ

КЛАССИФИКАЦИЯ ПЕРВИЧНЫХ ДАННЫХ

Определяют следующую классификацию первичных данных:

- **Мастер-данные (master-data)** определяют ключевые, представляющие особую ценность для организации или бизнеса и относительно редко изменяемые сущности.
- **Разделяемые справочники (reference data)** систематизируют и классифицируют другие данные, а также связывают между собой данные различных организаций. Сегодня на эту роль претендуют открытые данные (open data) любых уровней — от федеральных до местных.
- **Оперативные (транзакционные) данные (transactional data)** отражают информацию о ходе исполнения бизнес-процессов.
- **Исторические данные (historical data)** образованы из прошлых версий мастер-данных, справочников и транзакционных данных, возникших после завершения соответствующих бизнес-процессов. Такие данные явно привязаны ко времени свершения с указанием периода актуальности для записей справочников и реестров, конкретной исторической глубины. По сути, исторические данные — это один из важнейших ресурсов компании или организации, такие данные купить нельзя.

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.

Качество данных — это свойство данных удовлетворять предъявляемым к ним требованиям. Причем в соответствии с базовыми принципами качества данных (определяются стандартами серии ISO/TS 8000 «Data quality — Part 1: Overview», ГОСТ Р 56214–2014/ISO/TS 8000-1:2011 «Качество данных. Часть 1. Обзор») понятие качества данных затрагивает только те данные, которые участвуют в принятии какого-либо управленческого решения.

Качество данных зависит от контекста и потребностей потребителей данных:

- **качество** — степень соответствия совокупности присущих характеристик объекта требованиям;
- **требование** — потребность или ожидание, которое установлено, предполагается или является обязательным
- **качество данных** — степень, с которой набор характеристик, присущих данным, отвечает требованиям[54].

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.

Управление качеством данных

планирование, организация и контроль выполнения работ по применению стандартных методов управления качеством к данным с целью обеспечить их пригодность к использованию[55].

Программы качества данных преследуют 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 проведение измеримых улучшений качества данных на основе требований их потребителей.

54 ГОСТ Р 56215-2014/ISO/TS 8000-150:2011 Национальный стандарт Российской Федерации. Качество данных. Ч. 150. Основные данные. Структура управления качеством // Standartgost.ru. URL https://pdf.standartgost.ru/catalog/Data2/1/4293767/4293767872.pdf

55 DAMA-DMBOK: свод знаний по управлению данными / Dama International; пер. с англ. Г. Агафонова. 2-е изд. М.: Олимп-Бизнес, 2020. С. 562.

Принципы управления качеством данных

Управление качеством данных должно быть ориентированы на соблюдение следующих принципов:

- **Критичность.** Чем критичнее данные для организации и ее клиентов, тем больше внимания следует уделять им 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 следует контролировать и обеспечивать 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 (каждое звено в цепи передачи данных должно гарантированно обеспечивать высокое качество выходных данных).
- **Предупреждение.** Основное внимание нужно уделять 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 следует измерять объективно и согласованно, а методологию измерения и оценки показателей качества — доводить 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)[56].

56 DAMA-DMBOK: свод знаний по управлению данными / Dama International; пер. с англ. Г. Агафонова. 2-е изд. М.: Олимп-Бизнес, 2020. С. 566.

Условия управления качеством основных данных

Для успешного управления качеством основных данных организация должна выполнять следующие основные условия:

- **Вовлечение в работу сотрудников.** На всех уровнях управления качеством данных активное вовлечение в работу всех участников процесса способствует повышению качества данных любой организации. Несмотря на то что на качество данных в наибольшей степени влияет самый начальный уровень работы, контроль за управлением информацией на среднем уровне также необходим. Вовлечение в работу менеджеров самого высокого уровня управления качеством данных очень важно для оптимизации и повышения авторитета организации и всех происходящих в ней процессов.
- **Подход к процессу.** Концентрация на точности данных измерений и на внесении поправок — это не достаточная мера для повышения качества данных организации. Желаемое качество данных более эффективно достигается в том случае, если процесс управления качеством влияет и на сам источник данных.
- **Постоянное усовершенствование.** Качество данных постоянно совершенствуется благодаря обработке, проверке измерений и постоянной корректировке данных. Однако даже при осуществлении этих процессов ошибки неизбежны. Повторяемость ошибок можно избежать только при анализе и устранении коренных причин, мешающих процессам повышения качества данных. Организации должны совершенствовать не только процессы управления данными, но и все бизнес-процессы, в которых напрямую используются информационные данные.
- **Обмен основными данными.** Все процессы управления качеством основных данных совпадают или соответствуют автоматически контролируемым и проверяемым требованиям, обеспечивающим постоянный обмен основными данными характеристик между организациями и системами[57].

57 ГОСТ Р 56215-2014/ISO/TS 8000-150:2011 Национальный стандарт Российской Федерации. Качество данных. Ч. 150. Основные данные. Структура управления качеством // Standartgost.ru. URL https://pdf.standartgost.ru/catalog/Data2/1/4293767/4293767872.pdf

Измерение 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.

Единой универсальной классификации измерений качества данных до сих пор не выработано. Измерения включают как характеристики, оцениваемые по вполне объективно измеримым показателям (полнота, действительность, соответствие формату и т. п.), и те, которые в значительной степени зависят 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).

Далее приводится набор общепринятых измерений качества данных с определениями и описаниями подходов к их измерению:

1. Актуальность

Под актуальностью данных понимают совокупность характеристик, относящихся к соблюдению сроков или графиков их получения, синхронизации или обновления.

Показатели актуальности 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 волатильные данные сохраняют актуальность в течение короткого периода времени. 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.

2. 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 либо же через проверочные правила. Области допустимых значений определяются с учетом типа, формата, разрядности и точности/погрешности измерения ожидаемых на входе величин.

Также можно устанавливать допустимые сроки годности данных, например, поступающих с радиочастотных идентификаторов (RFID) или с каких-нибудь датчиков контрольно-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.

3. Полнота

Все ли требующиеся данные наличествуют? Полноту можно измерять 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: в обязательных для заполнения полях; в условно обязательных полях; в необязательных полях; отсутствие данных в полях неприменимых атрибутов. Измерения на уровне набора данных могут требовать сравнения с источником записей или историческими уровнями полноты данных.

4. 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 данными статистики по каждому региону (например, с 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.

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.

В данный перечень также могут быть включены и требования к размеру и составу наборов данных, которыми обмениваются различные системы, и динамические параметры согласования данных по времени. Можно определять требования согласованности:

- между различными подмножествами значений атрибутов внутри одной и той же записи (внутренняя непротиворечивость записи);
- между множествами значений атрибутов из различных наборов и записей (согласованность записей);
- между множествами значений одних и тех же атрибутов одной и той же записи в различные моменты времени (хронологическая согласованность).

Наконец, под согласованностью можно понимать 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 ± 5% (или ± 3σ).

6. Соответствие

Измерить степень близости данных к реальности 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 из надежного источника справочных данных (например, Dun & Bradstreet).

7. 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, уникальность проверяется тестированием структуры ключей.

8. 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 («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.

Взаимосвязи между параметрами качества данных представлены на рис. 27.

РИСУНОК 27
ВЗАИМОСВЯЗИ МЕЖДУ ПАРАМЕТРАМИ КАЧЕСТВА ДАННЫХ

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.

