Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Храмешин Сергей Николаевич
Должность: Научный руководитель
Дата подписания: 20.08.2025 12:36:22
Уникальный программный ключ:
9e604e7f43158d9188ee2ea08e1c2d1c69a0d1b

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

С. В. Рындина

ТЕХНОЛОГИИ АНАЛИЗА БОЛЬШИХ ДАННЫХ (ПРОДВИНУТЫЙ УРОВЕНЬ): УПРАВЛЕНИЕ И РУКОВОДСТВО ДАННЫМИ, ХРАНЕНИЕ И ОБРАБОТКА ДАННЫХ

Учебно-методическое пособие

ПЕНЗА 2023

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Пензенский государственный университет» (ПГУ)

С. В. Рындина

Технологии анализа больших данных (продвинутый уровень): управление и руководство данными, хранение и обработка данных

Учебно-методическое пособие

Пенза
Издательство ПГУ
2023

УДК 004.738.5(075)+339(075)
Р95

Р е ц е н з е н т
кандидат технических наук, доцент
А. А. Масленников

Рындина, Светлана Валентиновна.
Р95 Технологии анализа больших данных (продвинутый уровень): управление и руководство данными, хранение и обработка данных) : учеб.-метод. пособие / С. В. Рындина. – Пенза : Изд-во ПГУ, 2023. – 48 с.

Рассмотрена инфраструктура хранения данных и особенности ее формирования под разные бизнес-задачи с учетом масштаба и отраслевой специфики компании. Материал соответствует программе дисциплины «Технологии анализа больших данных (продвинутый уровень)» и может быть использован при написании выпускной работы магистра.

Издание подготовлено на кафедре «Цифровая экономика» и предназначено для обучающихся по магистерской программе «Цифровая экономика и бизнес-аналитика» направления подготовки 38.04.01 «Экономика».

УДК 004.738.5(075)+339(075)

© Пензенский государственный
университет, 2023

СОДЕРЖАНИЕ

Введение 4
1. Управление данными 5
1.1. Большие данные и их источники 5
1.2. Администрирование данных 8
1.3. Оценка зрелости управления данными в организации 13
1.4. Управление данными: организационные изменения 16
1.5. Управление данными как активом:
цепочка создания ценности 19
2. Хранение и обработка данных 22
2.1. Хранилища данных: архитектура хранилищ,
архитектура слоев данных, структуры данных 22
2.2. Типы хранилищ данных 29
2.3. ETL-приложения и BI-приложения 33
2.4. Альтернативы Data Warehouse: Data Lake и Data LakeHouse 39
3. Лабораторная работа «Проектирование хранилища данных» 44
Список литературы 45

Введение

В соответствии с учебным планом обучающиеся на втором курсе магистерской программы «Цифровая экономика и бизнес-аналитика» направления подготовки 38.04.01 «Экономика» изучают дисциплину «Технологии анализа больших данных (продвинутый уровень)».

Целью учебно-методического пособия является знакомство студентов с инфраструктурой хранения больших данных, принципами формирования структуры хранения и выбором состава компонентов инфраструктуры в соответствии с поставленными задачами.

Также рассмотрены вопросы, связанные с администрированием данных.

Задания для лабораторных работ имеют практическую направленность, позволяют познакомиться с основными компонентами стратегии управления данными.

В пособии использованы материалы учебного курса BI Consult [1] и материалы информационного сайта «Управление данными» [2].

Также автор пособия рекомендует главу из электронной книги «Учебник 4CDTO» авторства Олега Гиацинтова «Управление данными для эффективной цифровой трансформации» [3]. Эта глава в свободном доступе для зарегистрировавшихся на сайте компании DIS Group пользователей

1. Управление данными

1.1. Большие данные и их источники

Данные собирают, хранят и обрабатывают как разработчики цифровых решений (в том числе в составе гибридных продуктов, в которых физическая реализация дополняется цифровой компонентой), так и компании, использующие цифровые технологии в своей деятельности.

Данные – неотъемлемая часть реализации автоматизированных и автоматических бизнес-процессов компании, в которых используются различные информационные системы, цифровые сервисы, технологические решения:

– бухгалтерский учет;

– формирование финансовой и операционной отчетности;

– процессы управления компанией: планирование, мониторинг и контроль;

– управление заказами;

– сервисное обслуживание клиентов и т.п.

В организации есть и бизнес-процессы, генерирующие данные, и бизнес-процессы, потребляющие данные или результат их обработки, и те, у которых совмещается генерация с потреблением на разных этапах процесса.

Управление данными как активом организации подразумевает реализацию процессов по созданию, обеспечению качества, обработке и безопасности данных.

Источники больших данных (Big Data) достаточно разнообразны:

– социальные сети, блоги, СМИ;

– мобильные приложения, интернет-магазины, цифровые платформы, порталы услуг, цифровые сервисы (например, банковские), цифровые экосистемы (Сбер, МТС и т.п.);

– информационные системы компаний: данные транзакции, профили клиентов и т.п.;

– цифровое производство и производственное оборудование, оснащенное датчиками и сенсорами;

– интернет вещей (IoT) и подключенные к нему устройства, в том числе смарт-устройства, носимая электроника;

– метеорологические станции, измерители состава воздуха и водоемов, спутники.

Цифровая поверхность компаний независимо от масштаба их деятельности постоянно растет, доля проникновения цифровых сервисов и платформ в бизнес-архитектуру компании постепенно увеличивается.

Организации по мере роста использования цифровых решений в своей деятельности сталкиваются с проблемой управления цифровым активом, в который превращаются данные из различных информационных систем, сервисов и платформ, используемых компанией.

У каждого бизнеса свои особенности, собственный паттерн использования данных, но есть общие схемы извлечения ценности из данных:

– улучшение клиентского опыта, связанного с продуктами, услугами, сервисом и обслуживанием в компании;

– привлечение и удержание клиентов, своевременное реагирование на изменение предпочтений клиентов и на их ожидания или потребности;

– определение перспективных продуктов, услуг, функций, цифровых решений, которые позволяют бизнесу диверсифицировать свое ценностное предложение, получить конкурентное преимущество.

Программные решения для хранения и обработки больших данных используются сегодня во многих отраслях:

Органы государственного и муниципального управления. Предоставление государственных услуг в цифровом формате, консолидация данных различных государственных структур и работа в соответствии с принципом «одного окна», т.е. минимизация участия заявителей (граждан, юридических лиц и индивидуальных предпринимателей) в процессах сбора различных документов и справок, если таковые можно заменить запросами на подтверждение данных внутри периметра государственных учреждений. Формирование инициатив в здравоохранении, социальной сфере на основе данных об услугах и их получателях.

Банковская отрасль. Наиболее успешна в извлечении выгоды из больших данных: создание финансовых продуктов и услуг под потребности рынка на основе доступных данных, управление клиентами, эффективная оценка объектов финансирования, залогов и т.п. Управление кибербезопасностью банковской деятельности более эффективно при использовании больших данных, а выявление мошенничества аналитика больших данных – один из основных инструментов обнаружения подозрительных активностей и операций.

Промышленность. Благодаря обработке больших данных можно точнее прогнозировать спрос, более качественно обслуживать производственное оборудование, своевременно оказывать сервисную поддержку покупателям своих товаров: владея большими данными легко прогнозировать момент окончания ресурса эксплуатации того или иного компонента оборудования или готового изделия.

Здравоохранение. Большие данные позволяют формировать эффективные протоколы лечения с учетом индивидуальных особенностей пациентов, разрабатывать новые лекарственные препараты, медицинские изделия, а также улучшать диагностику заболеваний и лечение пациентов.

Сельское хозяйство. Благодаря данным с метеодатчиков прогноз погоды стал более точным, что помогает сельскохозяйственным предприятиям оперативно реагировать на ситуацию. Аналитика больших данных помогает подбирать культуры, дающие наибольший урожай с учетом климатических и почвенных особенностей, подбирать условия содержания скота (режим, корма и т.п.) для улучшения целевых показателей.

Городская инфраструктура. Вывоз мусора, управление дорожно-транспортной сетью, управление ресурсоснабжающими сетями и т.п. на основе больших данных позволяет уменьшить потери, увеличить отдачу, повысить качество обслуживания.

Логистика. Управление цепочками поставок и оптимизация маршрутов доставки на основе больших данных позволяет увеличить объем перевозок, уменьшить простои, снизить аварийность и т.п.

Ретейл. На основе больших данных происходит управление ассортиментом (в ассортиментной матрице остается то, что найдет своего покупателя), управление закупками, что позволяет снизить просрочку, списания, ликвидацию нераспроданных остатков, а также уменьшить затраты на хранение излишков.

Строительство. Управление застройкой на основе данных, прогнозирование необходимых рынку объектов недвижимости (жилая недвижимость, коммерческая, какого формата и т.п.).

Большие данные имеют следующие характеристики:

– большой *объем (volume)* данных, передаваемых из многих источников;

– *разнообразие (variety)* типов данных;

– *скорость (velocity)*, с которой в информационных средах появляются новые данные;

– *достоверность (veracity)*, насколько данным можно доверять;

– *ценность (value),* которую данные потенциально представляют для компании;

– *изменчивость (variability).*

Первые три характеристики были использованы при описании больших данных в 2001 г. Дугом Лейни – аналитиком консалтинговой фирмы Meta Group Inc. А в дальнейшем описание уточнялось и детализировалось.

Технологии больших данных обеспечивают возможность:

– обрабатывать массивы данных, хранящихся распределено из-за их масштабов;

– работать с потоковыми данными, которые образуются в информационных системах с большой скоростью в режиме реального времени (данных становится все больше с каждым тактом времени);

– работать со структурированными и плохо структурированными данными параллельно в разных аспектах.

Данные в организациях традиционно определяются как ресурс, и приоритетным направлением в его обслуживании является не увеличение отдачи от данных, а снижение затрат на сбор, ввод, хранение, обработку и т.п. Системное извлечение выгоды из владения данными при таком отношении к данным в организации не предусматривается.

С изменением фокуса рассмотрения данных с ресурса на актив меняются и приоритеты обслуживания.

1.2. Администрирование данных

Руководство данными (Data Governance) предоставляет принципы, политику, процессы, рамочную структуру, метрики и механизмы надзора для управления данными как активом для руководства деятельностью по управлению данными на всех уровнях [4].

Цель Data Governance – создать в компании возможности для управления данными как активом. Приоритетом становится экономическая выгода от данных:

– повышение ценности данных: за счет непрямой монетизации (снижение производственных затрат, повышение качества продуктов/услуг, расширение клиентской базы и т.п.) и за счет прямой монетизации (продажи данных в виде информационных продуктов);

– снижения рисков, связанных с данными (сокращение издержек, связанных с инцидентами нарушения конфиденциальности, безопасности и несоответствия нормативно-правовым требованиям).

Для этого программа Data Governance должна быть [4]:

– устойчивой (т.е. постоянным процессом, а не разовым проектом);

– встроенной в деятельность;

– измеримой.

Data Governance – это стратегическое управление данными, которое устанавливает процессы и процедуры и определяет лиц или должности, которые отвечают за точность и надежность данных (рис. 1.1).

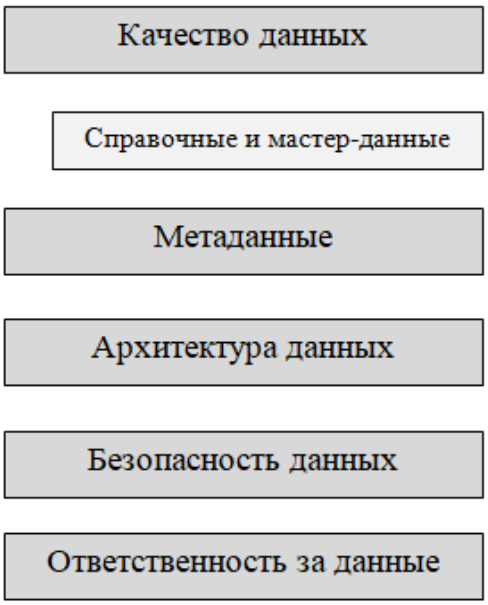

Рис. 1.1. Области Data Governance [5]

Data Governance включает управление качеством данных (Data Quality), а также ответственность за данные, т.е. определение сотрудников, которые осуществляют управление и надзор за процедурами и инструментами, используемыми для обработки, хранения и защиты данных.

Различают Data Governance – как формирование стратегии управления данными, и Data Management – как осуществление непосредственного управления данными согласно определенной стратегии.

То есть Data Management – это прямое управление процессами сбора, хранения и обновления данных.

