ХРАНИЛИЩА ДАННЫХ

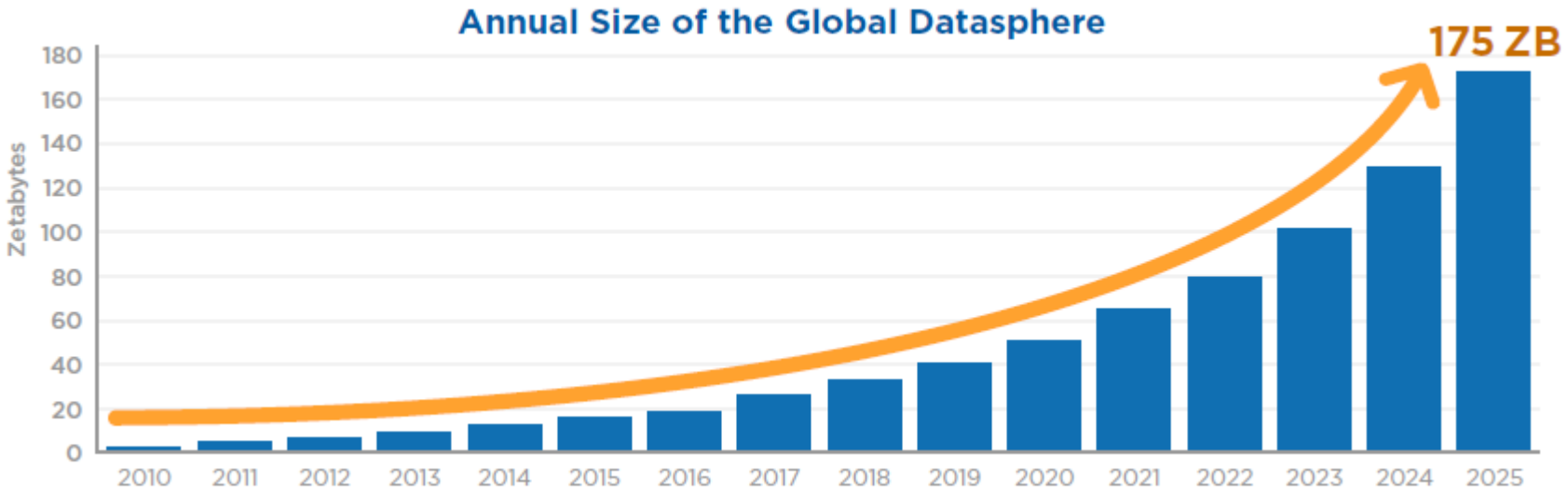
Содержание

- Введение
- Big Data
- Системы поддержки принятия решений
- Хранилища данных

Цифровая вселенная

К 2025 г. прогнозируется, что объем всех данных в мире составит 175 зеттабайт ($175*10^{21}$ байт).

Если записать на DVD, то можно достичь Луны 23 раза. Чтобы их «скачать» одному человеку понадобится 1,8 млрд. лет.



Почти 75% данных являются копиями. Используется менее 3% из 23% потенциально полезных данных. Только 5% информации в мире анализируется.

Объем (размер) данных

Название	Размер по ГОСТ 8.417-2002 (приставки по СИ)	Символ	Примечание: размер по стандартам МЭК
байт	8 бит	В	
килобайт	10^3 В	КВ	2^{10} =1024 байт
мегабайт	10^6 В	МВ	2^{20} байт
гигабайт	10^9 В	GB	2^{30} байт
терабайт	10^{12} В	ТБ	2^{40} байт
петабайт	10^{15} В	РВ	2^{50} байт
эксабайт	10^{18} В	ЕВ	2^{60} байт
зеттабайт	10^{21} В	ZB	2^{70} байт
йоттабайт	10^{24} В	YB	2^{80} байт

Цифровая вселенная

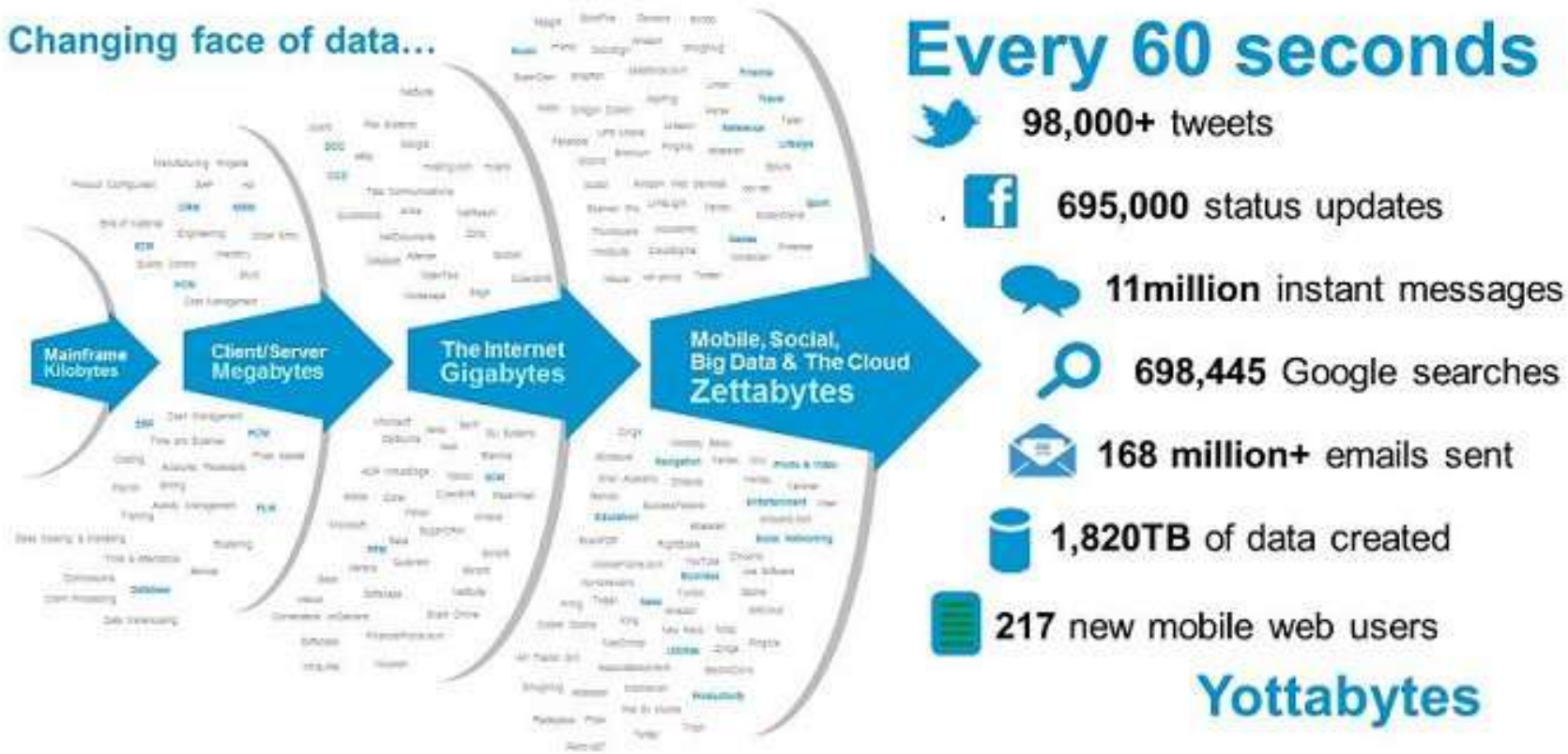
1. К 2025 году 75% населения будет иметь постоянный доступ в интернет. 6 млрд. пользователей будет взаимодействовать с данными ежедневно. Это 75% населения Земли.
2. Кратно возрастет количество умных гаджетов и домашних роботов (M2M). *Интернет вещей (IoT)*

Число устройств, подключенных к IoT, к 2025 году составит около 41,6 млрд. (системы наблюдения, промышленные, автомобильные и медицинские датчики и др.).

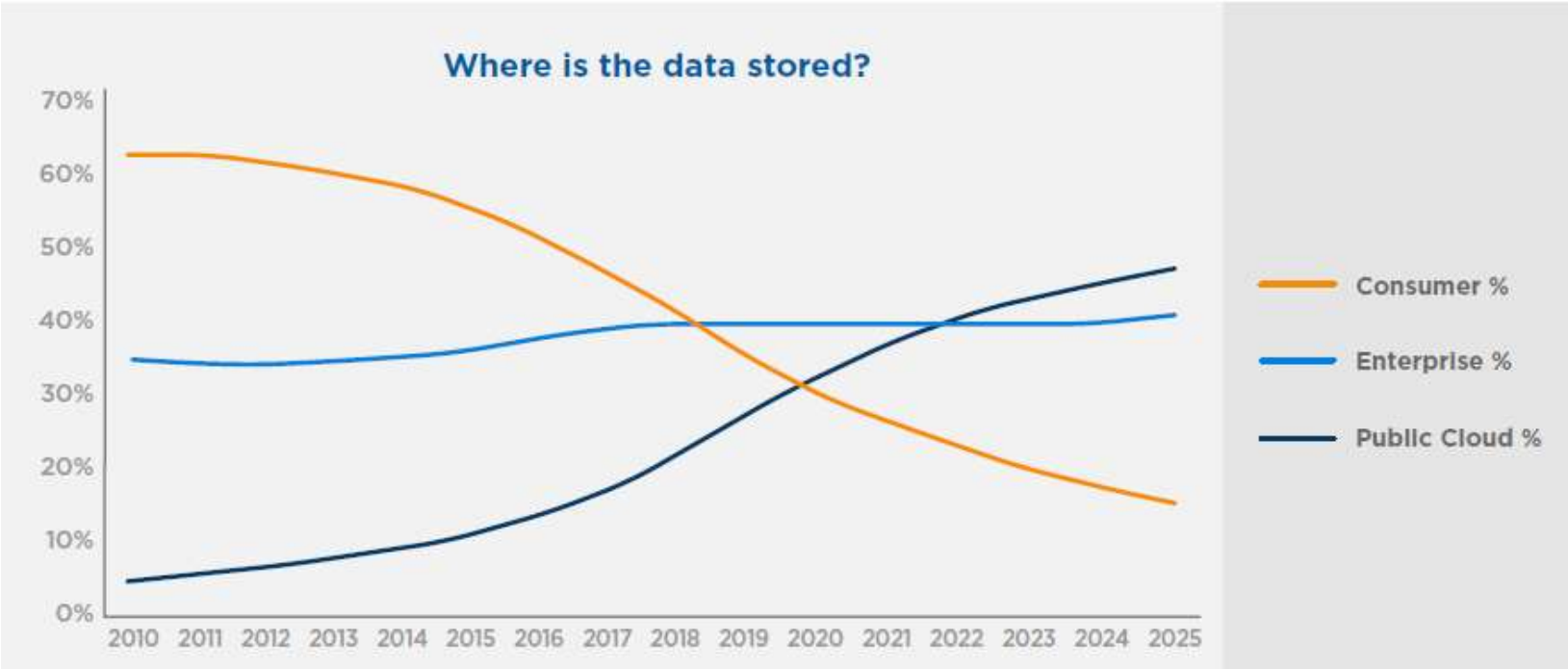
Объем данных, генерируемых этими устройствами, к 2025 году вырастет до 79,4 зеттабайта.