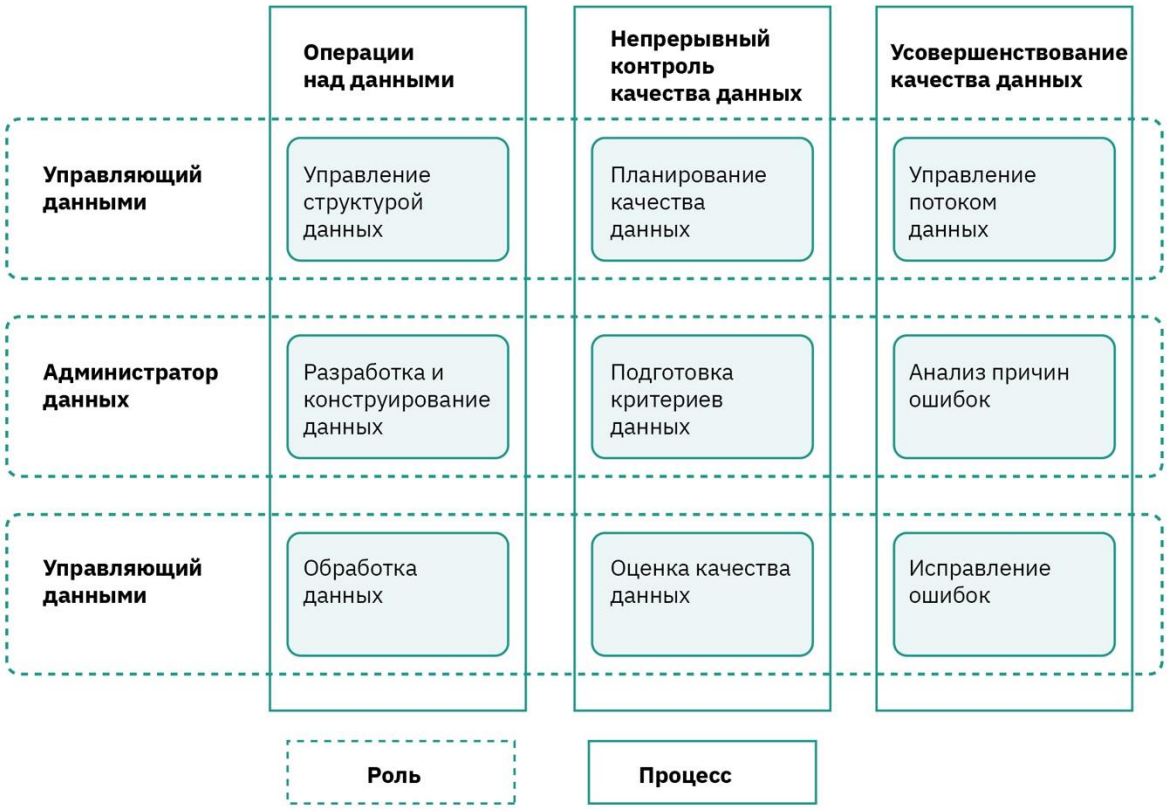
Наличие надежного процесса определения данных через метаданные также позволяет организации формализовать и документировать стандарты и требования, на соответствие которым будут проверяться данные в рамках контроля качества. Иными словами, качество данных определяется их соответствием ожиданиям.

Хорошее управление метаданными также способствует оптимизации усилий, направленных на повышение качества данных в масштабах организации. Репозиторий метаданных вполне можно использовать 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 выработки общеорганизационного консенсуса относительно приоритетов и драйверов улучшений.

Структура управления качеством основных данных

Структура управления качеством основных данных графически представлена в матрице данных 3x3 на рис. 28. Структура состоит из трех процессов верхнего уровня: операций над данными непрерывного контроля качества данных и обновления или уточнения качества данных. Каждый процесс верхнего уровня сегментируется на три процесса в соответствии с ролью человека, осуществляющего такой процесс. Процессы соотносятся друг с другом в зависимости от их очередности и входящих/исходящих данных.

РИСУНОК 28
СТРУКТУРА УПРАВЛЕНИЯ КАЧЕСТВОМ ДАННЫХ



Процессы, осуществляемые управляющим данными, дополняют процессы, осуществляемые под руководством администратора. Процесс управления структурой данных строится в соответствии с руководящими принципами действия такой структуры. Направление и управление потоком данных предусматривают наличие необходимой информации для анализа причин ошибок и плана повышения качества, определяющего задачи и цели процесса повышения качества, а также определяющего необходимые критерии работы.

К процессам, объединяемым под руководством администратора, относятся разработка и конструирование данных, подготовка критериев качества и анализ причины ошибок. Эти три процесса контролируют, координируют данные. Процесс конструирования данных помогает обеспечить качество операций над данными путем усовершенствования структуры данных. Процесс подготовки критериев качества обеспечивает методологию оценки качества данных. Процесс анализа причин ошибок позволяет избежать повторения таких ошибок.