Джон Лэдли (John Ladley) уточняет разницу между терминами Data Governance и Data Management следующим образом: руководить данными нужно для того, чтобы «делать правильные вещи» (doing the right things), а управлять данными нужно, чтобы «делать вещи правильно» (doing things right) [6].

Разделение функций руководства и управления данными на протяжении жизненного цикла данных и контента можно определить как надзор и исполнение (рис. 1.2).

В контексте управления данными выделяют следующие типы данных [8]:

Метаданные (Metadata) – это данные о данных. Они нужны для понимания и определения, какими данными оперирует предприятие. Метаданные определяют структуры, типы данных, доступы к ним и т.д.

Референс-данные (Reference data) – это относительно редко меняющиеся данные, которые определяют значения конкретных сущностей,

используемых при выполнении операций в рамках всего предприятия. К таким сущностям чаще всего относятся: валюты, страны, единицы измерения, типы договоров/счетов и т.д.

Рис. 1.2. Функции Data Governance и Data Management [7]

Мастер-данные (Master data) – это базовые данные, которые определяют бизнес-сущности, с которыми имеет дело предприятие. К таким бизнес-сущностям обычно относятся (в зависимости от предметной области и отраслевой направленности предприятия) клиенты, поставщики, продукция, услуги, договора, счета, пациенты, граждане и т.п. Кроме информации о той или иной мастер-сущности, в мастер-данные входят взаимосвязи между этими сущностями и иерархии. Обычно мастер-данные воспринимаются как ключевой нематериальный актив предприятия, так как от их качества и полноты зависит эффективность его работы. Мастер-данные, или иначе нормативно-справочная информация, – крайне важная основа управления данными. Большой проблемой становится разная гранулярность собираемых данных: в одних процессах – это минуты, в других дни, в третьих недели и т.д., причем степень агрегации не имеет под собой каких-либо внятных бизнес-оснований: разные именования одних и тех же сущностей, разные идентификаторы для одних и тех же объектов и т.п. Выработать единые правила сбора и хранения данных, единое представление этих данных чрезвычайно важно, если компания не только собирает данные, но и планирует превратить их в ценный актив.

Транзакционные данные (*Transactional data*) – это данные, которые образовались в результате выполнения предприятием каких-либо бизнес-транзакций. Например, для коммерческого предприятия: продажи продуктов и услуг, закупки, поступления/списания денежных средств, поступления на склад и т.п. Транзакционные системы широко используют мастер-данные при выполнении транзакций.

Исторические данные (*Historical data*) – это данные, которые включают в себя исторические транзакционные и мастер-данные. Чаще всего

такие данные аккумулируются в ODS (Operational data store, хранилище оперативных данных) и DWH (Data warehouse, хранилище данных) и служат для решения различных аналитических задач и поддержки принятия управленческих решений.

Взаимосвязь между Data Governance и Data Management можно отразить на схеме взаимодействия (рис. 1.3). Но это деление довольно условное, так как администрирование и управление данными находятся в цикле взаимосвязи друг с другом и непосредственно перетекают одно в другое.

Рис. 1.3. Соотношение между Data Governance и Data Management [9]

Администрирование данных (Data Governance), или руководство данными, включает:

– формирование инфраструктуры и выбор технологий для обеспечения информационных процессов организации и контроля качества информационных активов;

– настройку и поддержку процессов и политик, которые помогают обеспечить управление информационными активами внутри организации, а также надлежащее обращение с данными в процессах ввода, хранения, обработки, доступа и удаления;

– определение лиц (или должностей) в организации, которые имеют полномочия и несут ответственность за обработку и защиту определенных типов данных.

Администрирование данных обеспечивает для данных:

– безопасность и конфиденциальность;
– целостность;
– удобство в использовании,
– гибкость;

– доступность;

– разграничение ролей и обязанностей в обслуживании данных;

– управление внутренними и внешними потоками данных компании в целом.

Преимущества от администрирования данных в организации включают:

- Более низкие затраты на управление данными (Data Management) за счет переиспользования процедур.
- Управление соответствием регуляторным нормам и соблюдение комплаенс-процедур.
- «Прозрачность» любых действий с данными.
- Увеличение ценности данных в организации.
- Эффективность решений прошлых и текущих проблем на основе систем принятия решений, использующих данные.
- Усовершенствованный мониторинг и отслеживание механизмов качества данных и другой деятельности, связанной с данными.
- Общий рост выручки организации за счет монетизации данных.

Данные перестают считаться побочным продуктом, а обеспечение высокого качества данных превращается в цель бизнес-процессов (рис. 1.4).

Фреймворк функций управления данными в привязке к жизненному циклу представлен на рис. 1.5.

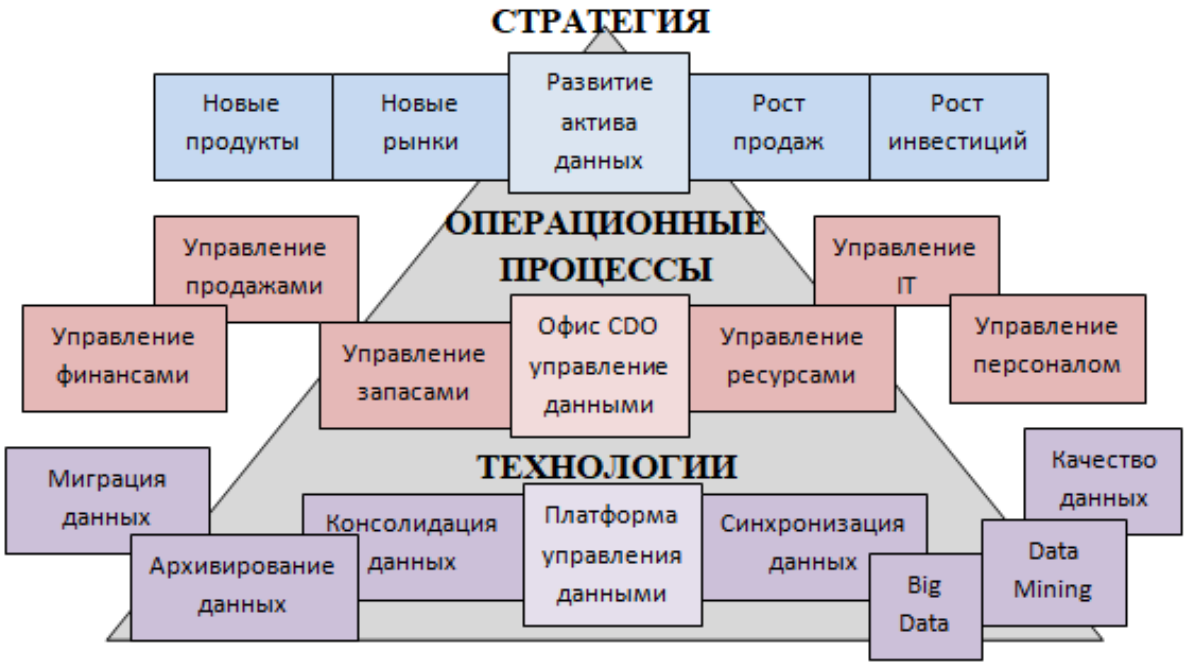

Рис. 1.4. Корпоративная функция управления данными [3]

Рис. 1.5. Фреймворк функций управления данными [10]

1.3. Оценка зрелости управления данными в организации

Управление данными в организации, как и управление процессами, требует оценки текущего уровня зрелости. Данные в том или ином виде для оперативной, финансовой и бухгалтерской отчетности собираются и обрабатываются в любой организации независимо от сферы деятельности и масштаба.

Однако при фокусировке на данных и внедрении подхода к управлению данными как активом необходимо предварительно определить ключевые точки изменений.

Для оценки зрелости необходимо оценить [3]:

– текущее состояние бизнес-процессов в области управления данными;

– наличие политик, правил, регламентов в этом направлении;

– наличие ответственных лиц, курирующих эту тему в целом по компании или по бизнес- или территориальным направлениям;

– наличие процессов мониторинга зависимости между длительностью и успешностью проектов и использованием данных в них;

– наличие программных средств по управлению данными.

Оценка зрелости управления данными может проводиться с опорой на модель процессной зрелости. Например, ГОСТ Р ИСО 13053-1–2015 [11, с. 3] определяет уровни процессной зрелости следующим образом:

– начальный (уровень 1): в организации отсутствует описание процессов;

– управляемый (уровень 2): процесс удовлетворения требований потребителя формальный;

– определенный (уровень 3): определены процессы во всей организации;

– управляемый на основе количественных данных (уровень 4): всеми процессами уровня 3 управляют и выполняют необходимые измерения;

– совершенствуемый (уровень 5): все процессы в организации постоянно улучшают.

Консалтинговая компания International Data Corporation (IDC) предложила модель оценки зрелости цифровой трансформации [12, 13], содержащую пять уровней (рис. 1.6), которую также можно использовать при оценке уровня зрелости компании в управлении данными:

1. Ad hoc (для этого случая, т.е. возникающий по случаю, не на постоянной основе).

2. Opportunistic (оппортунистический, т.е. в русле определенных интересов, удобный, выгодный).

3. Repetable (воспроизводимый).

4. Managed (управляемый).

5. Optimized (оптимальный).

Подробнее про уровни зрелости цифровой трансформации и процессной зрелости [14, с. 5–18].

Описание уровней процессной зрелости управления данными приведем согласно источнику [9].

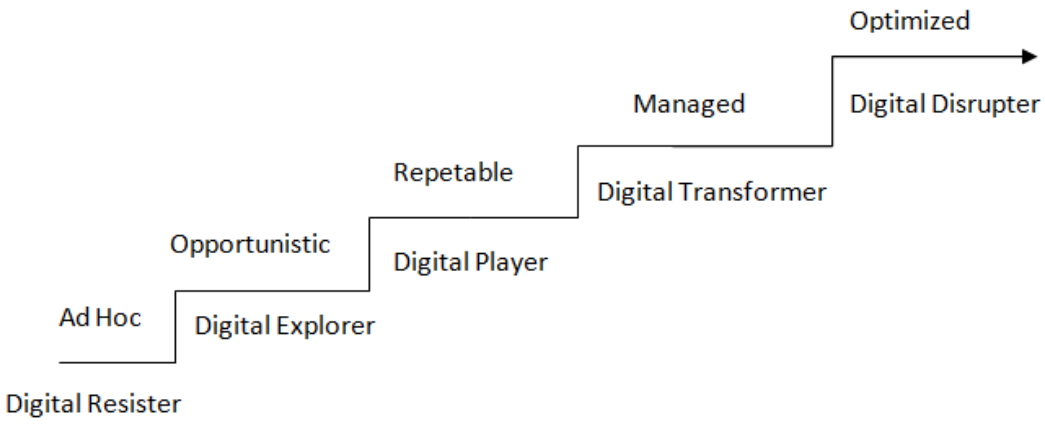

Рис. 1.6. Модель цифровой зрелости IDC

Уровень 1. Начальный (или бессистемный – ad hoc): успех зависит от компетенции отдельных сотрудников:

– слабое или отсутствующее руководство данными;
– набор инструментов крайне ограничен;
– роли определены раздельно по вертикалям;
– механизмы контроля отсутствуют или слабо согласованы;
– проблемы с качеством данных не решаются.

Уровень 2. Повторяемый: присутствует минимальная дисциплина выполнения процессов:

– появляются элементы руководства данными;
– внедрение согласованного набора инструментов;
– определены некоторые роли и процессы;
– есть понимание негативных последствий от проблем с качеством данных.

Уровень 3. Установленный: введены и используются стандарты:

– данные рассматриваются как ключевой фактор эффективной работы;
– масштабируемые процессы и инструменты;
– сокращение процессов, выполняемых вручную;
– результаты процессов, включая управление качеством данных, более прогнозируемые.

Уровень 4. Управляемый: измеряются характеристики процессов, связанных с данными, и осуществляется их контроль:

– централизованное планирование и руководство;
– управление рисками, связанными с данными;
– метрики эффективности управления данными;
– измеримое повышение качества данных.

Уровень 5. Оптимизированный: обеспечена возможность измерять степень достижения целей:

– высокий уровень прогнозируемости процессов;

– контроль над рисками и их минимизация;

– внедрены метрики оценки управления качеством данных и эффективности процессов.

Оценить текущее положение компании в модели зрелости системы управления данными можно на основе Microsoft Cloud Adoption Framework для Azure [15]. По нескольким аспектам управления данными (пользователи, процессы, политики и технологии) выделены четыре уровня: без управления, первая ступень (уровень), вторая ступень (уровень) и полное управление.

Эта модель привлекательна тем, что в отдельных аспектах управления данными компания может находиться на различных уровнях.