3. Каждый человек будет взаимодействовать с данными как минимум раз в 18 секунд.

Цифровая вселенная



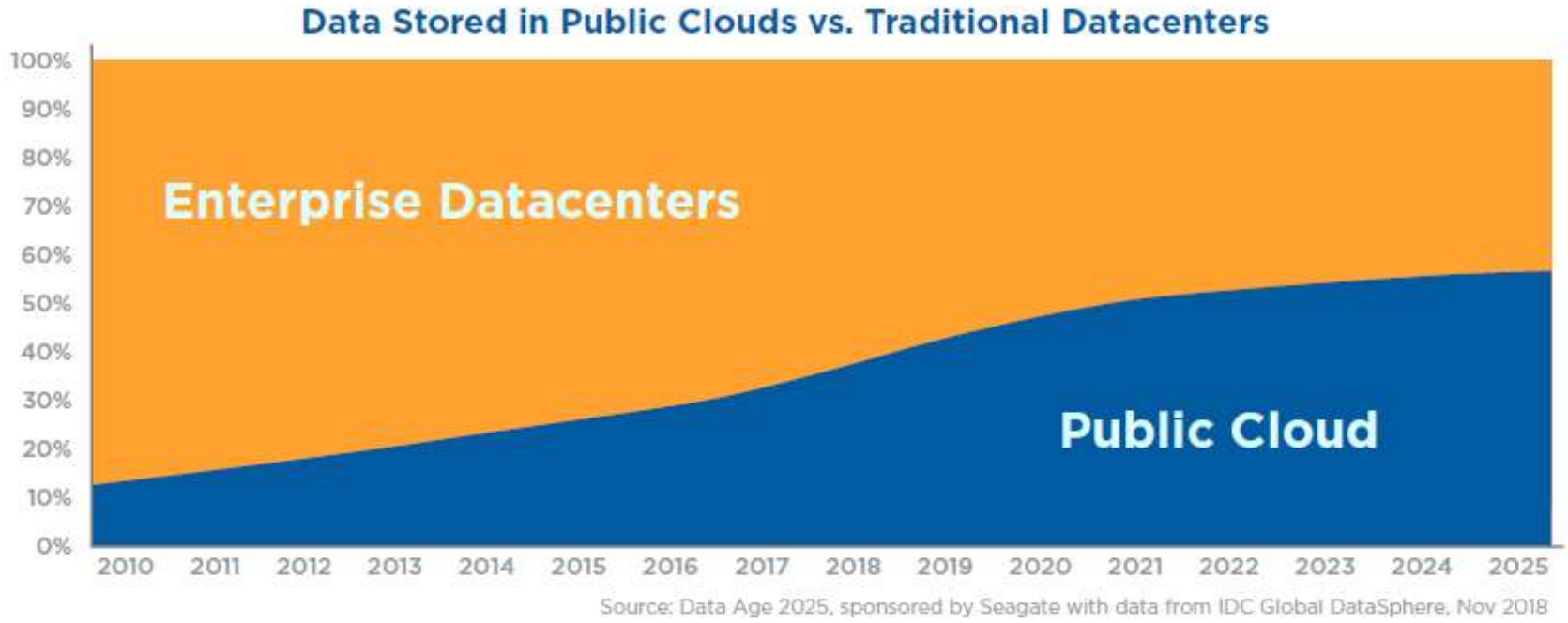
Где хранить?



К 2020 г. совокупной доступной емкости систем хранения будет достаточно для хранения менее чем 15%.

К 2025 г. 49% хранящихся в мире данных будут храниться в общедоступных облачных средах.

Где хранить?



Тип используемого хранилища зависит от типа данных и их применения: DVD, HDD, внешние дисковые массивы и ленты, RAID-массивы и т.п.

Облачные хранилища: Microsoft, Amazon, Dell EMC, Google и др.

Большие данные

Большие данные (*Big Data*) — совокупность подходов, инструментов и методов обработки данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста и распределения по многочисленным узлам вычислительной сети.

Данные

Данные – ресурс для получения информации. Они должны быть представлены в форме, пригодной для хранения, передачи и обработки.

Структурированные данные - организуют в ряды и колонки строго определенного формата, чтобы приложения могли извлекать данные и эффективно обрабатывать их. Обычно хранятся с применением СУБД (~ 20%).

Неструктурированные данные: офисная документация, графические данные, чертежи, веб-страницы, сообщения электронной почты и IM, видео- и аудиофайлы и другие мультимедийные активы (~ 80%) .

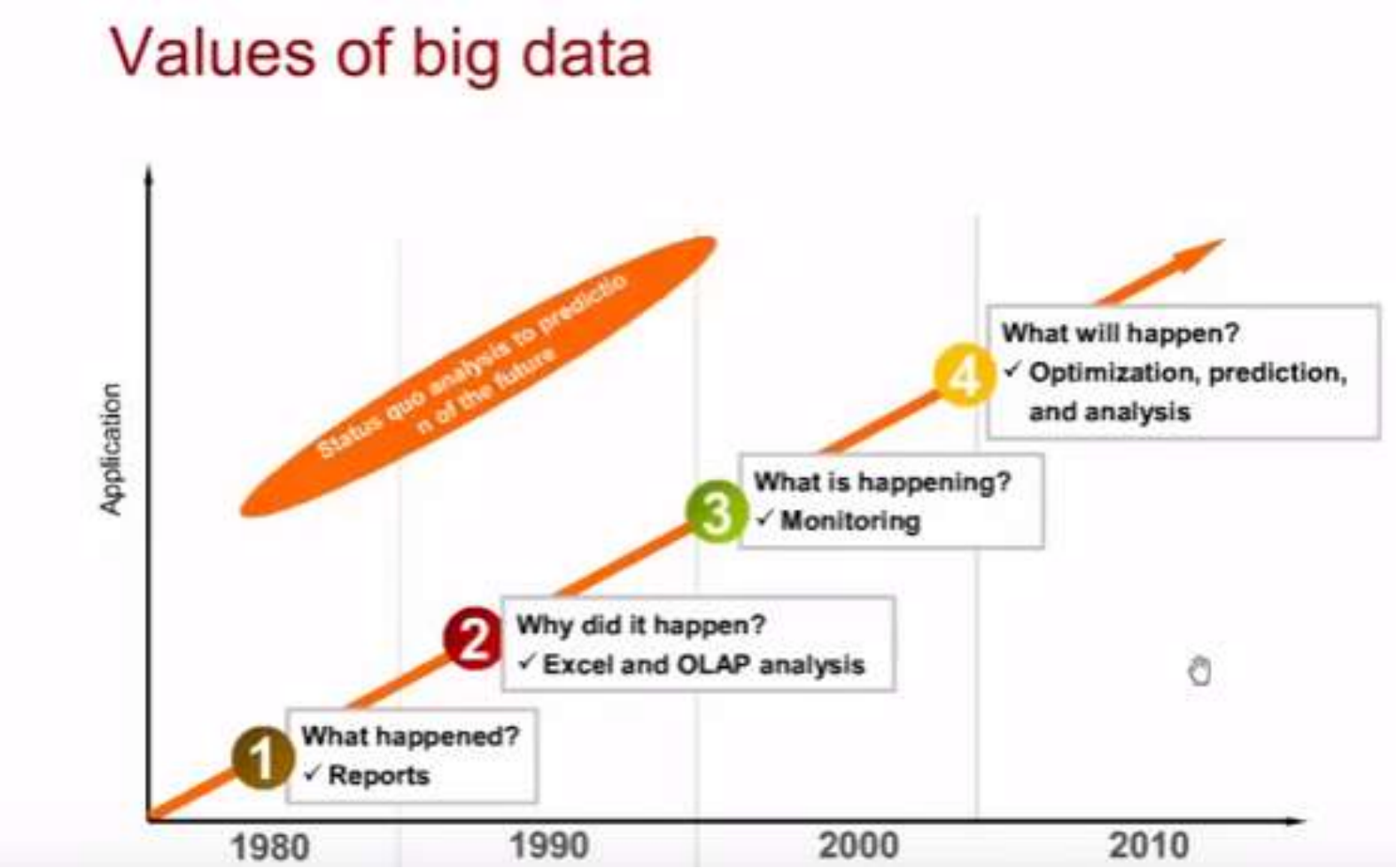
Big Data

В качестве определяющих характеристик для больших данных отмечают *«5 V's»*:

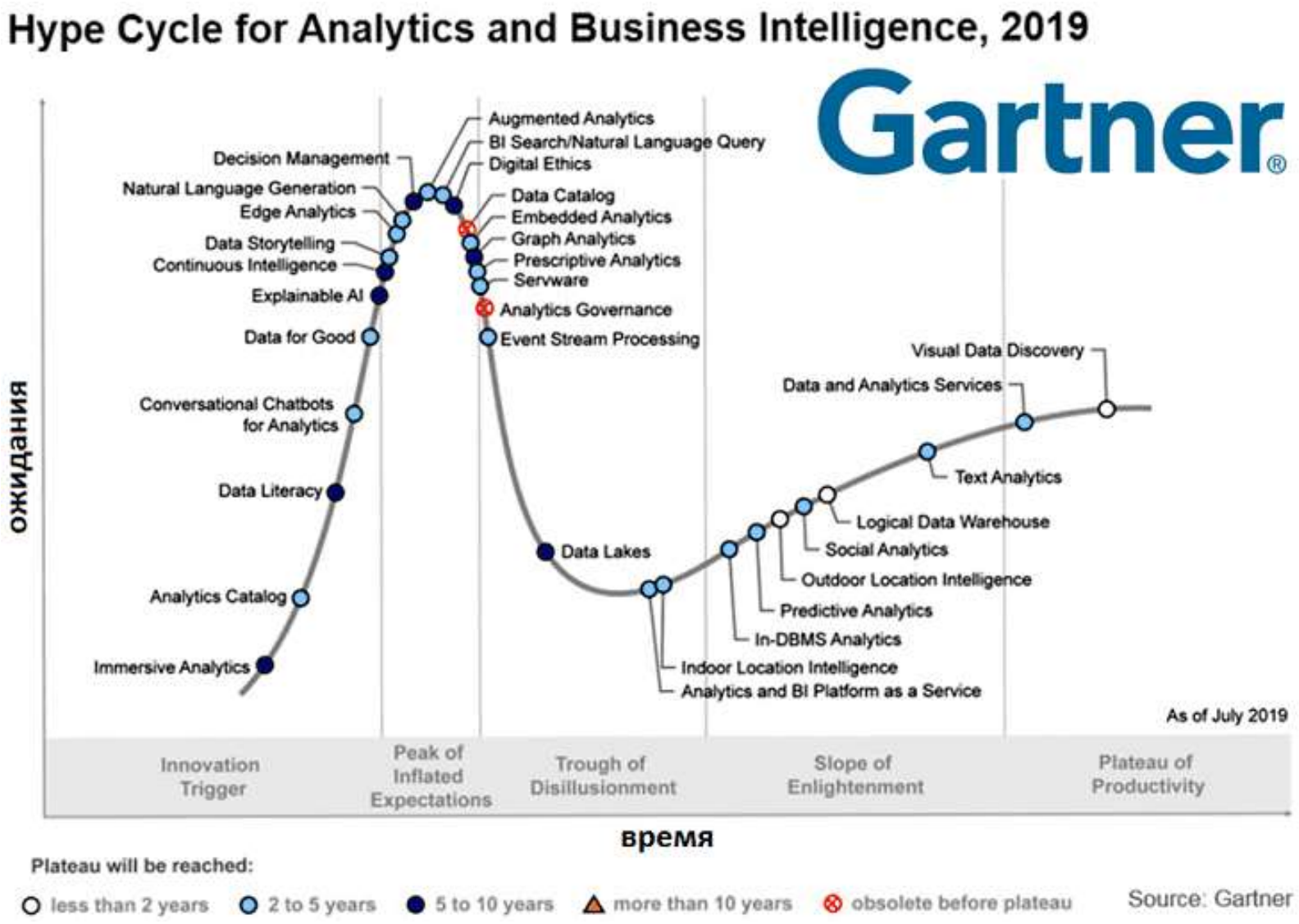
- ✓ объём (*volume*, физический объём данных),
- ✓ скорость (*velocity,* как скорость прироста, так и скорость обработки и получения результатов),
- ✓ многообразие (*variety*, одновременная обработка различных типов данных; P2P, P2M, M2M),
- ✓ точность (*veracity*, достоверность: верификация и валидация данных),
- ✓ ценность (*value*, экономический эффект для пользователей).

Конвейер монетизации данных

Эволюция работы с данными (повышение ценности данных)



Цикл зрелости технологий



Business Intelligence

Бизнес-аналитика (BI) – это общий термин, подразумевающий под собой разнообразные программные продукты и приложения, созданные для анализа первичных данных организации.

Бизнес-анализ как деятельность состоит из нескольких связанных между собой процессов:

- Получение информации из баз данных *(OLTP-системы)*,

Область детализированных данных. Цель – поиск информации (информационно-поисковые системы).

- Аналитическую обработку в реальном времени *(online analytical processing, OLAP)*,

Сфера агрегированных показателей. Цель – обобщение информации и многомерный анализ.

- Интеллектуальный анализ данных *(data mining)*

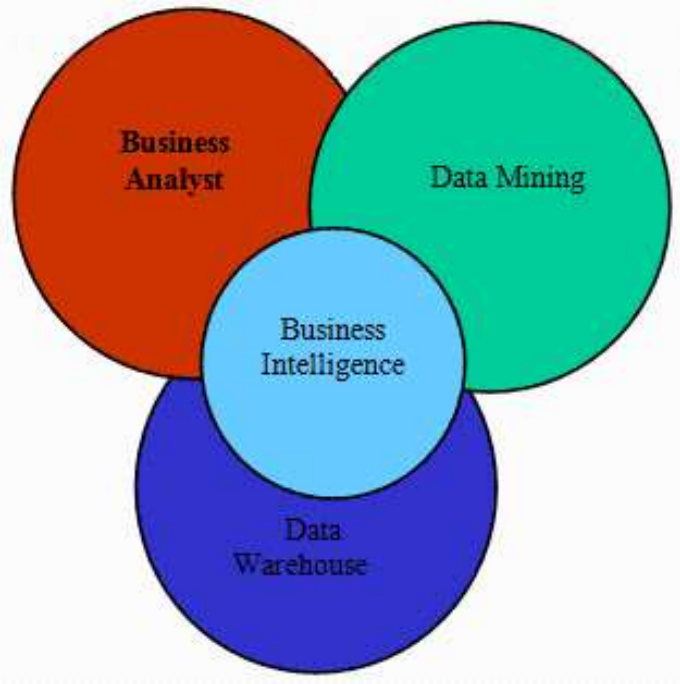
Сфера закономерностей. Цель – поиск закономерностей в накопленной информации, построение моделей и правил, которые объясняют найденные аномалии и/или прогнозируют развитие некоторых процессов.

- Составление отчетов *(reporting)*.

Business Intelligence

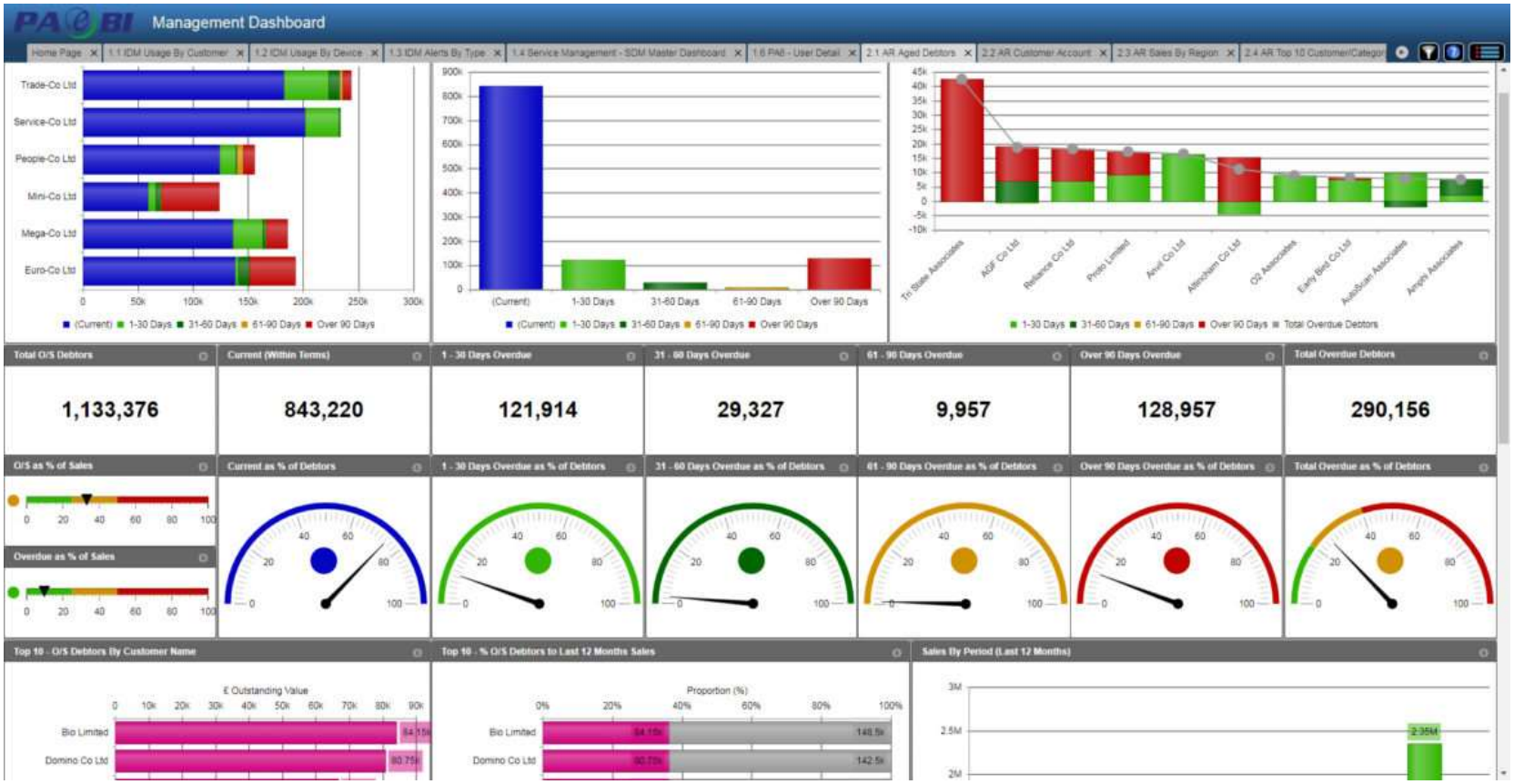
Инструменты Business Intelligence решают большой спектр задач:

- ✓ Моделирование бизнес-ситуаций
- ✓ Анализ нестандартных запросов и их решение
- ✓ Снижение нагрузки на сотрудников компаний, путем автоматизации их работы
- ✓ Улучшенные данные при увеличении объема этих данных
- ✓ Объективная оценка бизнеса
- ✓ Анализ использования финансовых ресурсов
- ✓ Прогноз и оценка инвестиционной и финансовой деятельности



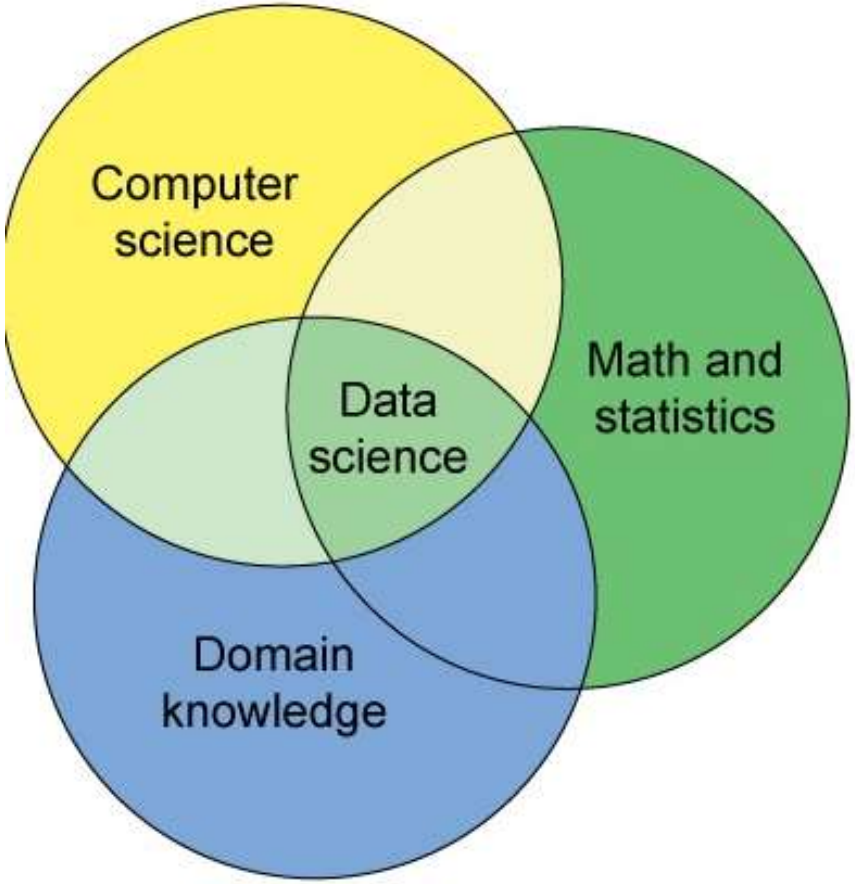
Business Intelligence

Интерфейс руководителя – панель индикаторов (dashboards)



Data Science

Наука о данных (*data science*) — раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме.