Процессы, осуществляемые техническим персоналом, сводятся к обработке данных, оценке качества данных и исправлению ошибок. Эти процессы следуют в определенной последовательности: сначала обработка данных: создание, поиск, обновление, преобразование, удаление. Все это осуществляется в соответствии с инструкциями и руководящими принципами. Для выявления ошибок, не отмеченных при обработке данных, процесс оценки качества данных осуществляется в режиме реального времени или периодически. В случае обнаружения ошибки при обработке данных исправление такой ошибки производится в соответствии с установленным способом выполнения подобных действий.

В целом политика по обеспечению качества данных довольно постоянна, и это позволяет организации ставить перед собой конкретные цели. Такая политика является одним из основных факторов контроля, влияющих на работу всей структуры[58].

58 ГОСТ Р 56215-2014/ISO/TS 8000-150:2011 Национальный стандарт Российской Федерации. Качество данных. Ч. 150. Основные данные. Структура управления качеством // Standartgost.ru. URL https://pdf.standartgost.ru/catalog/Data2/1/4293767/4293767872.pdf

Базовые стандарты, регламентирующие управление качеством данных

Для исключения разночтений при сборе простых характеристик нужной оценки качества данных применяются стандарты (начиная с ГОСТ Р 56215) регламентирующие понимание и применение терминов.

Стандарты группы ISO/TS 8000 и перевод (ГОСТ Р 56214-2014/ISO/TS 8000-1:2011 «Качество данных. Часть 1. Обзор») – более 20 спецификаций, к которым сейчас активно добавляются новые:

- части 1–99: «Качество общих данных»;
- части 100–199: «Качество основных данных»;
- части 200–299: «Качество данных в транзакциях»;
- части 300–399: «Качество данных о продукции».

Принципы стандартов 8000:

- качество применимо к данным, которые имеют определенное назначение, их учитывают при принятии какого-либо решения;
- качество данных затрагивает нужные и подходящие данные, уместные в подходящем месте в подходящее время;
- качество данных отвечает требованиям потребителя;
- качество данных предотвращает повторение дефектов данных и сокращает избыточные расходы.

Международные стандарты группы ISO 25000 — это три главных стандарта: 25010, 25012 и 25040. Они определяют качество программного продукта с акцентом на общей модели качества данных, представленных в структурированном виде для информационной системы, а также критерии качества «продукта данных» как специального вида программного продукта.

Инструменты управления качеством данных[59]

Для поддержки цикла обеспечения качества данных необходимы инструменты, помогающие сформировать единый взгляд на данные. Рынок средств управления качеством данных растет почти на 20% в год и к 2022 году составит 1,4 млрд долл. Причинами такого интереса являются увеличение объемов данных, рост количества источников, усиление регуляторного давления и рисков для репутации бренда.

Ключевые игроки рынка решений по управлению качеством данных: Ataccama, Alteryx, Experian, IBM, Informatica, Information Builders, Microsoft, Oracle, Pitney Bowes, SAP, SAS Institute, Syncsort, Talend, Trianz, Tamr, Tibco и Unidata. Инструментов, используемых для управления качеством данных, достаточно много, и условно их можно разделить на пять групп:

- универсальные решения: платформы от Informatica, IBM, MS Data Quaity Services, SAP, Unidata;
- специализированные системы: Informatica Data Quality, SAS DQ, SAP Data Services, IBM InfoSphere QualityStage, Unidata, Experian Data Quality, TIBCO Clarity; кроме этого, соответствующие решения предлагают компании Talend, Ataccama;
- управление реестрами и справочниками: Informatica Master Data Management, IBM InfoSphere MDM, Ataccama MDC, Talend, Unidata;
- интеграция и очистка данных: Informatica Power Center, SAS Data Management, Syncsort Trillium, TIBCO Clarity, Talend Data Quality, Alteryx, Pentaho, Unidata;
- управление метаданными и Data Governance: Informatica Enterprise Data Catalog и Axon, Alation, Collibra Data Governance Center, Unidata.