1.4. Управление данными: организационные изменения

Управление данными включает определение лиц (или должностей) в организации, которые имеют полномочия и несут ответственность за обработку и защиту определенных типов данных.

Изменения в организационной структуре включают:

1. Вовлечение руководства в управление данными.
2. Определение роли владельцев и распорядителей данных (Data Stewards).
3. Определение функций и обязанностей новых ролей.
4. Назначение ответственных сотрудников.

Состав ролей определяется потребностями конкретной организации и уровнем ее зрелости.

Традиционные роли в работе с данными:

– инженер данных (Data Engineer);

– бизнес-аналитик (Business Analyst);

– специалист по работе с данными (Data Scientist).

Инженер данных решает в компании следующие задачи:

– разрабатывает, тестирует и поддерживает инфраструктуру работы с данными: базы данных, хранилища и системы массовой обработки;

– проводит построение процессов извлечения, загрузки, преобразования, хранения и подготовки данных, включающих консолидацию

данных из различных источников, их предварительную очистку и форматирование, т.е. готовит данные для пользователей (потребителей данных), например бизнес-аналитиков или специалистов по данным.

Бизнес-аналитик решает задачи, связанные с обработкой уже подготовленных данных, но часть операций по преобразованию данных все же находится в его компетенции. Бизнес-аналитик:

– проводит фильтрацию, структурирование данных, избавляется от дубликатов и пропусков, выявляет фиктивные данные и т.п.;

– формирует отчеты с использованием различных визуализаций по данным;

– выдвигает гипотезы и проводит их проверку;

– автоматизирует процессы создания типовых для компании отчетов по данным.

Специалист по данным (исследователь данных):

– создает и обучает предиктивные модели с помощью алгоритмов машинного обучения и нейросетей;

– выявляет скрытые закономерности в бизнес-деятельности;

– способствует извлечению ценности из данных: принимает участие в формировании стратегии компании, принятии решений, транслируя инсайты, полученные благодаря исследованию данных.

Поиск бизнес-гипотез аналитиками (специалистами по данным и бизнес-аналитиками) развивается по двум сценариям:

– реактивному – как результат реакции на возникающие события – запросы пользователей, изменение законодательства;

– проактивному – предположение выстраивается на основе экспертного мнения, приводя к новой гипотезе, которая даст улучшение в бизнесе.

Инженер данных (дата-инженер) – это технический специалист, его основная зона ответственности – аппаратные и программные технологии работы с данными. Бизнес-аналитик и специалист по данным – это высокоуровневые бизнес-потребители данных, отличие в их задачах связано с тем, насколько глубоко требуется исследовать данные для решения бизнес-проблем. Иерархия извлекаемой ценности из данных и уровня требуемых компетенций растет от бизнес-аналитика до специалиста по данным.

Бизнес-аналитик занимается в основном подготовкой оперативной отчетности: описательным и дескриптивным анализом данных для установления текущего состояния процессов и направлений деятельности. Специалист по данным работает с будущим: прогнозирует, какими будут значения ключевых показателей при различных сценариях развития

бизнеса и внешнего окружения, а также выявляет драйверы изменений в нужном компании направлении. От исследователя данных требуется гораздо более интенсивное погружение в предметную область бизнеса, так как он постоянно работает с его сложными контекстами, выходящими далеко за пределы метаданных (описания состава и типов исходных данных). Усилия бизнес-аналитиков и специалистов по данным направлены на создание добавленной ценности в организации с помощью данных.

Для организаций, внедряющих Data Governance, рекомендуют включить в оргструктуру новую роль – распорядителя данных, или управляющего данными (Data Steward).

Функциональные обязанности распорядителя данных в разных организациях могут отличаться. Приведем примерный перечень задач, которые может решать распорядитель данных:

– определение требований к данным на основе запросов подразделения и способы их использования для достижения поставленных целей;

– обеспечение доступа к нужным данным пользователей в заданное время;

– создание источников данных для разных групп пользователей;

– ведение нормативно-справочной информации (НСИ) информационных систем, в том числе проверка и обеспечение качества и актуальности данных;

– определение источников данных и степени доверия к ним;

– обеспечение и гарантия качества собранных, хранящихся и используемых данных;

– контроль инцидентов, связанных с качеством данных;

– выполнение политик и стандартов, установленных программой управления данными;

– помощь инженерам и аналитикам данных в проектировании и реализации процессов и процедур для сбора, хранения, использования и безопасности данных.

Для роли распорядителя данных также существует иерархия. Разумеется, не все ее уровни обязательно присутствуют в организации, которая внедряет Data Governance: определяющим является соотношение масштаба организации и потенциальная польза от данных как актива.

Согласно [7] матрица распорядителей данных представлена следующими ролями:

Главные распорядители данных (Chief Data Stewards) могут возглавлять органы руководства данными вместо CDO или выступать в качестве CDO в виртуальной (основанной на комитетах) или распределенной организационной системе руководства данными.

Исполнительные распорядители данных (Executive Data Stewards) – это старшие руководители, входящие в состав Совета по руководству данными.

Распорядители корпоративных данных (Enterprise Data Stewards) осуществляют надзор (oversight) за отдельными областями (domain) данных предприятия в процессе выполнения всех связанных с этими областями бизнес-функций.

Распорядители бизнес-данных (Business Data Stewards) – это бизнес-специалисты, чаще всего признанные эксперты в той или иной предметной области, ответственные за соответствующее подмножество данных. Они работают с заинтересованными лицами (stakeholders) в части определения и контроля данных.

Владелец данных (Data Owner) – это распорядитель бизнес-данных, который обладает подтвержденными полномочиями на утверждение решений, касающихся его области данных.

Технические распорядители данных (Technical Data Stewards) – это ИТ-специалисты, работающие в одной из областей знаний управления данными. Среди них – специалисты по интеграции данных, администраторы баз данных, специалисты по бизнес-аналитике, аналитики качества данных или администраторы метаданных.

Координирующие распорядители данных (Coordinating Data Stewards) возглавляют и представляют команды распорядителей бизнес-данных и технических распорядителей данных в обсуждениях как на уровне команд, так и с участием исполнительных распорядителей данных. Координирующие распорядители данных особенно важны в крупных организациях.

Перечисленные роли включают функционал бизнес-аналитиков и инженеров данных, обобщая их всех ролью распорядителя, которой на разных уровнях иерархии управления данными соответствует различный объем и содержание должностных обязанностей, а также уровень ответственности за конечный результат превращения данных в ценный актив.

1.5. Управление данными как активом: цепочка создания ценности

При прямой монетизации данных как актива предполагается продажа информационных продуктов внешним потребителям.

Для прямой монетизации данных организация может использовать платформенную бизнес-модель.

Для этого необходимо создать платформу данных и цифровых продуктов, основанных на данных, на которую данные загружаются автоматически, трансформируются в доступную для конечных пользователей (или бизнес-пользователей) форму.

На платформе бизнес-пользователям доступны функции работы с данными без участия технических специалистов: отбор необходимых данных, выдвижение и проверка гипотез, решение бизнес-задач на основе данных.

Такая платформа может обслуживать не только внешних потребителей по запросу, но и внутренних потребителей данных, т.е. можно не разграничивать решения для внутренней и внешней монетизации данных и организовать универсальную платформу данных.

Для платформенного решения можно рассмотреть следующие роли участников [16]:

Производители / издатели данных (data producers/publishers) – это люди и системы, которые данные создают. Данные могут создаваться специально с целью предоставления для использования или генерироваться как побочный продукт какого-либо процесса.

Потребители / пользователи данных (data consumers) – это люди и системы, которые используют данные.

Брокеры данных (data brokers) выполняют функции посредников. Брокеры не производят данные, но предоставляют другим возможность их использования. Они являются дистрибьюторами данных и цифровых продуктов, основанных на данных

Владельцы данных (data owners) отвечают за принятие решений о данных, которые находятся в сфере их влияния. Владелец данных имеет возможность создавать, редактировать, изменять, совместно использовать и ограничивать доступ к данным, обеспечивать их контроль и качество. Владение данными также определяет способность владельца данных назначать, совместно использовать или передавать полномочия по владению данными третьей стороне. Данные нематериальны и этим отличаются от других активов организации. За материальными активами организации закреплены сотрудники, отвечающие за сохранность активов. Но наличие данных в системе не гарантирует, что данным как активу не был нанесен непоправимый урон (они оказались в публичном доступе, нарушена конфиденциальность, потеряно качество данных и т.д.). Кроме того, данные могут создаваться в одних подразделениях, а пользователи данных могут быть как внутренние – из других подразделений, так и внешние по отношению к организации. Для решения всех возникающих вопросов с данными как объектом монетизации и необходим владелец данных.

Цепочка поставок данных представлена на рис. 1.7.

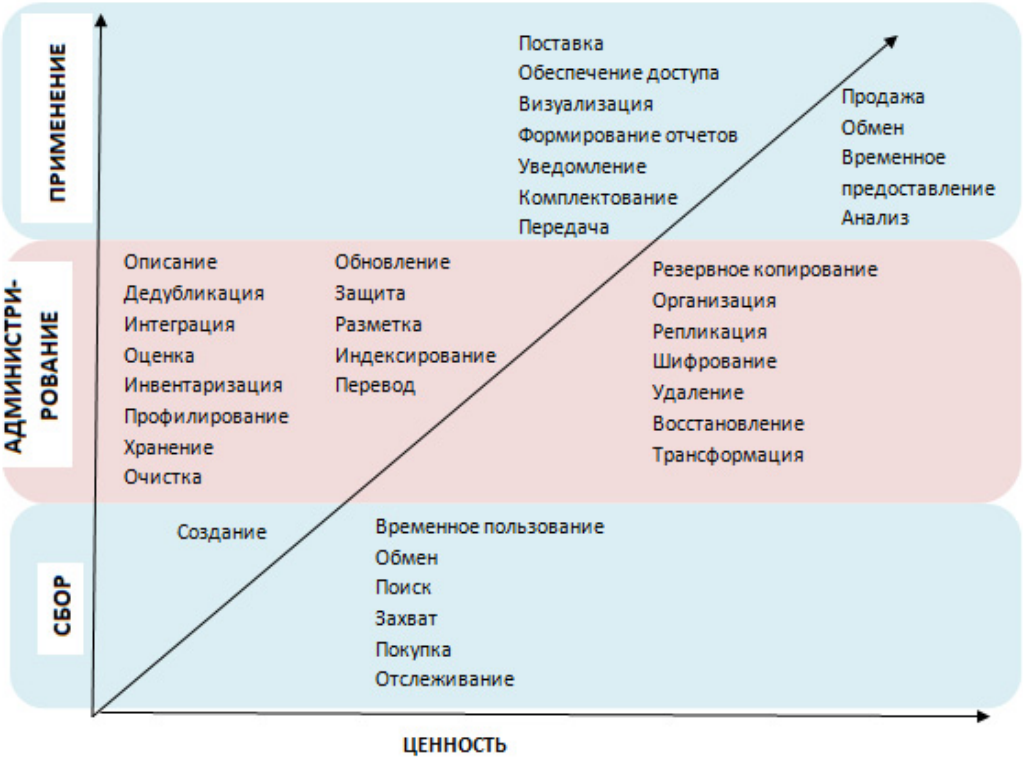

Рис. 1.7. Цепочка поставок данных [16, 17]

Цепочка ценности данных представлена на рис. 1.8.