Data Scientist (Big Data Analyst)

Data Scientist - это ученый-эксперт по данным, который занимается сбором данных и умеет находить в них логические закономерности, решает бизнес-задачи с использованием данных и математического моделирования.

Он изучает огромные массивы информации со сложной структурой (результаты исследований, рыночные тенденции, предпочтения клиентов и пр.).

Главная компетенция – умение видеть логические связи в массивах собранной информации, и на основании этого разрабатывать новые подходы и решения.

Навыки: работать с языками программирования (SAS, R и Python), работать со статистикой, использовать аналитические методы, методы интеллектуального анализа данных, приложения искусственного интеллекта, методы проектирования баз данных, а также уметь визуализировать данные.

Big Data Analyst

Объем мирового рынка технологий обработки больших данных и бизнес-аналитики в 2019 году достигнет 189,1 млрд. долл.
К 2022 году он вырастет до 274,3 млрд. долл.

Data Engineer (Big Data Developer)

Data Engineer специализируется на организации процесса сбора, очистки и предобработки данных («garbage in — garbage out»).

1) Понимание сути и сбор данных (источники, типы, структуры данных, организация доступа)
2) Построение архитектуры процесса обработки данных (обработка в режиме реального времени и офлайн).
3) Превращение моделей в готовый продукт или сервис (интеграция в сайт, связь с базой данных и т.п.).

Если аналитики Big Data отвечают за анализ больших данных, выявление взаимосвязей и построение моделей, то инженеры Big Data отвечают за хранение, преобразование данных и быстрый доступ к ним.

Знание современных технологий и подходов в области обработки данных: MapReduce, Hadoop, Spark, Aerospike, Redis, Storm и т.д.

Знание языков программирования (SAS, R и Python) и библиотек (Pandas, Numpy, Scikit-learn).

Знание технологии ETL.

Системы оперативной обработки информации

OLTP (On-Line Transaction Processing) — оперативная (т.е. в режиме реального времени) обработка транзакций.

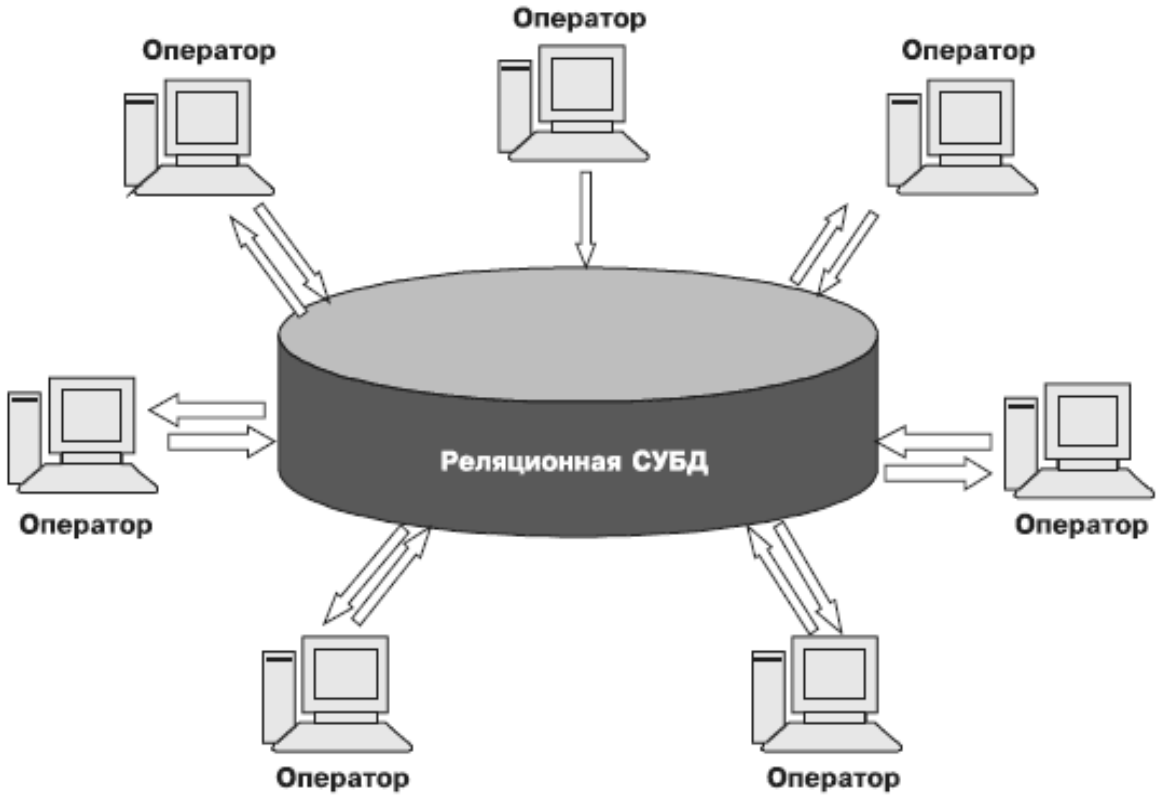
Транзакция —набор операций, который рассматривается как единое завершенное, с точки зрения пользователя, действие над некоторой информацией, обычно связанное с обращением к базе данных.

OLTP предполагает, что система работает с небольшими по размерам транзакциями, но идущими большим потоком, и при этом клиенту требуется от системы минимальное время отклика.

Главное требование — быстрое обслуживание (не более нескольких секунд) относительно простых запросов большого числа пользователей.

Обычно аналитические возможности OLTP-систем сильно ограничены.

Обобщенная структура системы OLTP



Примеры: системы складского учета, системы заказов билетов, банковские системы. Два общих свойства: очень большое число клиентов и непрерывное поступление информации.

Системы поддержки принятия решений (постановка задачи)

Со временем в OLTP начали аккумулироваться большие объемы данных.

Сбор данных – не самоцель!

Появилась потребность в ИС, которые позволяли бы проводить глубокую аналитическую обработку (поиск закономерностей, вывод из них правил, принятие решений и прогнозирование их последствий).

Системы поддержки принятия решений

Информационные системы поддержки принятия решений (**Decision Support System, DSS, СППР**) – ориентированы на аналитическую обработку данных с целью получения знаний, необходимых для разработки решений в области управления (BI).

В зависимости от данных, с которыми работают СППР, выделяют два основных их типа: **статические** и **динамические**.

- **статические СППР** (информационные системы руководителя, Executive Information Systems — EIS)

Содержат в себе предопределенные множества запросов и неспособны ответить на все вопросы, которые могут возникнуть при принятии решений. Результатом работы такой системы, как правило, являются многостраничные отчеты, которые нельзя изменить без привлечения программиста.

- **динамические СППР**.

Ориентированы на обработку нерегламентированных, неожиданных (ситуативных) запросов аналитиков к данным.

Отличия СППР и OLTP-систем

Свойство	OLTP-система	СППР
Цели использования данных	Быстрый поиск, простейшие алгоритмы обработки	Аналитическая обработка с целью поиска скрытых закономерностей, построения прогнозов и моделей
Уровень обобщения (детализации) данных	Детализированные	Как детализированные, так и обобщенные (агрегированные)
Требования к качеству данных	Возможны некорректные данные (ошибки регистрации, ввода и т.д.)	Ошибки в данных не допускаются, поскольку могут привести к некорректной работе аналитических алгоритмов
Формат хранения данных	Данные могут храниться в различных форматах в зависимости от приложения, в котором они были созданы	Данные хранятся и обрабатываются в едином формате
Время хранения данных	Как правило, не более года (в пределах отчетного периода)	Годы, десятилетия
Изменение данных	Данные могут добавляться, изменяться и удаляться	Допускается только пополнение; ранее добавленные данные изменяться не должны, что позволяет обеспечить их хронологию
Периодичность обновления	Часто, но в небольших объемах	Редко, но в больших объемах
Доступ к данным	Должен быть обеспечен доступ ко всем текущим (оперативным) данным	Должен быть обеспечен доступ к историческим (то есть накопленным за достаточно длительный период времени) данным
Характер выполняемых запросов	Стандартные, настроенные заранее	Нерегламентированные, формируемые аналитиком «на лету» в зависимости от требуемого анализа
Время выполнения запроса	Несколько секунд	До нескольких минут (важно, но не критично)
Число пользователей	Поддержкой большого числа пользователей	Небольшое число пользователей (аналитики)

Хранилища данных

Билл Инмон (1989г.): "**Хранилище данных** (Data Warehouse) - это предметно-ориентированное, привязанное ко времени и неизменяемое собрание данных для поддержки процесса принятия управляющих решений".

Хранилище данных (ХД) управляет данными, которые были собраны как из OLTP-систем, так и из внешних источников данных, и которые длительный период времени хранятся в системе.

Основные характеристики хранилищ данных

- Ориентация на предметную область. Учитывает специфику предметной области (клиенты, товары, продажи), а не прикладных областей деятельности (выписка счетов, контроль запасов, продажа товаров).
- Интегрированность и внутренняя непротиворечивость. Поскольку данные в хранилище поступают из разных источников, необходимо привести их к единому формату.
- Содержит исторические данные с привязкой ко времени (учет хронологии).
- Неизменяемость. Данные не обновляются в оперативном режиме, а лишь регулярно пополняются.
- Поддержка высокой скорости получения данных из хранилища.
- Предназначено для проведения анализа и принятия стратегических решений.
- Полнота (хранит как подробные сведения, так и частично и полностью обобщенные данные) и достоверность хранимых данных.
- Поддержка качественного процесса пополнения данных.
- Обслуживает относительно малое количество работников руководящего звена и аналитиков.

Виды данных

Детализированные данные поступают непосредственно из источников данных и соответствуют элементарным событиям, регистрируемым OLTP-системами. Такими данными могут быть ежедневные продажи, количество произведенных изделий и т.д. Это неделимые значения.

Агрегированные данные – обобщенные данные. Если обобщить данные в пределах недели или месяца и взять сумму, среднее, максимальное и минимальное значения за соответствующий период, то полученный ряд может оказаться более информативным, чем конкретные значения.