Кроме того, широкое распространение получили открытые экосистемы работы с данными, например, на основе языка R (cran.r-project.org/package=dlookr, cran.r-project.org/package=dataQualityR).

Рынок инструментов управления качеством данных весьма динамичный, и его мониторинг осуществляется на многих ресурсах, например: G2 Grid for Data Quality (www.g2.com/categories/data-quality). Имеются обзоры от аналитиков, например: www.gartner.com/reviews/market/data-quality-tools.

59 Волков Д., Незнанов А. Качество данных: от стратегии к практике // Lobanov Logist. URL: https://www.lobanov-logist.ru/library/all_articles/64180/

Методология оценки и повышения качества данных на примере нсуд

В отличие от вертикально интегрированных компаний (в них, как правило, присутствует централизованная координация работ по управлению данными), обеспечению качества государственных данных в таких масштабных проектах, как Национальная система управления данными (НСУД)[60], препятствуют противоречивые интересы различных ведомств, нормативные ограничения в принятии решений и продолжительность изменений нормативно-правовых актов. Интеграционные проекты для различных ведомственных информационных систем, сопровождаемые долгими согласованиями изменений между различными ведомствами, также усложняют процессы управления качеством данных.

Задача НСУД — обеспечить доверие граждан и бизнеса к государственным данным, для чего необходимы:

- **полнота данных** (наличие сведений обо всех объектах учета, но без избыточности);
- **достоверность** (актуальность и реальность данных на момент запроса, отсутствие ошибок);
- **согласованность** (соответствие эталонным данным);
- **уникальность** (исключение дублирования и избыточных атрибутов, использование эталонных атрибутов).

Нормативное правовое обоснование оценки качества данных:

- **до** внедрения СНУД:
  - законодательно качество данных не регулируется;
  - качество данных отрывисто определяется в нескольких нормативных документах, не связанных между собой;
- **после** внедрения НСУД:
  - разработаны методические рекомендации по оценке параметров качества данных;
  - в компоненты НСУД «зашиты» метрики качества данных по каждому отдельному атрибуту.

В настоящее время нет единого документа, которой бы закрепил положение о качестве данных, содержащихся в государственных информационных ресурсах (ИР).

Важно, что НСУД оценивается качество данных ИР, а не качество ИР, при этом параметры качества «безопасность» определяет информационная система или ИР, а не сами данные.

Объекты оценки качества данных в НСУД и фазы ведения ИР представлены на рис. 29 и рис. 30.

60 Распоряжение Правительства Российской Федерации от 03.06.2019 № 1189-р «Об утверждении Концепции создания и функционирования национальной системы управления данными и плана мероприятий ("дорожной карты") по созданию национальной системы управления данными на 2019–2021 годы».