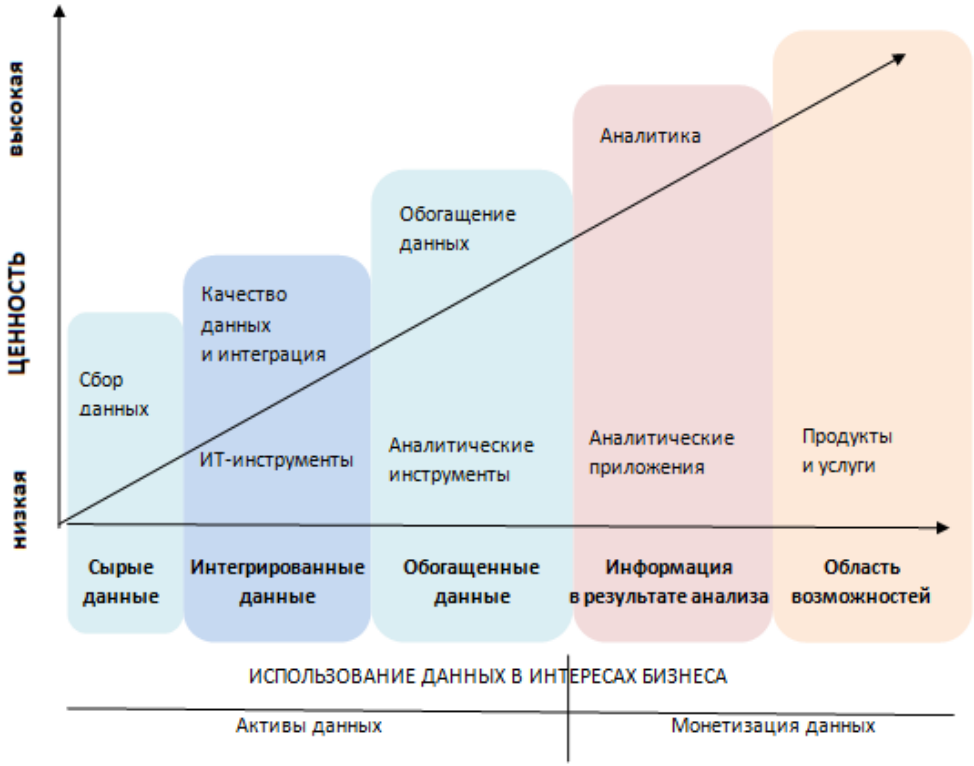

Рис. 1.8. Цепочка ценности данных [16, 18]

2. Хранение и обработка данных

2.1. Хранилища данных: архитектура хранилищ, архитектура слоев данных, структуры данных

Любое хранилище данных – это база данных, соединенная с источниками «сырых» данных при помощи инструментов интеграции данных, с одной стороны, и обладающая аналитическими интерфейсами – с другой. Хранилища данных предоставляют пространства для хранения, а также механизмы для преобразования, перемещения и представления данных конечному пользователю.

Инструменты интеграции данных, т.е. объединение данных из нескольких разрозненных источников для предоставления пользователям единого централизованного представления, а также аналитические интерфейсы будут рассмотрены в разд. 2.3.

Хранилище данных (Data Warehouse, DWH) – это система, в которой собраны данные из различных источников внутри компании (рис. 2.1), и эти данные используются для поддержки принятия управленческих решений.

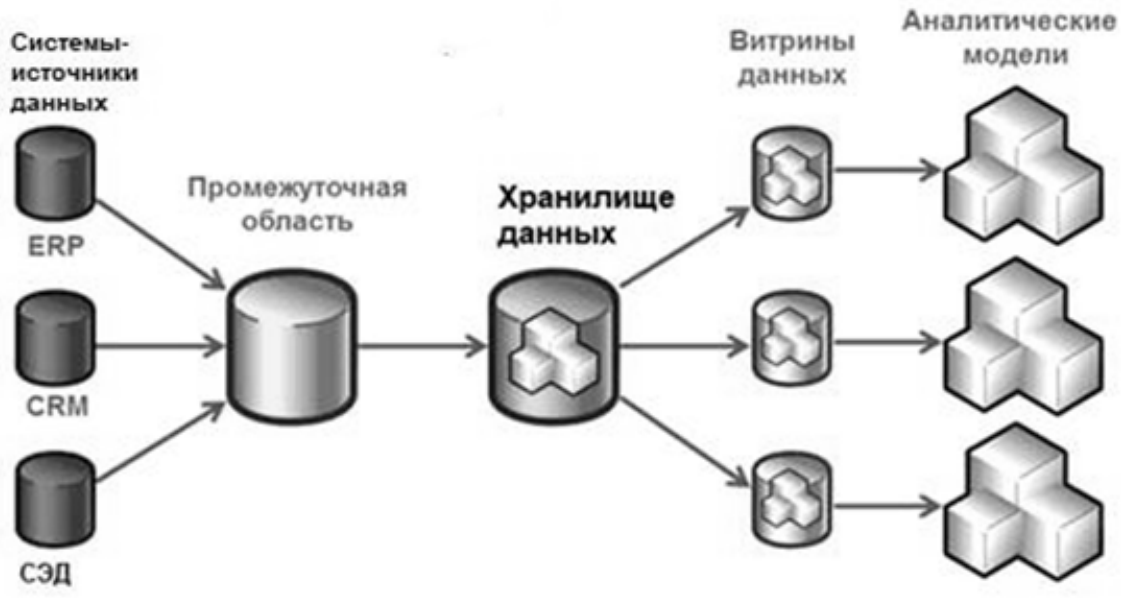

Рис. 2.1. Схема управления данными (ERP – Enterprise Resource Planning, планирование ресурсов предприятия; CRM – Customer Relationship Management, управление взаимоотношениями с клиентами; СЭД – Система Электронного Документооборота) [19]

Витрина данных (Data Mart) предоставляет информацию из корпоративного хранилища (DWH) по конкретной теме или доменной области, например: продажи, финансы или маркетинг. Витрины получают данные из меньшего количества источников, чем хранилища, так как сфокусированы на определенной предметной области. Data Mart также может включать содержимое из внешних источников, помимо DWH.

Промежуточная область или область временного хранения данных (*Staging Area*) предназначена для временного хранения данных, извлеченных из систем-источников; является промежуточным слоем между источниками данных и хранилищем данных. Промежуточная область преобразует данные из многообразных источников данных с различными типами и форматами данных в обобщенный структурированный формат, который проще запрашивать с помощью инструментов анализа и отчетности.

Традиционная архитектура хранилища данных имеет трехуровневую структуру [20]:

Нижний уровень содержит сервер базы данных, используемый для извлечения данных из множества различных источников, например из транзакционных баз данных, используемых для интерфейсных приложений.

Средний уровень содержит сервер OLAP (On-Line Analytical Processing, оперативная аналитическая обработка), который преобразует данные в структуру, лучше подходящую для анализа и сложных запросов. Сервер OLAP может работать двумя способами: либо в качестве расширенной системы управления реляционными базами данных, которая отображает операции над многомерными данными в стандартные реляционные операции (Relational OLAP), либо с использованием многомерной модели OLAP, которая непосредственно реализует многомерные данные и операции.

Верхний уровень – это уровень клиента. Этот уровень содержит инструменты, используемые для высокоуровневого анализа данных, создания отчетов и анализа данных.

Хранилища данных отличаются от OLTP-систем (Online Transaction Processing, онлайн-обработка транзакций), обрабатывающих транзакции в режиме реального времени. С помощью хранилища данных нагрузка на анализ отделяется от нагрузки на транзакции. Хранилища данных в значительной степени являются системами, ориентированными на чтение. Они обеспечивают гораздо больший объем чтения данных по сравнению с записью и обновлением. Это обеспечивает гораздо более высокую производительность анализа.

Отличия операционных баз данных от хранилищ:

- Операционная база данных создается для хорошо известных задач и рабочих нагрузок, таких как поиск определенных записей, индексация и т.д.
- Операционные базы данных поддерживают одновременную обработку нескольких транзакций. Оперативный запрос к базе данных позволяет читать и изменять транзакции.
- Операционная база данных поддерживает текущие данные.
- Хранилище данных поддерживает исторические данные.
- В хранилище данных нет частых обновлений (обычно обновления происходят с некоторой периодичностью, когда загружаются все новые записи за период).
- Наиболее частый запрос к хранилищу – на чтение данных.
- Хранилище данных содержит консолидированные исторические данные.

Таким образом, хранилище данных [21]:

- *служит в качестве всеобъемлющего накопителя данных*, это единый репозиторий для всех корпоративных бизнес-данных, появляющихся в организации;
- *отражает исходные данные*. Хранилище данных получает данные из исходных пространств хранения наподобие Яндекс.Метрики, CRM, IoT-устройств и т.д. И внутри, и вне компании постоянно генерируются новые релевантные данные. Требуется специальная инфраструктура управления этим потоком данных, прежде чем они попадут в хранилище;
- *хранит структурированные данные*. Данные в хранилище стандартизованы и структурированы. Благодаря этому конечные пользователи могут запрашивать их через интерфейсы бизнес-аналитики и формы отчетов;
- *хранит данные, ориентированные на предметную область.* Основная цель хранилища – это бизнес-данные, которые могут относиться к различным предметным областям. Чтобы понять, к чему относятся данные, они структурируются на основе конкретной предметной области, называемой *моделью данных*. Примером предметной области может быть регион продаж или суммарные продажи определенного товара. Кроме того, к данным добавляются метаданные, подробно объясняющие, откуда взялся каждый элемент информации;
- *фиксирует зависимость от времени.* Собираемые данные обычно являются историческими. Большая часть хранимых данных чаще всего разделена по временным периодам.

Через длительные промежутки времени (например, раз в несколько лет) в хранилище могут происходить ревизии, позволяющие избавляться от нерелевантных данных. Но данные, размещенные в хранилище, обычно остаются неизменными: их практически не удаляют, редко проводят с ними какие-либо манипуляции, кроме чтения.

Чтобы структурировать хранилище данных, используются схемы данных «звезда» и «снежинка».

Схема типа «звезда» (star) включает таблицу фактов и ряд денормализованных таблиц измерений. Таблица фактов содержит агрегированные данные, которые будут использоваться для составления отчетов, а таблица измерений описывает хранимые данные. Таблица фактов использует только одну ссылку для присоединения к каждой таблице измерений (рис. 2.2), что значительно упрощает написание сложных запросов.

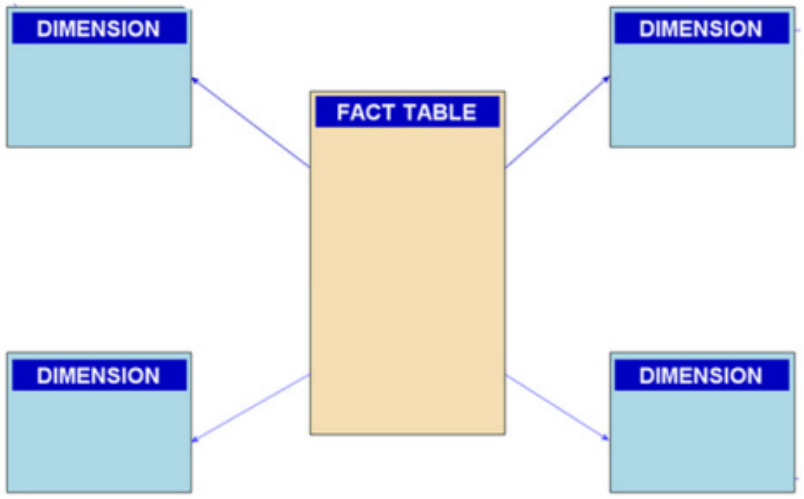

Рис. 2.2. Схема «звезда»

Схема типа «снежинка» (snowflake) использует нормализованные данные. Исходная денормализованная таблица измерений из схемы «звезда» в схеме «снежинка» разветвляется на отдельные нормализованные таблицы измерений, связанные между собой. Таким образом, таблицы измерений в схеме «снежинка» могут быть связаны не только с таблицей фактов, но и с таблицами измерений (рис. 2.3). Например, таблица измерений продукта в схеме «звезда» может быть нормализована в products таблицу (продуктов), product_category таблицу (категорий продуктов) и product_manufacturer таблицу (производителей продуктов) в схеме «снежинка». Схема «снежинки» использует меньше дискового пространства и лучше сохраняет целостность данных. Однако для реализации аналогичных по сложности запросов в схеме «снежинка»

происходит гораздо больше соединений таблиц, чем в схеме «звезда», что сказывается на производительности запросов.

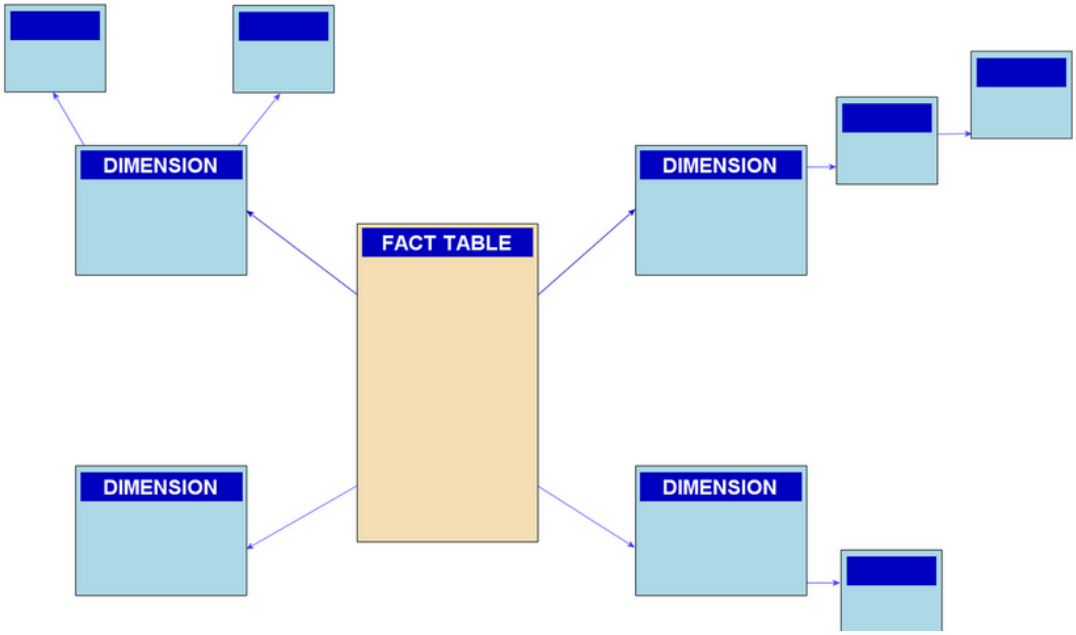

Рис. 2.3. Схема «снежинка»

Классическую архитектуру слоев данных рассмотрим на примере Яндекс Go [22].