Так как один и тот же набор детализированных данных может породить несколько наборов агрегированных данных, то объем данных существенно возрастает. Однако, если бы они вычислялись в процессе выполнения запросов, время выполнения запроса увеличилось бы в несколько раз.

Метаданные

Метаданные («данные о данных») необходимы для описания значения и свойств информации с целью лучшего ее понимания, использования и управления ею.

Примеры: Книга (аннотация, глоссарий, оглавление, номера страниц, об авторах и издательстве), Фотография (дата, формат, размер, координаты).

Одно из основных назначений метаданных — повышение эффективности поиска.

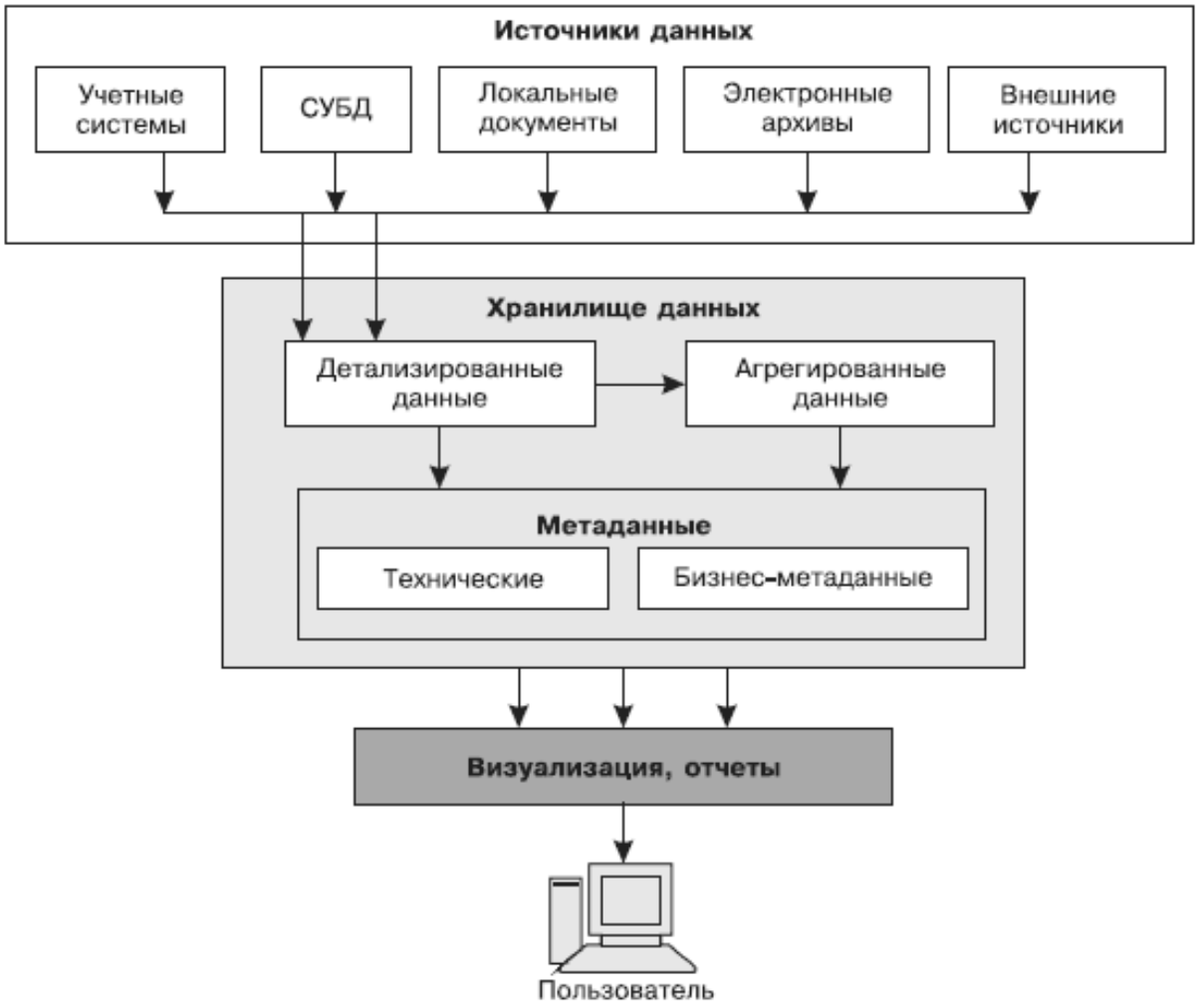
Метаданные хранятся отдельно от данных в **репозитарии метаданных**.

Метаданные

Два уровня метаданных: технический (административный) и бизнес-уровень.

- **Технический уровень** содержит метаданные, необходимые для обеспечения функционирования хранилища (статистика загрузки данных и их использования, описание модели данных и т.д.).
- **Бизнес-метаданные** описывают объекты предметной области — атрибуты объектов и их возможные значения, соответствующие поля в таблицах и т.д.

Обобщенная концептуальная схема хранилища данных



Извлечение данных (ETL)

Извлечение данных из разнотипных источников и перенос их в ХД с целью дальнейшей аналитической обработки связано с рядом проблем:

- Исходные данные расположены в источниках самых *разнообразных типов и форматов*, созданных в различных приложениях. Для анализа данные должны быть преобразованы в единый универсальный формат, который поддерживается ХД и аналитическим приложением.
- Данные в источниках обычно *излишне детализированы*, тогда как для решения задач анализа в большинстве случаев требуются обобщенные данные.
- Исходные данные, как правило, являются «*грязными*» (отсутствующие, неточные или бесполезные данные), что мешает их корректному анализу.

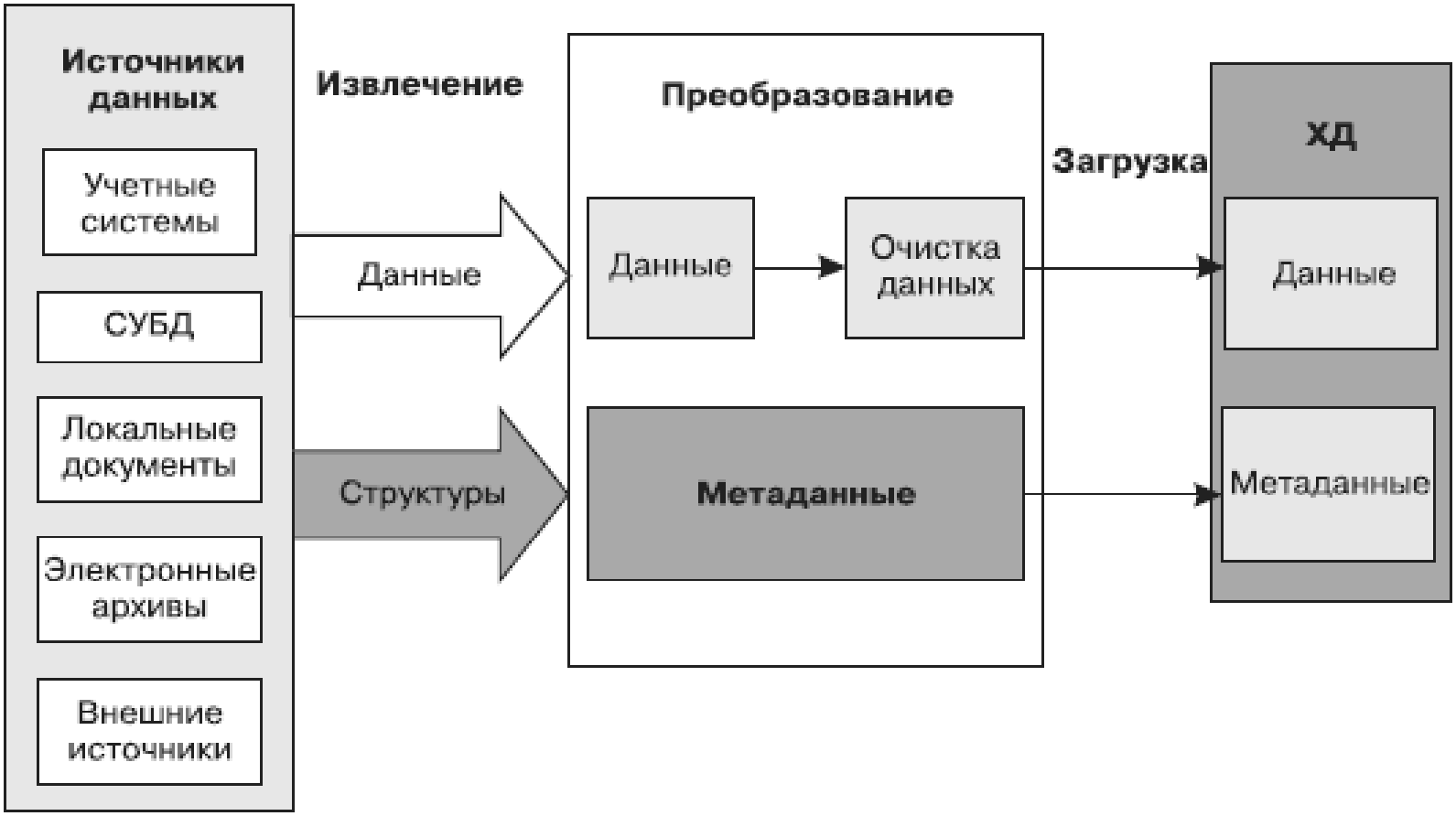
Извлечение данных (ETL)

ETL (extraction, transformation, loading – извлечение, преобразование и загрузка данных) – комплекс методов, реализующих процесс переноса исходных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных.

Приложения ETL извлекают информацию из источников, преобразуют ее в формат, поддерживаемый системой хранения и обработки, а затем загружают в нее преобразованную информацию.

Популярные ETL-инструменты: IBM Infosphere Information Server, Oracle Data Integrator, Microsoft SQL Server Integrated Services, CloverDX Data Integration Software, SAS Data Integration Studio, SAP BusinessObjects Data Integrator.

Обобщенная структура процесса ETL



Извлечение данных

Данные извлекаются из одного или нескольких источников и подготавливаются к преобразованию.

Из источников должны извлекаться не только сами данные, но и информация, описывающая их структуру, из которой будут сформированы метаданные.

Процесс извлечения данных может выполняться ежедневно, или еженедельно. Иногда требуются извлечение данных в режиме реального времени: например, системы, анализирующие биржевые операции.

Преобразование данных

Процесс преобразования данных включает в себя:

- Преобразование типов данных
- Преобразования, связанные с нормализацией схемы данных
- Преобразования ключей.
- Преобразования, связанные с обеспечением качества данных в ХД (очистка данных).

Очистка данных

Данные в ХД должны быть:

- точными – должны содержать правильные количественные значения метрик;
- полными – пользователи должны иметь доступ ко всем релевантным данным;
- согласованными – никакие противоречия в данных не допускаются (агрегаты должны точно соответствовать детализированным данным);
- уникальными – одни и те же объекты должны иметь одинаковые наименования и идентифицироваться одинаковыми ключами;
- актуальными – пользователи должны знать, с какой частотой данные обновляются (т.е. на какую дату данные действительны).