РИСУНОК 29
ОБЪЕКТ ОЦЕНКИ КАЧЕСТВА ДАННЫХ В НСУД

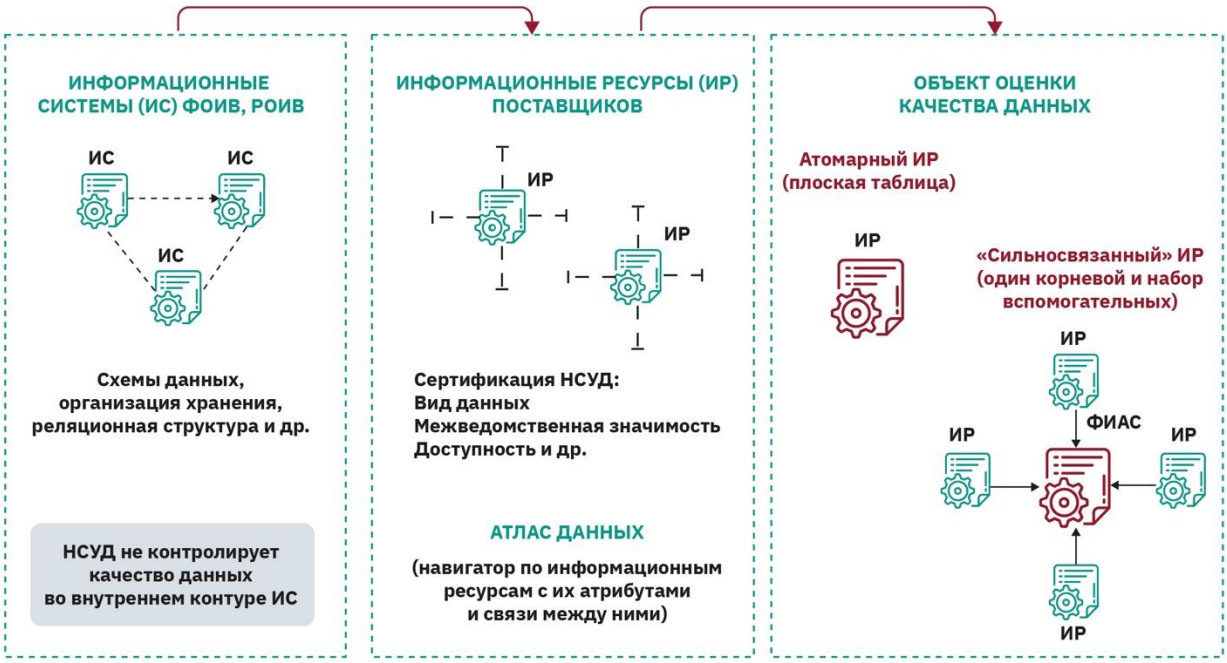
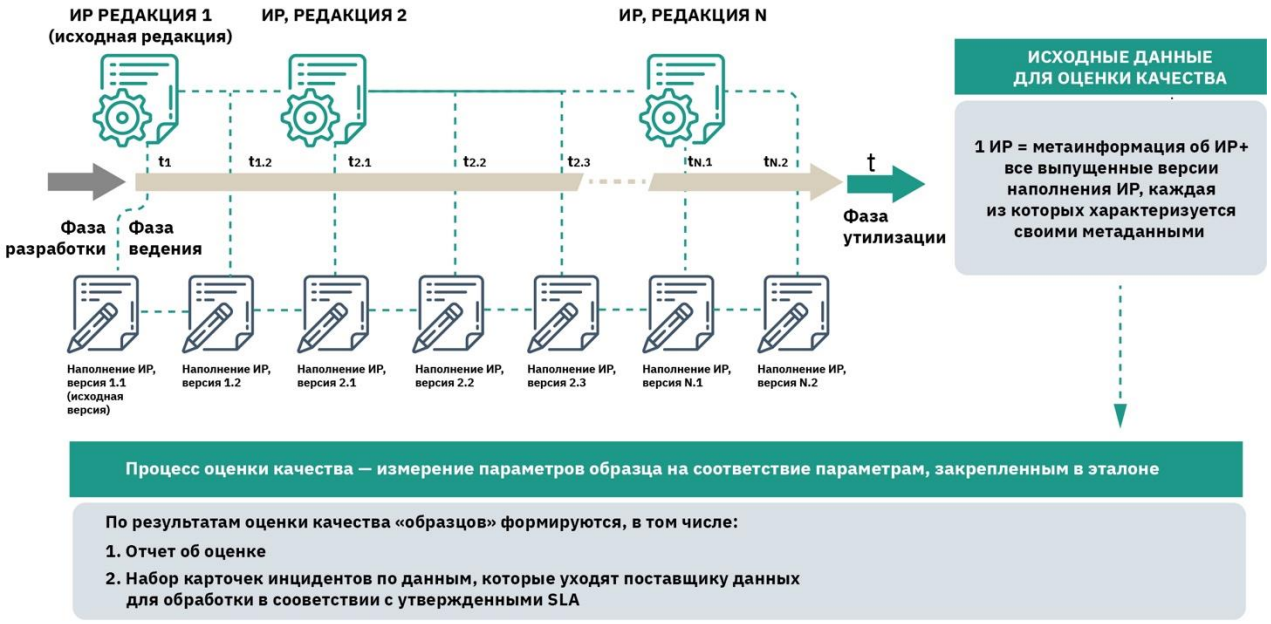


РИСУНОК 30
ФАЗЫ ВЕДЕНИЯ ИНФОРМАЦИОННЫХ РЕСУРСОВ (ИР)



ПАРАМЕТРЫ КАЧЕСТВА ДАННЫХ В НСУД

В НСУД выделены 7 параметров качества с возможностью их измерения, разработана система подсчета и метрики, позволяющие провести оценку качества данных (табл. 6).