Первый слой – RAW. RAW – это не аббревиатура, а название «сырой».

RAW – слой сырых данных, в этот слой загружаются файлы, логи, архивы (рис. 2.4). Форматы могут быть абсолютно различные: tsv, csv, xml, syslog, json и т.д.

Рис. 2.4. Слой RAW [22]

Следующий слой ODS – Operational Data Store (хранилище оперативных данных). Данные в оперативной области полностью соответствуют данным в системах-источниках (рис. 2.5). При загрузке данных производится обогащение служебной информацией, например: системной датой загрузки, идентификацией загрузочного процесса. В слой ODS загружаются данные в формате, приближенном к реляционному. Данные здесь могут являться результатом предобработки данных из RAW перед загрузкой в DDS

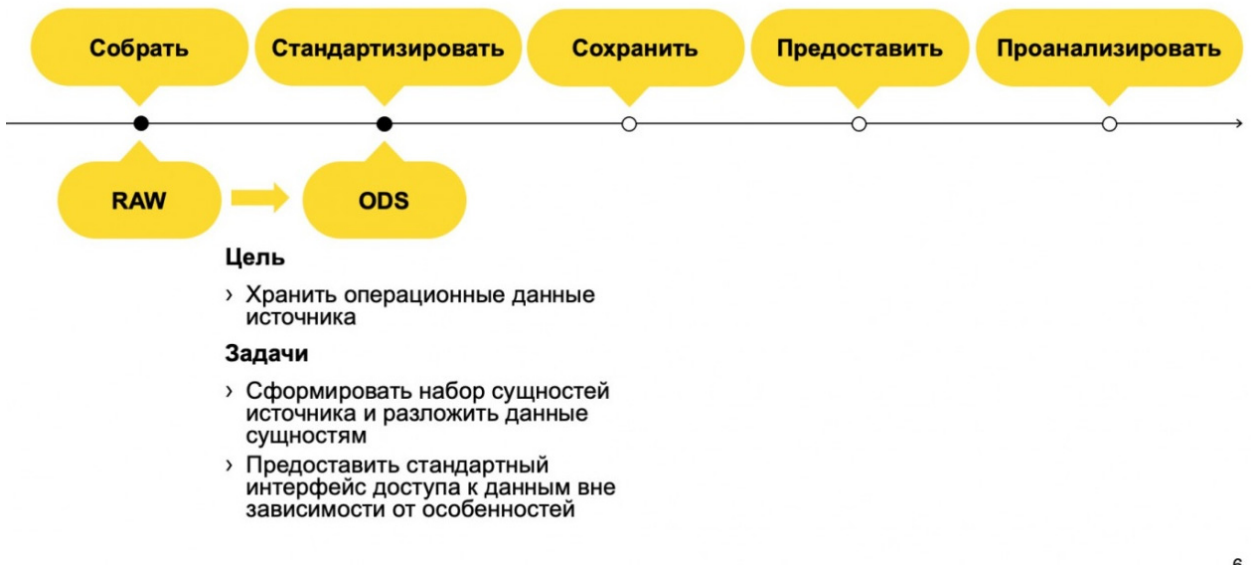

Рис. 2.5. Слой ODS [22]

DDS – Detail Data Store (хранилище детальных данных). Здесь создается консолидированная модель детальных данных. DDS – ядро хранилища.

В этом слое хранятся преобразованные и очищенные детальные данные, полученные из систем-источников, и основные классификаторы.

Данная область содержит следующие типы сущностей:

– справочники и классификаторы;
– сущности, содержащие фактические значения;
– сущности, описывающие связи.

Справочники и классификаторы определяют:

– участников основных бизнес-процессов – клиентов, поставщиков, филиалы, услуги, продукты и т.п.;
– базовые справочники – дата и время, валюта, страны и т.п.;
– прочие справочники, отражающие потребности бизнеса в необходимой аналитике данных, определяющие, в разрезе каких справочников необходимо анализировать фактические данные.

Сущности, содержащие фактические значения, – транзакционные данные из систем источников. Например, информация о совершенных телефонных звонках, выставленных счетах, проводках, проданных товарах и т.п.

Сущности, содержащие связи, определяют взаимосвязи между остальными сущностями, например, клиент-услуга.

Область детальных данных не содержит никаких агрегатов. Только детальные, очищенные и структурированные в соответствии с моделью данные (рис. 2.6).

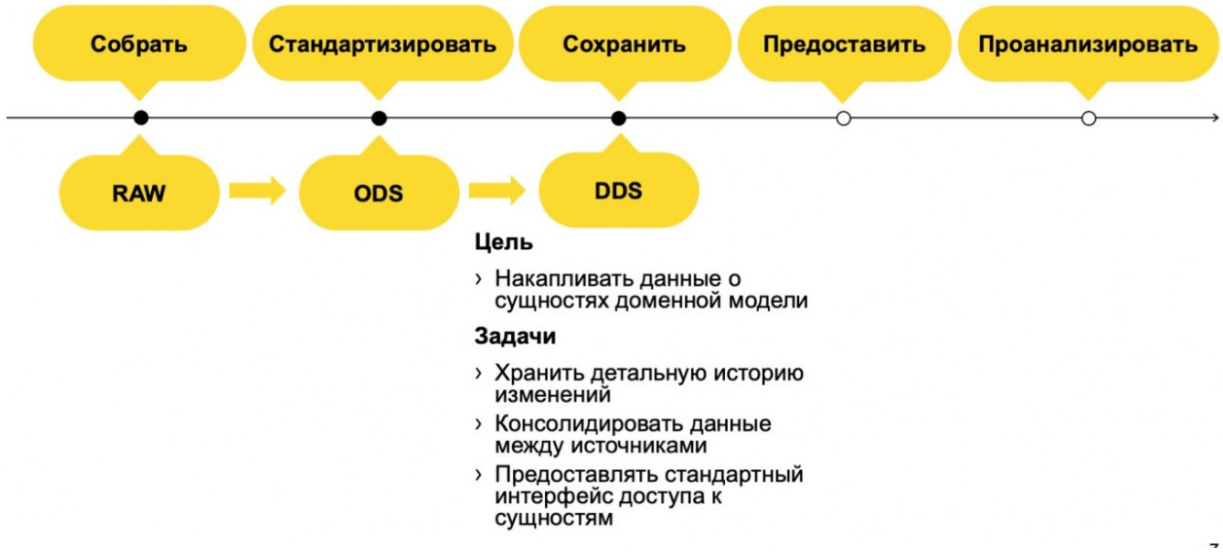

Рис. 2.6. Слой DDS [22]

Следующий слой – CDM (Common Data Marts, общие витрины с данными). CDM – плоские, удобные для аналитиков витрины с оптимизированным доступом на чтение (рис. 2.7).

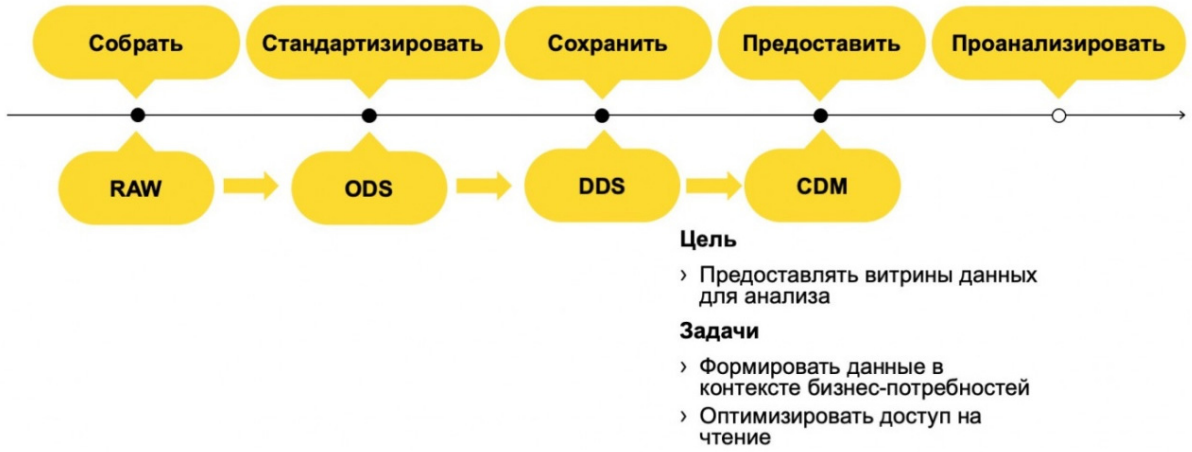

Рис. 2.7. Слой CDM [22]

Слой REP (repository, репозиторий) – это слой абстракции, инкапсулирующий в себе все, что относится к способу хранения данных, чтобы защитить от внешнего вмешательства и неверного использования. Его назначение в отделении бизнес-логики от деталей реализации доступа к данным (рис. 2.8).

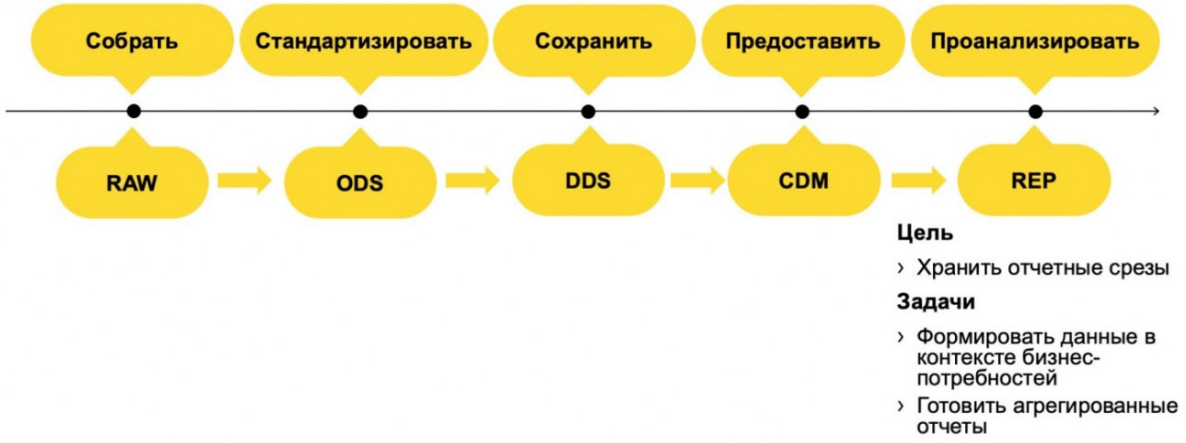

Рис. 2.8. Слой REP [22]

В данной области хранятся агрегаты данных, которые в основном строятся для сущностей, описывающих участников бизнес-процессов. Например, агрегаты строятся для данных по продажам товаров, оказанию услуг, клиентам и т.п. Данные агрегируются в разрезе времени – от часа, дня к неделе, месяцу. Для каждого агрегата может быть определена своя степень агрегации данных.

2.2. Типы хранилищ данных

Хранилище данных внутри компании может быть организовано по-разному [21].

Хранилище данных на внутренних мощностях компании – это локальное выделенное оборудование и программное обеспечение (ПО) компании, использующиеся для единого хранения данных. Хранилище данных можно соединить с источниками данных через API, чтобы постоянно получать информацию и преобразовывать ее в процессе.

При такой реализации хранилища используется дорогостоящая технологическая инфраструктура (и аппаратная, и программная) и требуется команда дата-инженеров для создания и поддержки всей платформы данных. Данные компании обрабатываются безопасным образом, хранилище данных можно преобразовывать в другие стили архитектур платформ данных, изменять их масштаб с сохранением конфиденциальности данных.