Очистку данных можно разделить на следующие типы:

- конвертация и нормализация данных (согласование форматов данных, например, даты),
- обнаружение одинаковых имен атрибутов, с различными по смыслу значениями;
- стандартизация написания имен (ФИО), представления адресов (Улица", "Ул."), устранение дубликатов;
- замещение кодов значениями (например, почтового индекса наименованием населенного пункта);
- исключение ненужных атрибутов (например, комментариев);
- стандартизация наименований таблиц, индексов и т.д.

Загрузка данных

Основная цель - быстрая загрузка данных в хранилище.

Загрузка данных, основанная на использовании команд SQL, является медленной, поэтому загрузка с помощью встроенных в СУБД средств импорта/экспорта является предпочтительной.

При загрузке данных должна быть гарантирована ссылочная целостность данных, а агрегаты должны быть построены и загружены одновременно с детализированными данными.

Архитектуры хранилищ данных

Под архитектурой ХД понимают совокупность программно-аппаратных компонент, совокупность технологических и организационных решений, предпринимаемых для создания, разработки и функционирования ХД.

Шесть уровней архитектуры хранилища данных

Документы	ETL	Ведение НСИ	SRD	Тематическая витрина данных	Сценарный анализ
Унаследованные системы		Ведение метаданных		Региональная витрина данных	Статистический анализ
Транзакционные системы		Центральное хранилище данных		Витрина данных подразделения	Многомерный анализ
Файлы		Оперативный склад данных		Прикладная витрина данных	Отчетность
Архивы		Зоны временного хранения		Функциональная витрина данных	Планирование
Источники данных	**Извлечение, преобразование, загрузка**	**Хранение данных**	**Выборка, реструктуризация, доставка**	**Предоставление данных**	**Бизнес-приложения**

Уровень хранения данных

- *НСИ* – нормативно-справочная информация (набор классификаторов, справочников, словарей, стандартов, регламентов, используемых предприятием).
- *Оперативный склад данных* необходим тогда, когда требуется быстрый доступ к пусть неполным, не до конца согласованным данным (например, в банках фрод-операции). Часто для этих целей создается **озеро данных** (Data Lake). Оно позволяет хранить данные в исходных форматах (часто нужно для расширенной аналитики). Обычно на платформе Hadoop.
- *Зоны временного хранения* нужны для реализации специфического бизнес-процесса, например, когда перед загрузкой данных контролер данных должен просмотреть их и дать разрешение на их загрузку в хранилище. Для этих зон требуется создание дополнительных средств администрирования, мониторинга, обеспечения безопасности и аудита.

Системы SRD

SRD (Sample, Restructure, Deliver) - выборка, реструктуризация и доставка данных.

SRD выполняет выборку из единого ХД и имеет дело с очищенными данными, структуры которых должны быть приведены в соответствие с требованиями различных приложений.

SRD должно доставить данные в различные витрины в соответствии с правами доступа, графиком доставки и требованиями к составу информации.

Витрины данных

Витрина (киоски) данных (data marts) — срез ХД, представляющий собой массив тематической, узконаправленной информации, ориентированный, например, на пользователей одного департамента.

Витрины данных должны иметь структуры данных, максимально отвечающие потребностям обслуживаемых задач.

Витрины данных следует группировать по территориальным, тематическим, организационным, прикладным, функциональным и иным признакам.

Достоинства витрин данных

- Аналитики видят и работают только с теми данными, которые им реально нужны.
- Для реализации витрин данных не требуется мощная вычислительная техника.
- Относительно небольшой объем хранимых данных, на организацию и поддержку которых не требуется значительных затрат.
- Корпоративная информационная система может эффективно наращиваться за счет добавления новых витрин данных.
- Использование витрин данных позволяет снизить нагрузку на централизованное ХД.

Озера данных (Data Lakes)

Проблема: все данные разнородные и неструктурированные, перед загрузкой в базы их нужно долго обрабатывать. В итоге работа с Big data оказывается слишком сложной и дорогой.

Озеро данных представляет собой огромное хранилище, в котором разные данные хранятся в «сыром», т.е. неупорядоченном и необработанном виде.

Когда данные сохранены, с ними можно работать — извлекать по определенному шаблону в классические базы данных или анализировать и обрабатывать прямо внутри data lake.

Альтернатива витринам данных (Джеймс Диксон, 2015 год)

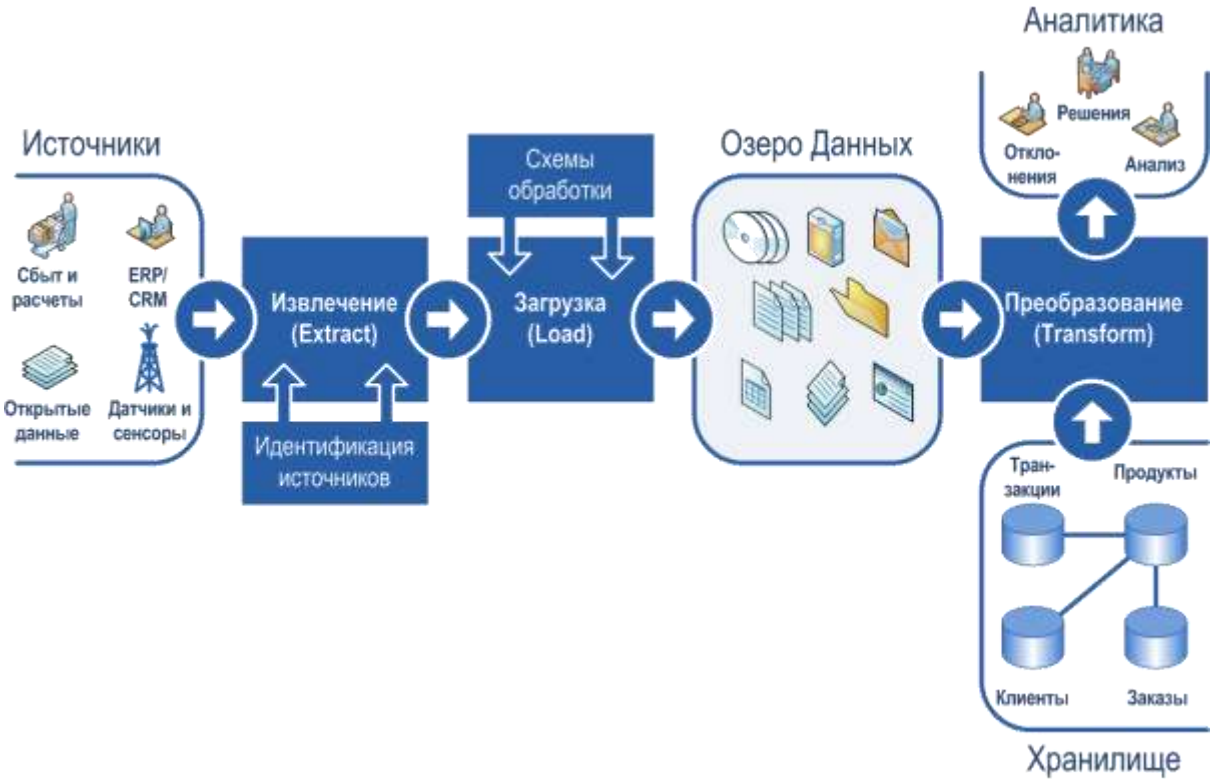
Озера данных (Data Lakes)

Принципы

- Все данные имеют ценность (сейчас или в будущем)
- Храним столь долго, сколько нужно
- Если ценность в будущем – храним в том виде, в котором есть
- Преобразовываем только тогда, когда возникает необходимость
- Приложения и пользователи интерпретируют данные по собственному усмотрению

Озера данных (Data Lakes)

- Не обязательный прием только структурированных данных
- Преобразование данных не является шагом препроцессинга, а превращается в «постпроцессинг по запросу»
- Резкое снижение требований к инфраструктуре в части скорости обработки информации



Озера данных (Data Lakes)

	Хранилище данных	Озеро Данных
Поддерживаемые бизнес-сценарии	• Единая аналитика и отчетность по предустановленным формам • Формирование единого подхода к данным различной природы • Встроенный анализ сервисов самообслуживания	• Исследование взаимосвязей и поиск закономерностей • Программируемый доступ к первичным данным • Поддержка глубокого погружения в массивы накопленных данных
Основные особенности	• Высокая эффективность исполнения типовых запросов • Низкое время отклика при выполнении типовых запросов • Согласованность задач в пределах корпорации	• Обработка очень больших массивов данных • Масштабируемость в объёмах хранения данных, компромиссность в скорости их обработки • Возможность независимого решения множества задач с использованием накопленных данных
Требования к ИТ-инфраструктуре	• Хранение больших объёмов данных с приемлемыми (средними) издержками • Стандартизованные инструменты анализа и унифицированный язык запросов • Высокоэффективное управление хорошо структурированными данными	• Хранение огромных объёмов данных с низкими издержками • Возможность разработки (в том числе – конструирования по шаблонам) новых алгоритмов анализа и преобразования данных • Одинаково хорошее управление как структурированными, так и неструктурированными данными

Озера данных (Data Lakes)

	Хранилище данных	Озеро Данных
Ракурс: сотрудники		
Организация работы	• ИТ-подразделение: аналитики создают модели данных, кубы и формы отчётов • Бизнес-подразделения: сотрудники используют данные хранилища через инструменты визуализации и отчёты	• Анализ данных ведётся в условиях тесного взаимодействия аналитиков данных и разработчиков, при непосредственном участии специалистов бизнес-подразделений в постановке аналитических задач
Навыки сотрудников	• Знание SQL и технологий баз данных • Понимание структуры бизнеса	• Продвинутые техники анализа данных и интерпретации результатов • Навыки программирования
Ракурс: процессы		
Модели данных	Строгая детерминированная схема обработки данных	Гибкая схема чтения и комбинирования данных
Качество данных	• Однократная проверка данных при их первичной загрузке с помощью ETL • Требования к модели данных определяются до их загрузки в хранилище • Обязательная нормализация и индексация данных	• Проверка данных и «очистка от загрязнений» при каждом их преобразовании • Модели данных создаются под каждый пользовательский запрос, система правил может постоянно расширяться
Быстродействие	Долгая загрузка, быстрая обработка	Быстрая загрузка, долгая обработка

Озера данных (Data Lakes)

Недостатки: любые данные, попадающие в data lake, попадают туда практически бесконтрольно. Это значит, что определить их качество невозможно. Если у компании нет понимания типов структур данных и методов их обработки, в нем быстро накапливаются огромные объемы неконтролируемых данных, чаще всего бесполезных.

В итоге озеро превращается в «болото» данных — бесполезное, пожирающее ресурсы компании и не приносящее пользы.