ТАБЛИЦА 6
ПАРАМЕТРЫ КАЧЕСТВА НСУД И ИХ ПОКАЗАТЕЛИ

ПАРАМЕТРЫ КАЧЕСТВА И ИХ ОПРЕДЕЛЕНИЯ	ПОКАЗАТЕЛИ
ЦЕЛОСТНОСТЬ аспект оценивания качества данных, характеризующий степень корректности связей между данными (записями и иными агрегатами данных).	– доля обнаруживаемых стандартными средствами системы управления базами данных нарушений целостности; – отношение количества идентифицируемых записей в массиве данных и максимальному количеству неповторяющихся автоматически генерируемых значений ключей.
ПОКРЫТИЕ (ОХВАТ) аспект оценивания качества данных, характеризующий возможность получения из массива данных записи данных о произвольно выбранном объекте предметной области. Для оценки параметров качества данных необходимо внедрение показателей, согласно которым можно будет определить качество данных.	– доля объектов учета, данные о которых включены в информационный ресурс, в общей массе объектов учета; – доля территории, сведения о которой содержаться в информационном ресурсе, в общей площади территории, выступающей в качестве объекта учета; – соотношение интервала времени наблюдений, результаты которого зафиксированы в информационном ресурсе, к общей протяженности интервала времени наблюдений или времени существования интересующих объектов учета.
ПОЛНОТА аспект оценивания качества данных, характеризующий уровень заполнения необходимых атрибутов в записях данных или массиве данных. Для оценки параметров качества данных необходимо внедрение показателя, согласно которому можно будет определить качество данных	– доля незаполненных (нулевых) значений в атрибутах, обязательных к заполнению, в общем массиве данных ИР (показатель может быть отнесен к конкретному записи, к конкретному атрибуту, к произвольной записи или произвольному атрибуту в массиве или наборе данных, к массиву или набору данных в целом).
ТОЧНОСТЬ аспект оценивания качества данных, характеризующий степень соответствия данных: 1) целям использования; 2) данным, полученным из авторитетного источника, или данным, полученным в результате измерения.	– формальная точность; – нормативная точность; – инструментальная точность.

ПАРАМЕТРЫ КАЧЕСТВА И ИХ ОПРЕДЕЛЕНИЯ	ПОКАЗАТЕЛИ
СВОЕВРЕМЕННОСТЬ (АКТУАЛЬНОСТЬ) аспект оценивания качества данных, характеризующий соответствие данных реальному положению дел, пригодность и доступность данных для использования в установленные или произвольные моменты времени.	– периодичность актуализации данных; – доля формально актуальных данных/записей данных в составе массива данных; – вероятность получения актуальных данных в произвольный момент времени; – коэффициент готовности информационного ресурса/массива данных (отношение времени пребывания в актуальном состоянии к времени пребывания в неактуальном состоянии); – длительность подготовки и согласования данных; – среднее время ожидания обслуживания/средняя продолжительность операции синхронизации изменений в реплицируемом ресурсе; – вероятность считывания данных, находящихся в неактуальном состоянии.
УНИКАЛЬНОСТЬ (ДУБЛИРОВАНИЕ) аспект оценивания качества данных, характеризующий степень избыточности массива данных.	– доля дублирующих записей информационного ресурса или массива, содержащих сведения об одних и тех же идентифицируемых объектах учета, различающихся значениями единичных атрибутов, не несущих идентифицирующей нагрузки; – доля дублирующих атрибутов, имеющих одинаковую и близкую семантику; – доля не синхронизируемых из единого источника справочников, имеющих общий источник и назначение в общем числе используемых справочников, используемых для замещения значений в информационных ресурсах одного обладателя; – доля записей в группе информационных ресурсов или массивов данных одного обладателя, содержащих противоречивые сведения об одних и тех же идентификационных объектов учета; – доля дублирующих записей в группе информационных ресурсов или массивов данных одного обладателя, содержащих сведения об одних и тех же идентифицируемых объектах учета, незначительно различающихся атрибутивным составом; – доля атрибутов, идентификаторы которых не включены в систему стандартных имен.
СОГЛАСОВАННОСТЬ аспект оценивания качества данных, характеризующий степень соответствия данных установленным образцам и правилам представления и обозначения.	– вероятность использования альтернативного обозначения или сокращения для сущностей, снабженных стандартизированным обозначением; – вероятность использования нестандартных наименований объектов учета.