Виртуальное хранилище данных – это несколько виртуально соединенных баз данных, к которым можно выполнять запросы как к единой системе. Данные могут оставаться в своих источниках: физически они никуда не перемещаются. Данным, хранящимся в виртуальном хранилище, необходимо ПО для преобразования, чтобы сделать их пригодными для использования конечными пользователями и инструментами отчетности. Обслуживание ПО и оборудования разных баз данных требует финансовых затрат и участия специалистов. Сложные запросы данных могут занимать слишком много времени, если данные, требуемые для запроса, хранятся в разных базах. Виртуальные хранилища данных используются в компаниях, имеющих «сырые» данные в стандартизованном виде, и не нуждающиеся в сложной аналитике.

Облачное хранилище данных – это центральный репозиторий для информации, который размещается в облаке. В этом случае база данных предоставляется как управляемый сервис (managed service) поставщика облачных услуг, оптимизированный для аналитики, масштабирования и удобства использования. Облачные хранилища состоят из слоев вычислений, хранения и клиента (сервиса). Поддержкой инфраструктуры облачного хранилища данных занимается поставщик облачных услуг. Стоимость такого сервиса зависит от объема памяти и вычислительных мощностей, необходимых для выполнения запросов. Требуется контроль со стороны компании за репутацией поставщика услуг: стоит ли выбранный поставщик услуг доверия в вопросе защиты от утечек и взломов, так как бизнес-данные – это уязвимая информация. Облачные хранилища – это решение для тех компаний, которые хотят передать процессы управления данными (реализацию управляемой интеграции данных, обслуживание хранилища и поддержку бизнес-аналитики) на аутсорсинг.

Построение хранилища данных (ХД) на территории заказчика (on premises) – инвестиционно емкий проект. Облачные ХД были призваны удешевить стоимость развертывания хранилища, а также справиться с постоянно растущими объемами исходных данных. Но повсеместного перехода с традиционных ХД на облачные не произошло. Основная причина в том, что возможности передачи данных растут медленнее, чем емкости хранилищ.

Популярным способом увеличения емкости сервера является подключение внешнего хранилища. Существует ряд аппаратных компонентов, программного обеспечения и протоколов, которые в конечном счете придают решениям для хранения данных их особые свойства. Архитектурные решения для подключения внешних хранилищ: файловое хранилище NAS (Network Attached Storage), блочные хранилища SAN (Storage

Area Network) или DAS (Direct Attach Storage), различающиеся типом подключения.

Основные типы хранения данных (рис. 2.9):

- *Файлы* (именованные области данных). Это наиболее популярный тип хранения данных, в котором структура обеспечивает хранение данных, одинаковое для пользователя и для накопителя.
- *Блоки.* Это одинаковые по размеру области, для которых структура данных задается пользователем. Характерной особенностью является оптимизация скорости доступа за счет отсутствия слоя преобразования блоков в файлы.
- *Объекты.* Данные хранятся в плоской файловой структуре в виде объектов с метаданными.

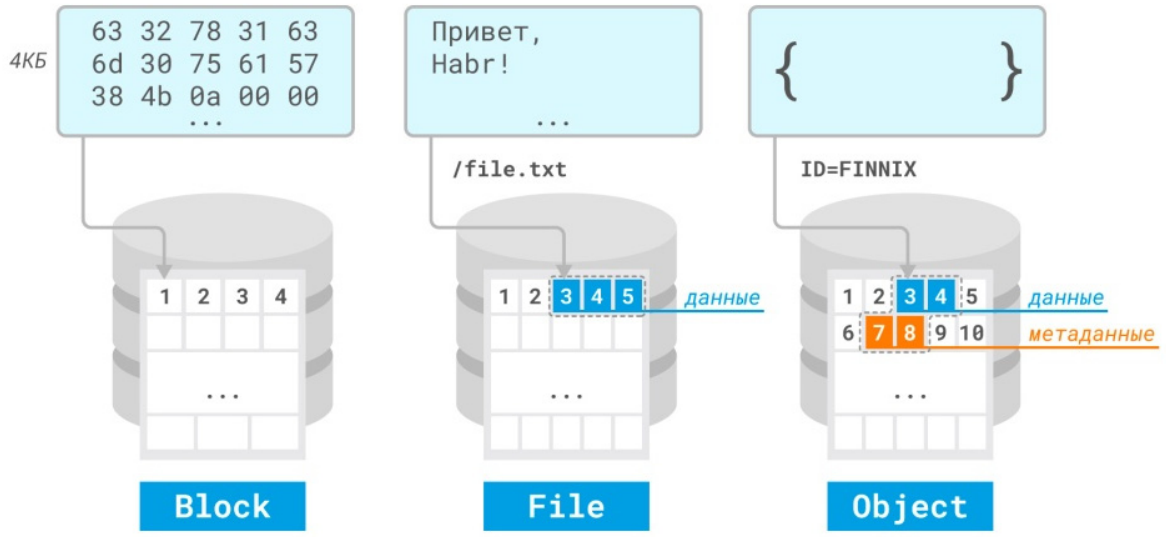

Рис. 2.9. Формы хранения данных [23]

На основе формата хранения данных различают следующие типы хранилищ данных [24]:

- *Блочные хранилища.* Они обладают набором инструментов, которые обеспечивают повышенную производительность: хост-адаптер шины разгружает процессор и освобождает его ресурсы для выполнения других задач. Поэтому блочные системы хранения часто используются для виртуализации, а также хорошо подходят для работы с базами данных. Недостатками блочного хранилища являются высокая стоимость и сложность в управлении, ограниченный объем метаданных. Любую дополнительную информацию приходится обрабатывать на уровне приложений и баз данных.

- *Файловые хранилища.* Достоинством файловых хранилищ является простота: файлу присваивается имя, он получает метаданные, а затем «находит» себе место в каталогах и подкаталогах. Файловые хранилища обычно дешевле по сравнению с блочными системами, а иерархическая топология удобна при обработке небольших объемов данных. Поэтому с их помощью организуются системы совместного использования файлов и системы локального архивирования. Сложности возникают по мере накопления большого количества данных – находить нужную информацию становится затратно. По этой причине файловые системы не используются в дата-центрах, где важна скорость. Как и у блочных хранилищ – ограниченный объем метаданных.
- *Объектные хранилища.* Способны работать с петабайтами информации, так как хорошо масштабируются. Объектные хранилища, благодаря своей возможности работать с растущими объемами данных, стали стандартом для большинства самых популярных сервисов в облаке.

Варианты подключения серверов данных [23]:

- *DAS* – это исторически первый вариант подключения носителей, применяемый до сих пор. Накопитель, с точки зрения компьютера, в котором он установлен, используется монопольно, обращение с накопителем происходит поблочно, обеспечивая максимальную скорость обмена данными с накопителем с минимальными задержками. Также это наиболее дешевый вариант организации системы хранения данных. Для профессиональных целей, а также промышленного применения чаще всего используется несколько накопителей, объединенных в RAID-массив (Redundant Array of Independent Disks, избыточный массив независимых (самостоятельных) дисков) программно либо с помощью аппаратной карты RAID для достижения отказоустойчивости и/или более высокой скорости работы, чем единичный накопитель. RAID – технология виртуализации данных для объединения нескольких физических дисковых устройств в логический модуль для повышения отказоустойчивости и (или) производительности.

Также есть возможность организации кэширования наиболее часто используемых данных на более быстром, но менее емком твердотельном накопителе для достижения и большой емкости, и большой скорости работы дисковой подсистемы компьютера.

- *SAN* – сеть хранения данных, является технологией организации системы хранения данных с использованием выделенной сети. Подключение серверов происходит с использованием специализированного оборудования. Накопители используются в блочном режиме, передача данных через протоколы iSCSI (Internet Small Computer System Interface,

интерфейс малых компьютерных систем интернета) и NVMe (Non-Volatile Memory Host Controller Interface Specification) – интерфейс доступа к твердотельным накопителям, подключенным по шине PCI Express и -oF (NVM Express over Fabrics).

Выбор между блочно-ориентированными хранилищами DAS и SAN основан на их возможностях. DAS может быть очень мощным и часто более дешевым, чем SAN. Однако DAS не может быть легко расширен, так как количество подключенных компьютеров ограничено физическим количеством портов SAS (Serial Attached SCSI, последовательный компьютерный интерфейс, разработанный для подключения различных устройств хранения данных) на DAS. Поэтому многие компании предпочитают выбирать блочные хранилища, подключенные через SAN.

- *NAS* – это сетевое хранилище или сетевой сервер со своей собственной файловой системой. Для передачи данных используются протоколы. NFS (Network File System, сетевая файловая система) – подходит для систем Linux или SMB (Server Message Block, блок сообщений сервера) – используется для Windows. Принципиально базируется на DAS, но ключевым отличием является предоставление общего файлового доступа. Так как работа ведется по сети – сама система хранения может быть сколько угодно далеко от потребителей. Также по сравнению с SAN упрощается работа клиентов, поскольку сервер NAS берет на себя все вопросы по общему доступу.

2.3. ETL-приложения и BI-приложения

Обычно в компании есть несколько специализированных цифровых решений, обслуживающих конкретные бизнес-процессы: маркетинг и продажи обычно имеют дело с CRM-системами, маркетинг дополнительно изучает трафик веб-ресурсов компании с помощью инструментов веб-аналитики (например, Яндекс.Метрики), платежные транзакции привязаны к POS-терминалу, производство обслуживает ERP-система, бухгалтерия готовит всю отчетность в 1С, а на проходной данные собирает СКУД (система контроля удаленного доступа). Состав систем в зависимости от масштаба и отраслевой принадлежности компании может быть иной. Но это не отменяет того факта, что данные в компании часто находятся в различных базах данных и приложениях. Эти системы могут размещаться локально, в облаке, на IoT-устройствах и т.д. Каждая система содержит информацию, связанную с конкретными операциями компании,

и если ничего не предпринять, то ни у одной команды или отдела не будет единого представления о данных.

Для решения проблемы требуется консолидация данных. Консолидация данных, или иначе процесс интеграции данных с использованием технологии ETL, включает в себя объединение данных из разрозненных источников, удаление их избыточности, очистку от любых ошибок и их агрегирование в одном хранилище данных [25].

Общее хранилище данных – это новая система данных для хранения, обработки и отображения копий выбранных данных из различных систем в едином представлении для пользователей.

Хранение данных в одном источнике обеспечивает лучшую целостность данных, т.е. данные в хранилище более точные и непротиворечивые. Благодаря этому пользователи могут выполнять сложные запросы и качественную аналитику. Однако в случае создания единого хранилища данных компании приходится нести дополнительные расходы как на хранение, так и на обслуживание.

Два способа интеграции данных в одно хранилище – ETL и ELT – включают процессы extract (извлечения), transform (преобразования) и load (загрузки). Различие между ETL и ELT в том, в какой последовательности и в какой среде выполняются преобразование и загрузка данных.

При использовании технологии ETL данные преобразуются на временном этапе подготовки до того, как попадут в хранилище данных, в то время как ELT позволяет преобразовывать данные после их загрузки в хранилище данных.

Инструменты ETL используются для интеграции данных, удовлетворяющих требованиям систем управления реляционными базами данных и/или традиционных хранилищ данных с поддержкой OLAP, т.е. данные должны быть структурированы и стандартизированы.

ELT – альтернатива традиционному методу интеграции данных ETL. По сути, ELT меняет местами два последних этапа процесса ETL, т.е. после извлечения из баз данных данные загружаются напрямую в центральный репозиторий, где происходят все преобразования. Промежуточная область, в которой происходит преобразование данных в процессе ETL, отсутствует.

Такая методика стала возможной благодаря современным технологиям, позволяющим хранить и обрабатывать огромные объемы данных в любом формате. К таким технологиям относится, например, Apache Hadoop – это open source программное обеспечение, созданное для непрерывного получения данных из различных источников вне зависимости от их типа. Облачные хранилища данных также поддерживают ELT, поскольку они разделяют ресурсы хранения и вычислений, а также обладают высокой масштабируемостью.

Ключевые этапы процессов ETL и ELT [26]: извлечение, преобразование и загрузка.

Извлечение (Extract) – первый этап интеграции данных. Извлечение и копирование данных из различных источников систем ERP и CRM, баз данных SQL и NoSQL, приложений SaaS, веб-страниц, неструктурированных файлов, электронных писем, мобильных приложений и т.д. Данные извлекаются одним из трех способов:

– полное извлечение применяется к системам, которые не могут отличать новые или измененные записи. В таких случаях единственным способом получения данных из системы является извлечение всех записей – как старых, так и новых;

– частичное извлечение с уведомлениями об обновлениях – самый удобный способ извлечения данных из систем-источников. Он возможен, если в системах есть уведомления об изменениях записей, благодаря чему отсутствует необходимость загрузки всех данных;

– инкрементное извлечение или частичное извлечение без уведомлений об обновлениях – это способ извлечения только тех записей, которые были изменены.