Решение: наладить процесс управления данными — data governance.

- отсекать источники с заведомо недостоверными данными;
- ограничить доступ на загрузку для сотрудников, у которых нет на это прав;
- проверять некоторые параметры файлов, например не пропускать в озеро картинки, которые весят десятки гигабайт.

Обычно за это отвечает **Chief Data Officer, CDO**

Реляционные хранилища данных

В отличие от OLTP-систем РХД проектируются так, чтобы добиться минимального времени выполнения запросов на чтение (у OLTP минимизируется время выполнения запросов на изменение данных).

Типичная структура РХД существенно отличается от структуры обычной реляционной БД. Как правило, эта структура денормализована (это повышает скорость выполнения запросов) и может допускать избыточность данных.

Измерения и факты в РХД

Измерения — это категориальные атрибуты, наименования и свойства объектов, участвующих в некотором бизнес-процессе. Измерения могут быть и числовыми, если какой-либо категории (например, наименованию товара) соответствует числовой код, но в любом случае это данные дискретные, то есть принимающие значения из ограниченного набора. Измерения качественно описывают исследуемый бизнес-процесс.

Факты — это данные, количественно описывающие бизнес-процесс, непрерывные по своему характеру, то есть они могут принимать бесконечное множество значений. Примеры: цена товара, их количество, сумма продаж, зарплата сотрудников и т.д.

В основе технологии РХД лежит принцип, в соответствии с которым измерения хранятся в плоских таблицах так же, как и в обычных реляционных БД, а факты — в отдельных специальных таблицах этой же БД.

Схема построения «Звезда»

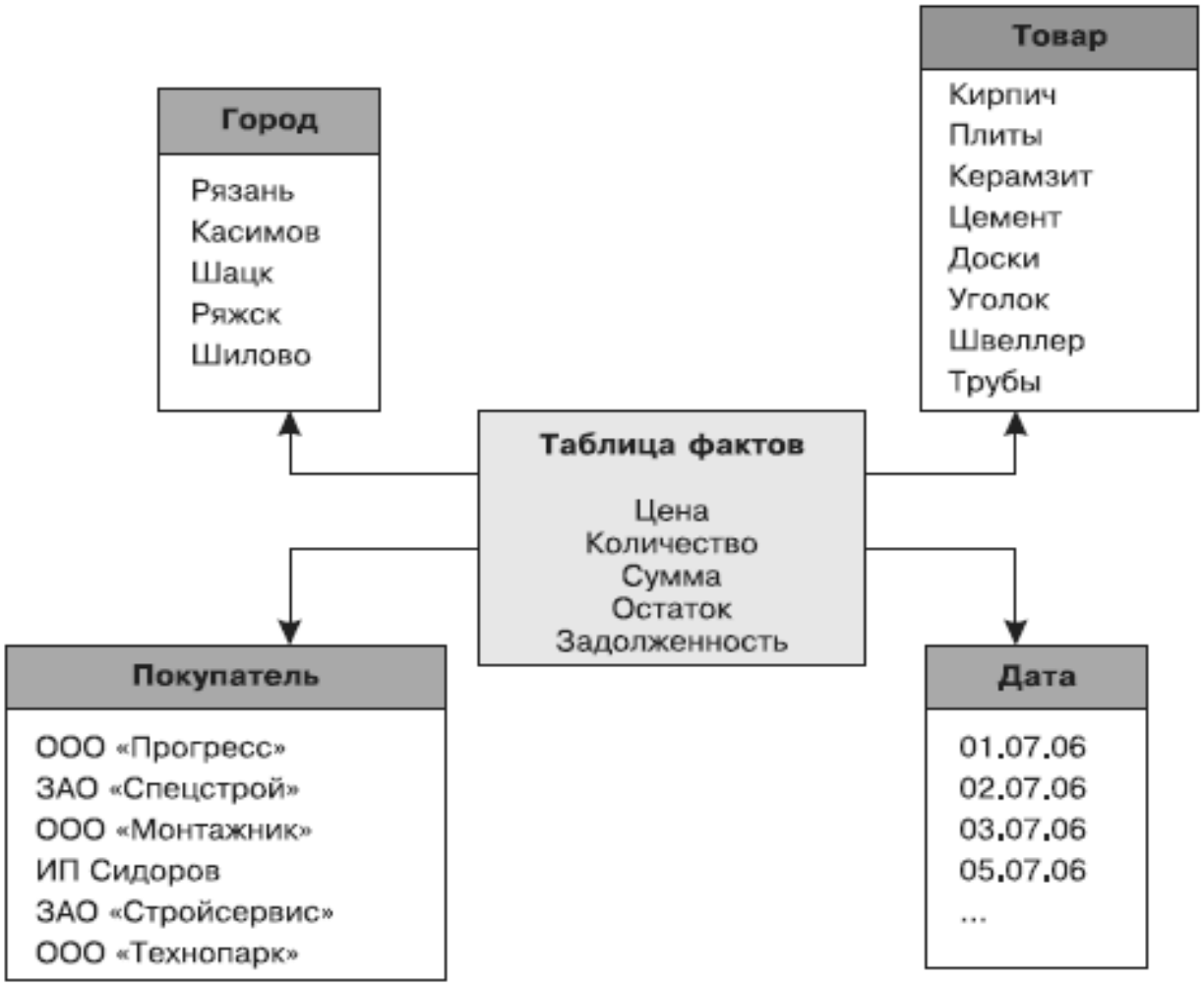


Схема построения «Звезда»

Преимущества схемы «звезда»:

- простота и логическая прозрачность модели;
- простая процедура пополнения измерений, поскольку приходится работать только с одной таблицей.

Недостатки схемы «звезда»:

- наличие иерархий в данных вызовет рост избыточности, замедлит обработку измерений (т.к. одни и те же значения могут встречаться несколько раз в одной и той же таблице) и повысит вероятность возникновения противоречий (например, из-за ошибок ввода).

Схема построения «Снежинка»

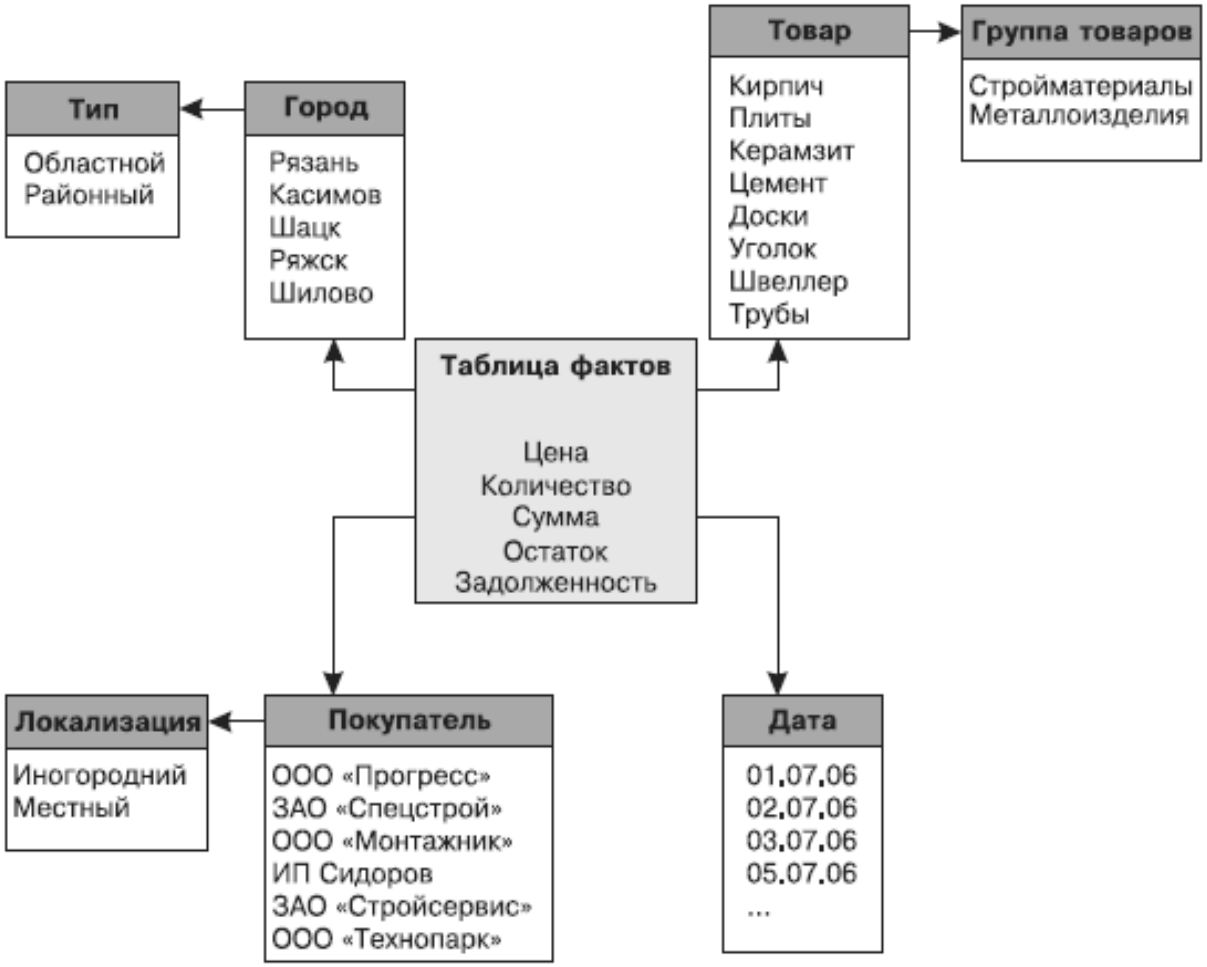


Схема построения «Снежинка»

Преимущества схемы «снежинка»:

- она ближе к представлению данных в многомерной модели;
- процедура загрузки из РХД в многомерные структуры более эффективна и проста, поскольку загрузка производится из отдельных таблиц;
- намного ниже вероятность появления ошибок несоответствия данных;
- большая, по сравнению со схемой «звезда», компактность представления данных, поскольку все значения измерений упоминаются только один раз.

Недостатки схемы «снежинка»:

- достаточно сложная для реализации и понимания структура данных;
- усложненная процедура добавления значений измерений.

Преимущества и недостатки РХД

Основные преимущества:

- практически неограниченный объем хранимых данных;
- поскольку реляционные СУБД лежат в основе построения многих систем OLTP, которые обычно являются главными источниками данных для ХД, использование реляционной модели позволяет упростить процедуру загрузки и интеграции данных в хранилище;
- при добавлении новых измерений данных нет необходимости выполнять сложную физическую реорганизацию хранилища;
- обеспечиваются высокий уровень защиты данных и широкие возможности разграничения прав доступа.

Главный недостаток РХД – невысокая производительность, из-за большого числа таблиц агрегатов.