Этапы обследования качества данных представлены на рис. 31.

РИСУНОК 31
ЭТАПЫ ОБСЛЕДОВАНИЯ КАЧЕСТВА ДАННЫХ

Критерии оценки качества данных для параметра «Покрытие» представлены в табл. 7.

ТАБЛИЦА 7
КРИТЕРИИ ОЦЕНКИ КАЧЕСТВА ДАННЫХ ДЛЯ ПАРАМЕТРА «ПОКРЫТИЕ»

ПОКАЗАТЕЛЬ КАЧЕСТВА ДАННЫХ ДЛЯ ПАРАМЕТРА «ПОКРЫТИЕ»	ЗНАЧЕНИЕ ПОКАЗАТЕЛЯ КАЧЕСТВА	РЕКОМЕНДАЦИИ К ИСПОЛЬЗОВАНИЮ ДАННЫХ
Доля объектов учета, данные о которых включены в информационный ресурс, в общей массе объектов учета	97,9 %	Данные рекомендованы к использованию в межведомственном взаимодействии
Доля территории, сведения о которой содержаться в ИР, к общей площади территории, выступающей в качестве объекта учета	96,5 %	Данные рекомендованы к использованию в межведомственном взаимодействии
Соотношение интервала времени наблюдений, результаты которого зафиксированы в информационном ресурсе, к общей протяженности интервала времени наблюдений или времени существования интересующих объектов учета	40 %	Данные, не рекомендуемые к использованию

КЛЮЧЕВАЯ РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Волков, Д. Незнанов А. Качество данных: от стратегии к практике / Д. Волков, А. Незнанов // Lobanov Logist. – URL: https://www.lobanov-logist.ru/library/all_articles/64180/ (дата обращения 19.10.2021)

2. Национальная система управления данными. Единые требования по управлению государственными данными. Часть 6: Требования по оценке качества государственных данных. Книга 1: Показатели качества государственных данных. Версия 1.1. // Аналитический центр при Правительстве Российской Федерации. – URL: https://digital.ac.gov.ru/upload/iblock/3e6/ч6.кн1.Показатели%20качества%20государственных%20данных.pdf (дата обращения 19.10.2021)

3. ГОСТ Р 56215-2014/ISO/TS 8000-150:2011 Качество данных. Ч. 150. Основные данные. Структура управления качеством = Data quality. Part150. Master data: Quality management framework: национальный стандарт Российской Федерации: издание официальное: утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 11 ноября 2014 г. № 1519-ст: введен впервые: дата введения 2015-07-01 / подготовлен Федеральным бюджетным учреждением «Федеральный центр каталогизации». – Москва: Стандартинформ, 2020. – 21 с.

4. Об утверждении Концепции создания и функционирования национальной системы управления данными и плана мероприятий ("дорожной карты") по созданию национальной системы управления данными на 2019–2021 годы: распоряжение Правительства Российской Федерации от 03.06.2019 № 1189-р.

5. DAMA-DMBOK: свод знаний по управлению данными / [Dama International]; [перевод с английского Г. Агафонова]. – 2-е изд. – Москва: Олимп-Бизнес, 2020. – 828 с.