При использовании метода ETL нужно заранее планировать, какие элементы данных необходимо извлекать для дальнейшего преобразования и загрузки. ELT позволяет извлекать все данные мгновенно, а позже принимать решение, какие данные нужно преобразовывать и анализировать.

Преобразование (Transform) – это второй этап в ETL и третий этап в ELT. В фазу преобразования входит последовательность действий, нацеленных на подготовку данных для изменения их под параметры другой системы или для достижения нужного результата. Преобразования могут включать в себя следующие действия:

– сортировку и фильтрацию данных для избавления от нерелевантных элементов;

– устранение дубликатов и очистку;

– удаление или шифрование для защиты уязвимой информации;

– соединение или разделение таблиц и т.д.

В ETL все эти операции происходят вне целевой системы, на этапе подготовки. Все преобразования могут происходить только один раз, из-за чего ETL оказывается довольно негибким. Метод ELT, напротив, гибок и удобен для преобразований, поскольку данные передаются напрямую в хранилище данных, где они могут валидироваться, структурироваться и преобразовываться различным образом и в любой момент.

Загрузка (Load) – это второй этап в ELT и третий этап в ETL. Этот этап заключается в загрузке данных в целевую систему хранения данных,

чтобы к ним могли получать доступ пользователи. Поток процесса ETL подразумевает импорт ранее извлеченных и уже подготовленных данных из промежуточной области (базы данных) в целевое хранилище данных. ELT направляет массив «сырых» данных непосредственно в целевое место хранения, минуя промежуточный уровень. Это сильно экономит время цикла «извлечение – доставка». Как и в случае с извлечением, данные могут загружаться или полностью, или частично.

Процессы интеграции данных сталкиваются с проблемами в данных двух видов:

– случайные ошибки, возникшие на уровне ввода, переноса данных, или из-за багов;

– различия в справочниках и детализации данных между смежными ИТ-системами. Контролируемые различия в структуре данных, это последствия оптимизации структуры данных под цели конкретной системы.

Задачи, которые решаются в процессах ELT, на этапе преобразования данных [27]:

– привести все данные к единой системе значений и детализации, попутно обеспечив их качество и надежность;

– обеспечить аудиторский след при преобразовании (Transform) данных, чтобы после преобразования можно было понять, из каких именно исходных данных и сумм собралась каждая строчка преобразованных данных.

Для решения этих задач преобразование данных в процессе ETL включает валидацию данных, т.е. выявление ошибок и пробелов в данных (табл. 2.1) и маппинг данных (data mapping), т.е. составление схемы обмена данными.

Аналитические интерфейсы, инструменты построения аналитических моделей – это слой извлечения ценности из данных построенного хранилища.

BI-приложения (business intelligence, бизнес-аналитика) и ML-платформы (machine learning, машинное обучение) обслуживают аналитические процессы в компании. Различие между ними заключается в том, какие виды аналитики в компании они поддерживают.

Business intelligence – это методика анализа данных, описывающая текущие показатели операционной деятельности и подводящая итоги деятельности по историческим данным. Сводные отчетные показатели, динамика изменения из прошлого в настоящее, визуализация зависимостей между факторами деятельности и т.п. – это основной функционал BI-приложений. Этот тип обработки данных также называется дескриптивной аналитикой. При помощи дескриптивной аналитики компании

могут исследовать состояние рынка в своей отрасли, а также собственные внутренние процессы. Анализ исторических данных помогает выявлять проблемы и потенциал бизнеса.

Таблица 2.1

Типичные проблемы с данными на этапе преобразования в процессе ETL [27]

Тип данных	Внутри поля	По отношению к другим полям	Совместимость форматов при передаче между системами
Перечисление и текст	Не из списка разрешенных значений. Отсутствие обязательных значений. Несоответствие формату	Не из списка разрешенных значений для связанного элемента. Отсутствие обязательных элементов для связанного элемента. Несоответствие формату для связанного элемента	Символы, допустимые в одном формате, недопустимы в другом. Кодировка. Обратная совместимость (элемент справочника был изменен в целевой системе без добавления маппинга). Новые значения (нет маппинга). Устаревшие значения (не из списка разрешенных в целевой системе)
Числа и порядки	Не число. Не в границах разрешенного интервала значений. Пропущено порядковое значение	Не выполняется отношение между полями (например: НДС и Выручка). Элементу присвоен неправильный порядковый номер. Разница за счет разных правил округления значений	Переполнение. Потеря точности и знаков. Несовместимость форматов при конвертации не в число
Даты и периоды		День недели не соответствует дате. Сумма единиц времени не соответствует из-за разницы рабочие/нерабочие/праздничные/сокращенные дни	Несовместимость формата даты при передаче текстом. Ошибка точки отсчета и точности при передаче числом

ML-приложения строят предикативные модели, которые относятся к прогнозной аналитике, занимающейся предсказаниями на основании обработки данных событий прошлого. Прогнозная аналитика не проводит анализ исторических событий, а делает прогнозы о бизнес-тенденциях будущего.

То есть данные хранилища используются этими приложениями по-разному: BI-приложения делают анализ настоящего, наводят фокус на текущую деятельность компании, а ML-приложения позволяют заглянуть в будущее. Это более продвинутый вид аналитики.

Третий вид аналитики – предписывающая аналитика. Она нацелена на поиск решений бизнес-задач и предлагает по итогу анализа программу действий. Этот вид аналитики реализуется при сочетании многих инструментов: BI-приложений, ML-приложений и самостоятельно разрабатываемых моделей и алгоритмов анализа. Третий вид аналитики скорее не про доступность инструментов анализа, а про мастерство и углубленность в предметную область специалистов по данным (исследователей данных).

Инструменты BI в основном дифференцируются по обслуживанию потребностей в конкретной отрасли. Нужно определить, какой тип отчетности требуется компании для удобного извлечения ценной информации, можно выбрать визуальное или текстовое представление.

Традиционно BI-система проектировалась исключительно для нужд руководства. При востребованности бизнес-аналитики только на уровне управления количество пользователей и типов данных ограничено, необходимость в полной автоматизации отсутствует и требуется технический персонал в качестве посредника между инструментом отчетности и конечным пользователем (руководством).

Современные компании и поставщики решений используют BI-приложения с самообслуживанием. Такой подход позволяет бизнес-пользователям, которых в организации, рассматривающей данные как ценный актив, становится все больше, и руководству получать отчеты, автоматически генерируемые системой. Автоматизированные решения требовательны к формату используемых данных и рассчитаны на стандартный подход к отчетности, чем ограничивают гибкость в настройке. Однако у такого подхода есть преимущество: для работы с системой не нужно постоянное участие технического персонала. Не обладающие техническими знаниями пользователи смогут создать отсчет самостоятельно или получить доступ к выделенному разделу хранилища данных.

2.4. Альтернативы Data Warehouse: Data Lake и Data LakeHouse

Хранилище данных и озеро данных – это две связанные, но в корне различные технологии. В хранилище данных содержатся структурированные данные. Озеро данных (Data Lake) позволяет хранить данные в любом формате: структурированные данные из реляционных баз данных (строки и колонки), полуструктурированные данные (csv, файлы логов, xml, json), неструктурированные данные (почтовые сообщения, документы, pdf) и даже бинарные данные (видео, аудио, графические файлы).

Рассмотрим архитектурные особенности озера данных в сравнении с хранилищами данных [29].

Источники данных. Озера и хранилища данных могут использовать неограниченное количество источников данных. Однако перед размещением данных в хранилище нужно спроектировать схему данных. В хранилище можно загружать только структурированные данные. В озерах данных могут храниться неструктурированные и полуструктурированные данные, например журналы веб-серверов, данные навигации, социальных сетей и показания датчиков.

Предварительная обработка. Перед размещением данных в хранилище обычно требуется их предварительная обработка. Инструменты ETL используются для предварительной очистки, фильтрации и структурирования наборов данных для размещения их в хранилище. Озера данных хранят любые данные. Озера данных демонстрируют гибкость благодаря возможности выбора: проводить предварительную обработку или нет. С озерами данных используются инструменты ELT, которые сразу загружают данные в озеро, а преобразуют их только по требованию.

Качество данных. Хранилище данных более надежно, потому что обработка данных происходит до их размещения в нем, т.е. заранее применяются функции, такие как сортировка, суммирование и проверка, чтобы гарантировать точность данных. В озеро данных могут быть переданы на хранение ошибочные и непроверенные данные, данные с дубликатами и т.п.

Производительность. Хранилище данных оптимизировано для обеспечения наивысшей скорости выполнения запросов. Бизнес-пользователи предпочитают хранилища данных, чтобы более эффективно создавать отчеты. Архитектура озера данных отдает больший приоритет объему хранилища и снижению затрат, чем производительности. Озера данных позволяют получить намного больший объем хранилища по меньшей цене и при этом использовать данные с достаточно приемлемой скоростью.

Сравнение озер данных и хранилищ данных по основным параметрам их функционирования представлено в табл. 2.2.

Таблица 2.2

Сравнительные характеристики для озер данных и хранилищ данных [29]

Характеристика	Хранилище данных	Озеро данных
Данные	Реляционные данные из транзакционных систем, операционных баз данных и групп бизнес-приложений	Любые данные, в том числе структурированные, частично структурированные или неструктурированные
Схема	Часто разрабатывается еще до реализации хранилища данных, но может создаваться и на этапе анализа (схема при записи или схема при чтении)	Записывается в момент анализа (схема при чтении)
Соотношение цены и производительности	Самая высокая скорость выполнения запросов за счет использования локального хранилища	Ускоренное выполнение запросов за счет использования недорогого хранилища и разделения процессов вычисления и хранения
Качество данных	Данные, прошедшие строгую проверку для использования в качестве главной и достоверной версии	Любые данные, проверенные и непроверенные (необработанные данные)
Пользователи	Бизнес-аналитики, разработчики и специалисты по работе с данными	Бизнес-аналитики (используют проверенные данные), специалисты по работе с данными, разработчики и инженеры, создатели архитектур данных
Аналитика	Пакетные отчеты, бизнес-аналитика и визуализация	Машинное обучение, исследовательская аналитика, изучение данных, потоковая передача, операционная аналитика, большие данные, профилирование

Параллельно с хранилищем данных в компании может быть и озеро данных для хранения неиспользуемых в аналитике данных на случай

проверки или в расчете на потенциальную ценность этих данных в будущем, или для исполнения законодательных требований.

Из-за меньшей степени защищенности данных и отсутствия предварительного преобразования данных содержание озера данных обходится дешевле, озера данных более гибкие и легче масштабируются.

Озера часто выполняют вспомогательную роль в подготовке аналитики. Часть компаний использует озера данных по прямому назначению – для продвинутой и ML-аналитики. Остальные – как область для временного хранения копии исходных данных перед их ETL-обработкой или как расширение хранилища данных.

Однако неверно выстроенные процессы сбора данных могут превратить озеро в бесполезное «болото» данных (Data swamp), в котором данные есть, но инструменты для извлечения данных или аналитики работают в нем настолько медленно, что результат обработки данных перестает быть актуальным для организации.

Необходимо заранее определить, для чего именно собираются данные, и выбрать подходящий под цели способ организации озера данных [30]:

– использование Hadoop (HDFS);
– организация собственного S3-хранилища;
– использование готового решения от провайдера.

Использование Hadoop. Hadoop – это программная среда, предназначенная для вычислений Big Data, используется для хранения и обработки любых типов данных (структурированных, слабо структурированных и неструктурированных). А архитектура платформы основана на трех компонентах: HDFS (Hadoop Distributed File System распределенная файловая система Hadoop для хранения больших файлов с возможностью потокового доступа к информации), MapReduce (модель распределенных вычислений, разработанная Google) и YARN.

Преимущества HDFS:

– большой размер блока по сравнению с другими файловыми системами;
– файлы пишутся однократно, что исключает внесение в них любых произвольных изменений;
– принцип WORM (Write-once and read-many) полностью освобождает систему от блокировок типа «запись-чтение»; запись в файл в одно время доступен только одному процессу, что исключает конфликты множественной записи;
– HDFS оптимизирована под потоковую передачу данных;

– самодиагностика – каждый узел данных через определенные интервалы времени отправляет диагностические сообщения узлу имен, который записывает логи операций.