Таким образом, выбор реляционной модели целесообразен если:

- Значителен объем хранимых данных.
- Иерархия измерений несложная (немного агрегированных данных).
- Требуется частое изменение размерности данных (можно ограничиться добавлением новых таблиц).

Многомерные хранилища данных

Большинство бизнес-процессов описывается множеством атрибутов. Если собрать всю информацию в таблицу, то она окажется сложной для визуального анализа. Более того, она может оказаться избыточной, т.к. в плоской таблице хранятся многомерные данные.

По Эдгару Кодду **многомерное концептуальное представление** – множественная перспектива, состоящая из нескольких независимых измерений, вдоль которых могут быть проанализированы определенные совокупности данных.

Одновременный анализ по нескольким измерениям определяется как **многомерный анализ**.

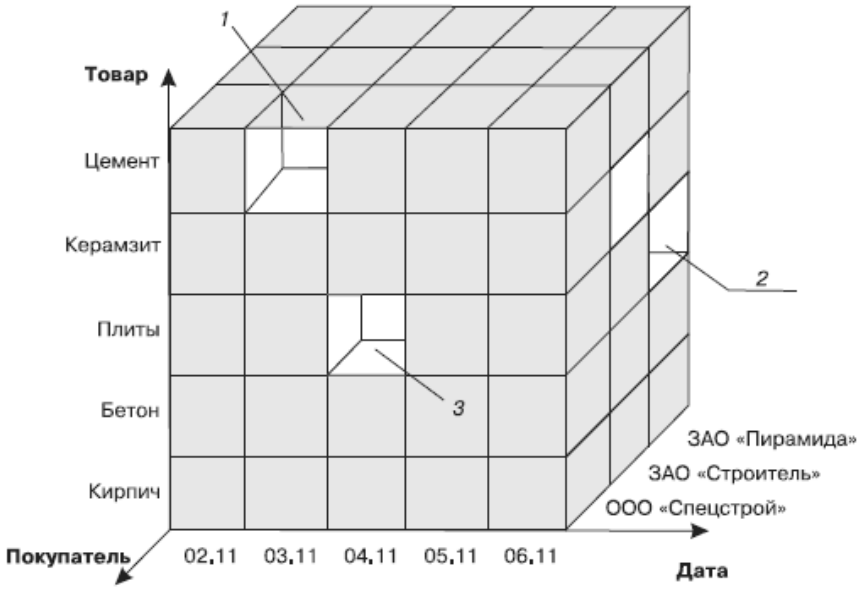
Измерения и факты в МХД

Измерение – это множество объектов одного или нескольких типов, организованных в виде иерархической структуры и обеспечивающих информационный контекст числового показателя. Измерение принято визуализировать в виде ребра (оси) многомерного куба. В качестве одного из измерений используется время.

Факт (мера) - это числовая величина, которая располагается в ячейках гиперкуба. Она количественно характеризует процесс.

Каждому набору значений измерений (например, «дата — товар — покупатель») будет соответствовать ячейка с фактами (мера, measure) связанными с данным набором.

Принцип организации многомерного куба



Преимущества и недостатки МХД

Основные преимущества:

- более наглядная структура, чем совокупность таблиц РХД.
- возможности построения аналитических запросов к системе более широки.
- уменьшение продолжительности поиска, т.к. агрегированные данные вычисляются предварительно и хранятся в многомерных кубах, поэтому тратить время на вычисление агрегатов при выполнении запроса не нужно.

Недостатки:

- требуется больший объем памяти.
- многомерная структура труднее поддается модификации; при необходимости встроить еще одно измерение требуется выполнить физическую перестройку всего многомерного куба.

Таким образом, применение МХД целесообразно в тех случаях, когда объем используемых данных сравнительно невелик, а сама многомерная модель имеет стабильный набор измерений.

Гибридные хранилища данных

Гибридные хранилища сочетают высокую производительность многомерной модели, и возможность хранить большие массивы данных, присущую реляционной модели.

Главным принципом построения ГХД является то, что детализированные данные хранятся в РХД, а агрегированные – в МХД.

ГХД оказываются наиболее подходящими, если данные, поступающие из OLTP-системы, имеют большой объем (несколько десятков тысяч записей в день и более) и высокую степень детализации, а для анализа используются в основном обобщенные данные.

Преимущества:

Построение OLAP-куба выполняется по запросу. Такой подход позволяет избежать взрывного роста данных. При этом можно достичь оптимального времени исполнения клиентских запросов.

Недостатком ГХД является усложнение администрирования из-за более сложного регламента его пополнения, поскольку при этом необходимо согласовывать изменения в реляционной и многомерной структурах.

Управление жизненным циклом информации

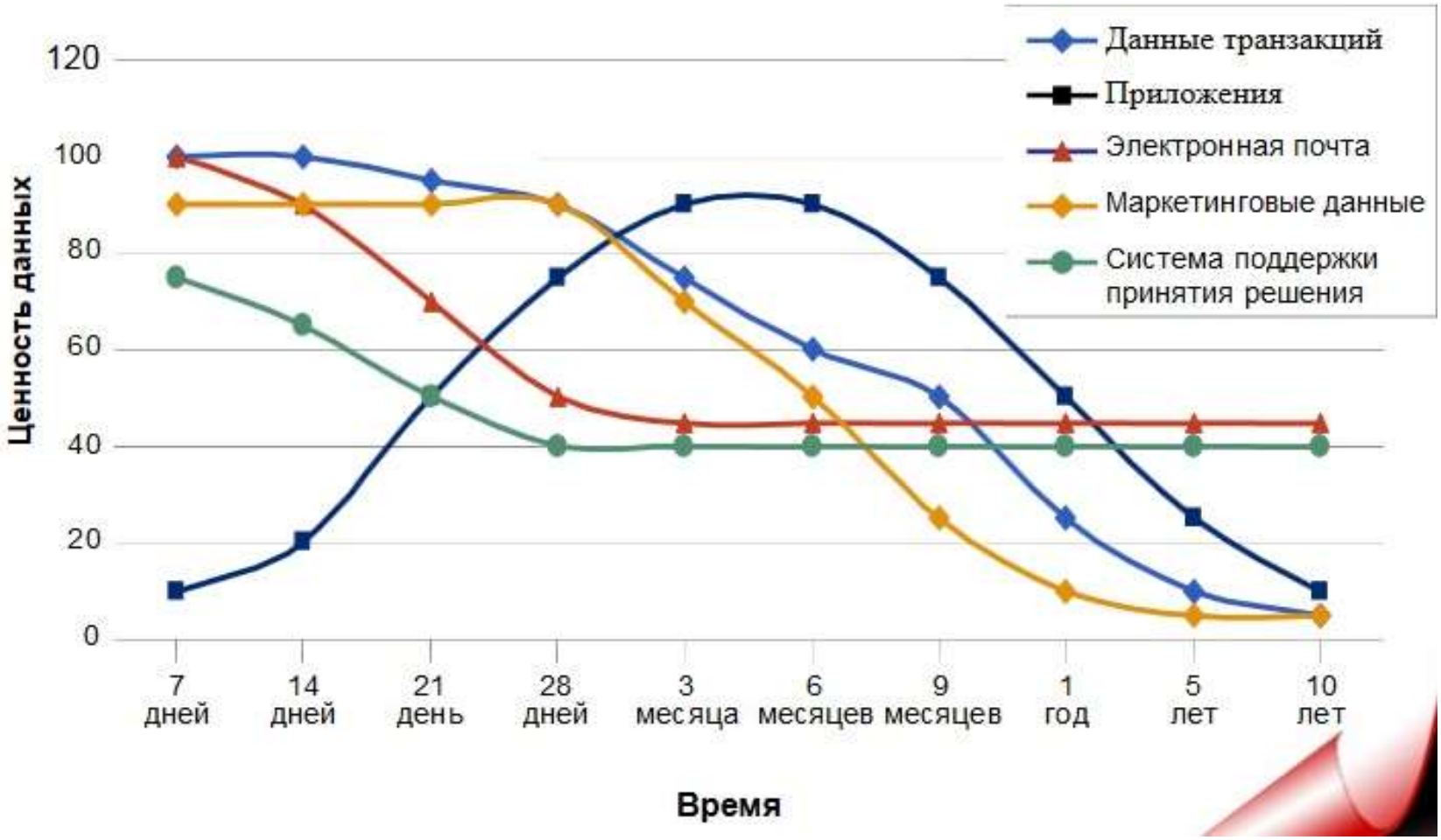
Жизненный цикл информации – это изменение ценности информации с течением времени. Например, в заказе на покупку ценность информации меняется с момента размещения заказа до истечения срока гарантии.

Управление жизненным циклом информации (Information Lifecycle Management – ILM) – это набор политик, процессов, практик, сервисов и инструментов, используемых для того, чтобы соотнести ценность информации с точки зрения бизнеса с наиболее подходящей и эффективной по стоимости инфраструктурой, начиная с момента создания информации и заканчивая ее размещением.

Проблемы клиента

- В настоящее время расходы на хранение данных составляют более 15% ИТ-бюджетов
- Ежегодно объемы данных растут более чем на 50%
- В большинстве случаев дисковые устройства хранения используются менее чем на 50%, 40% из них являются избыточными
- В мире существуют более 20 тысяч нормативных актов, включающих требования к хранению данных

Управление жизненным циклом информации



Многоуровневое хранение

Многоуровневое хранение – подход к определению различных уровней хранения для снижения затрат на хранение. Каждый уровень имеет различные степени защиты, производительности, частоты доступа к данным и пр. Информация хранится и передается между уровнями, исходя из ее ценности с течением времени.

Пример классификации информации, уровня обслуживания и политики жизненного цикла

Класс данных	Идентификация класса по аттрибутам						Регламент			
	приложение	владелец	файлы	путь	объем	дата создания	дата последнего доступа	скорость доступа	доступность	Политика жизненного цикла
критичные данные для бизнеса	SAPWebAS	*	*	/sap		*	*	15ms	99.99%	
	DB2	db2admin	*.dat	/db2		*	*	15ms	99.99%	
	-	domain\accounting *	*.xls	/home			<6 месяцев	40ms	99%	Перевести в класс важных файлов, если не было доступа в течение 6 месяцев
Важные файлы	-	domain\accounting *	*.xls	/home			>6 месяцев	120ms	99%	
Электронная Почта	LotusDomino	logistics*	*	*	>20MB	<6мес	>30дней			перевести в класс архив почты, если не было доступа в течение 30 дней и размер сообщения больше 20MB
Архив почты								3min	98%	
временые файлы	-	-	*.tmp, *.log, *.dmp, *.~*, tmp*	*		-	>7 дней	-	-	удалить
	-	-	*	/tmp		-	>7 дней	-	-	удалить
дублированные файлы	-	-					>30дней			удалить
ненужные файлы			*.mp3	*						удалить