Ограничения HDFS:

– сервер имен является центральной точкой всего кластера, и его отказ повлечет сбой системы целиком;

– нельзя дописывать или оставлять открытыми для записи файлы в HDFS, за счет этого в классическом дистрибутиве Apache Hadoop невозможно обновлять блоки уже записанных данных;

– нет поддержки реляционных моделей данных.

Развертывание собственного S3-хранилища. Этот вариант требует существенных затрат на создание инфраструктуры, на персонал, который будет разрабатывать, а потом обслуживать хранилище. По мере роста объема данных дополнительные сервера необходимо будет арендовать или докупать.

Готовое облачное хранилище от провайдера. Специалисты компании обслуживают как работу хранилища, так и инфраструктуру под ним. Гибкость облака позволит легко масштабировать кластеры под рост объема данных. Облачное хранилище легче связать с другой инфраструктурой проекта. При выборе провайдера существуют законодательные ограничения. В соответствии с постановлением Правительства РФ № 728 от 26.06.2018 хранение данных пользователей российских компаний допускается только на территории РФ. В законе «О персональных данных» (152-ФЗ) прописаны критерии безопасности хранения данных, в том числе ограничения и даже запрет на передачу персональных данных на территории других государств.

Хранилище данных и озеро данных обслуживают разные аналитические ниши. Однако есть попытка объединить их преимущества в одном решении для хранения данных.

Эволюция хранения данных связывается с новой гибридной архитектурой Data LakeHouse (DLH, хранилище и озеро данных), объединяющих два инструмента подготовки данных для аналитики. Термин data lakehouse условно можно перевести как «хранилище и озеро данных».

Как правило, хранилища и озера работают изолированно друг друга. Появление концепции гибридной архитектуры – это попытка объединить данные, собранные в хранилища и озера, и аналитические выводы, полученные на их основе.

Гибридная архитектура может не только расширить возможности хранилищ в части аналитики неструктурированных данных, но и устранить ограничения озер в обеспечении качества данных.

LakeHouse имеет следующие основные особенности [31]:

– *поддержка транзакций* – конвейеры данных способны одновременно считывать и записывать данные;

– *принудительное применение и управление схемой*, включая поддержку классических моделей DWH, такие как схемы звезды и снежинки, с обеспечением целостности и полноты данных, а также надежные механизмы управления и аудита;

– *совместимость с BI* – LakeHouse позволяет использовать инструменты бизнес-аналитики непосредственно в исходных данных, повышая их актуальность, а также уменьшая задержку и затраты, связанные с необходимостью выполнения операций над двумя копиями данных как в озере данных, так и в хранилище;

– *изоляция хранения от вычислений* по разным кластерам, что облегчает масштабирование для большего количества одновременных пользователей и объемов данных;

– *открытость стандартизованных форматов хранения данных*, таких как Apache Parquet или Iceberg, которые предоставляют API (Application Programming Interface, программный интерфейс приложения), поэтому различные инструменты и механизмы, включая ML-системы и библиотеки Python/R (языки программирования для аналитики данных), могут эффективно обращаться к данным напрямую;

– *многообразие различных типов данных* – от неструктурированных до структурированных: различные типы данных, включая изображения, видео, аудио, JSON-структуры и текст.

– *поддержка разнообразных рабочих нагрузок* – от алгоритмов машинного обучения до SQL-запросов и распределенных вычислений; хотя все эти рабочие нагрузки требуют разных технологий реализации, все они полагаются на один и тот же репозиторий данных;

– *сквозная потоковая передача событий* в режиме реального времени, что устраняет необходимость в отдельных streaming-системах.

В отличие от классических DWH, которые ориентированы на BI-инструменты, парадигму LakeHouse можно назвать AI-центричной (Artificial Intelligence, искусственный интеллект). Потенциально Lake House дает более гибкие и широкие возможности, но из-за этого является менее управляемым по сравнению с Data Warehouse.

3. Лабораторная работа
Проектирование хранилища данных

Цель работы: для отраслевой компании (отрасль выбирается по согласованию с преподавателем) научиться проектировать хранилище данных.

Формируемые знания, умения и навыки: научиться составлять регламентирующую документацию по созданию и функционированию хранилища данных, определять подходящие решения на отечественном рынке и обосновывать целесообразность включения их в инфраструктуру управления данными в компании. Получить навыки формирования регламентирующей документации по хранилищам данных с учетом отраслевой принадлежности компании.

Необходимо:

1. Сформировать первичную документацию по хранилищу данных, витринам, OLAP-кубам для отраслевой компании.
2. Провести сбор данных о бизнес-процессах (БП) компании и метриках.
3. Представить БП в виде укрупненных графических схем (например, BPMN-нотация).
4. Составить перечень источников данных для хранилища.
5. Построить модель данных для хранилища данных.
6. Разработать документацию по витринам данных.
7. Выработать требования к аналитическим OLAP-кубам (перечень измерений, метрик, разграничение прав доступа).
8. Определить состав основных управленческих отчетов.

Контрольные вопросы и задания

1. Какие бизнес-процессы являются ключевыми в компании?
2. Какие источники первичных данных могут эксплуатироваться в отраслевой компании? Какие данные они хранят (предметные области, сущности, атрибуты/характеристики сущностей, связи)?
3. Какая схема использовалась для построения модели данных в хранилище?
4. Какие витрины данных были спроектированы для отраслевой компании?
5. Какие сотрудники должны получить права доступа к витринам данных и построению отчетов по данным? Какие ограничения у предоставленных различным категориям сотрудников прав доступа?
6. Какие управленческие отчеты следует построить для отраслевой компании?

Список литературы

1. Современная архитектура хранилища данных» [BI Consult] : учеб. курс. URL: https://biconsult.ru/services/uchebnyy-kurs-sovremennaya-arhitektura-hranilishcha-dannyh (дата обращения: 10.03.2022).
2. Управление данными. URL: https://data-management.ru/blog/2019/10/data-governance/ (дата обращения: 10.03.2022).
3. Гиацинтов О. Управление данными для эффективной цифровой трансформации // 4CDTO [Электронная книга компании DIS GROUP]. URL: https://dis-group.ru/materials/upravlenie-dannymi-dlya-effektivnoj-czifrovoj-transformaczii/?ysclid=lfzfnmy4hw341968987
4. Управление данными: DAMA DMBOK2. Глава 3. Руководство данными Data Governance [Проект Data Literacy Project по переводу DAMA-DMBOK2]. URL: https://dataliteracy.ru/dmbok
5. Разделяй и властвуй, или Зачем управлять данными [Блог GlowByte]. URL: https://habr.com/ru/company/glowbyte/blog/708626/ (дата обращения: 10.03.2022).
6. Ladley J. Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program : 2nd ed. Academic Press, 2020.
7. DAMA-DMBOK: свод знаний по управлению данными. 2-е изд. М. : Олимп-Бизнес, 2020.
8. Что такое «система управления мастер-данными» и зачем она нужна [Блог «Терминология IT»]. URL: https://habr.com/ru/post/324148/
9. Инструменты Data Governance [Блог «Хранение данных»]. URL: https://habr.com/ru/post/599173/ (дата обращения: 13.03.2023)
10. Управление данными в госсекторе. Навигатор для начинающих / под ред. О. М. Гиацинтова, В. А. Сазонова, М. С. Шклярук. – М. : РАНХиГС, 2022.
11. ГОСТ Р ИСО 13053-1-2015. Статистические методы. Количественные методы улучшения процессов Шесть сигм. Ч. 1. Методология DMAIC (введ. 2016-07-01). М. : Стандартинформ, 2016. 27 с.
12. IDC Benchmark: Digital Maturity Check for Self-Assessment. URL: https://news.sap.com/2015/12/idc-benchmark-digital-maturity-check-for-self-assessment/ (дата обращения: 13.03.2023).
13. Digital Transformation Drives Supply Chain Restructuring Imperative. URL: https://www.opentext.com/file_source/OpenText/en_US/PDF/opentext-idc-digital-transformation-supply-chain.pdf (дата обращения: 03.03.2023).
14. Рындина С. В. Бизнес-процессы цифрового предприятия : учеб.-метод. пособие. Пенза : Изд-во ПГУ, 2020. 80 с.

15. Управление аналитическим пространством [Microsoft Cloud Adoption Framework для Azure]. URL: https://learn.microsoft.com/ru-ru/azure/cloud-adoption-framework/scenarios/cloud-scale-analytics/govern (дата обращения: 13.03.2023).

16. Кузнецов С., Константинов А., Скворцов Н. Ценность ваших данных : практическое руководство. М. : Альпина ПРО, 2022. 574 с. URL: https://znanium.com/catalog/product/1905847 (дата обращения: 13.03.2023).

17. Laney D. B. Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage; Routledge; 1st ed., 2017. (дата обращения: 13.03.2023).

18. Kotorov R. The Data Value Chain: Steps for Monetizing Your Data. Integration Developer News, 2017. URL: https://www.idevnews.com/stories/6998/The-Data-Value-Chain-Steps-for-Monetizing-Your-Data (дата обращения: 13.03.2023).

19. Вичугова А. Каталоги, хранилища и витрины данных: принципы, практики и инструменты проектирования [Школа больших данных]. URL: https://www.bigdataschool.ru/blog/why-you-need-data-catalogs-for-datamarts-and-dwh.html?ysclid=lg10qbbijk872111404 (дата обращения: 13.03.2023).

20. Архитектура хранилищ данных: традиционная и облачная [Блог «Хранилища данных»]. URL: https://habr.com/ru/post/441538/ (дата обращения: 13.03.2023).

21. Enterprise Data Warehouse: EDW Components, Key Concepts, and Architecture Types : офиц. сайт AltexSoft. URL: https://www.altexsoft.com/blog/enterprise-data-warehouse-concepts/ (дата обращения: 13.03.2023).

22. Как мы внедрили свою модель хранения данных – highly Normalized hybrid Model. Доклад Яндекса [Блог компании Яндекс]. URL: https://habr.com/ru/company/yandex/blog/557140/ (дата обращения: 13.03.2023).

23. Хранение данных. Или что такое NAS, SAN и прочие умные сокращения простыми словами [Блог компании Southbridge]. URL: https://habr.com/ru/companies/southbridge/articles/517502/ (дата обращения: 13.03.2023).

24. Как устроены хранилища данных: обзор для новичков [Блог компании 1cloud.ru].URL: https://habr.com/ru/company/1cloud/blog/345154/ (дата обращения: 13.03.2023).

25. Data Integration: Approaches, Techniques, Tools, and Best Practices for Implementation [Официальный сайт AltexSoft]. URL: https://www.altexsoft.com/blog/data-integration/ (дата обращения: 13.03.2023).

26. ETL vs ELT: Key Differences Everyone Must Know : офиц. сайт AltexSoft. URL: https://www.altexsoft.com/blog/etl-vs-elt/ (дата обращения: 13.03.2023).

27. Основные функции ETL-систем [Блог «Хранилища данных»]. URL: https://habr.com/ru/articles/248231/ (дата обращения: 13.03.2023).

28. Complete Guide to Business Intelligence and Analytics: Strategy, Steps, Processes, and Tools : офиц. сайт AltexSoft]. URL: https://www.altexsoft.com/blog/business/complete-guide-to-business-intelligence-and-analytics-strategy-steps-processes-and-tools/ (дата обращения: 13.03.2023).

29. Каковы различия между хранилищем данных, озером данных и банком данных? [Блог по аналитике компании Amazon]. URL: https://aws.amazon.com/ru/compare/the-difference-between-a-data-warehouse-data-lake-and-data-mart/ (дата обращения: 13.03.2023).

30. Как и где хранить большие данные? Технологии хранения big data [Блог по технологиям сообщества Selectel]. URL:https://selectel.ru/blog/big-data-storage/?ysclid=lg7pxt2i4d306246280 (дата обращения: 13.03.2023).

31. Вичугова А. DWH + Data Lake или что такое LakeHouse. [Школа больших данных]. URL: https://www.bigdataschool.ru/blog/what-is-lakehouse.html?ysclid=lg8d7zcir618019512 (дата обращения: 13.03.2023).

Учебное издание

Рындина Светлана Валентиновна

Технологии анализа больших данных (продвинутый уровень): управление и руководство данными, хранение и обработка данных

Редактор *А. Г. Темникова*
Технический редактор *М. Б. Жучкова*
Компьютерная верстка *М. Б. Жучковой*

Подписано в печать 26.09.2023.
Формат 60×84 1/16. Усл. печ. л. 4,42.
Тираж 5. Заказ № 468.

Издательство ПГУ.
440026, г. Пенза, ул. Красная, 40.
Тел.: (8412) 66-60-49, 66-67-77; e-mail: iic@pnzgu.ru

ПензГУ