Кафедра Прикладной математики

Института информационных технологий

РТУ МИРЭА

Дисциплина «Большие данные»

2022-2023 у.г.

Лекция №7
Технологии аналитики и визуализации больших данных

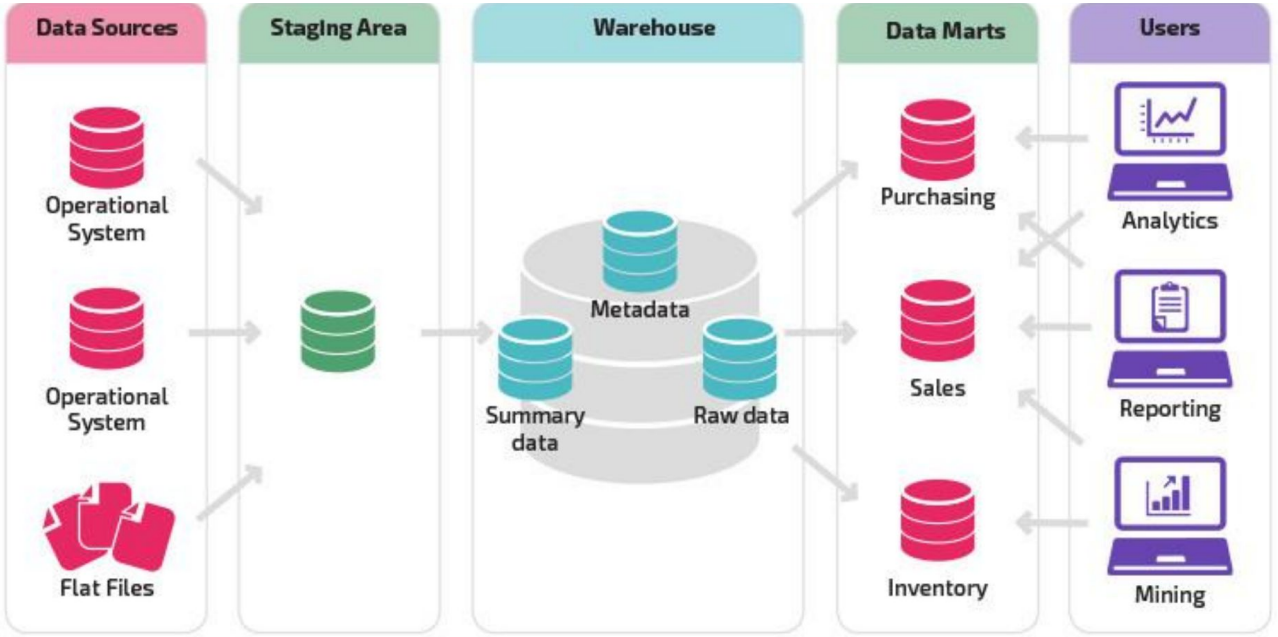
1. Технологии анализа данных: понятие аналитики данных, интеллектуальный анализ данных, математические методы анализа данных.
2. Аналитические базы данных. Организация хранилищ данных.
3. OLAP системы, витрины данных.
4. Системы параллельных вычислений (Massive Parallel Processing).
5. Способы графического представления информации: график разброса, график линий, столбчатая диаграмма, гистограмма, круговая диаграмма, карты, графовая визуализация, объемная визуализация, OLAP куб.
6. Системы визуализации данных: дашборды, динамические представления данных на основе фильтров

Часть 1. Аналитические базы данных. Организация хранилищ данных

Классический конвейер обработки больших данных

Классически поток обработки больших данных состоит из следующих стадий:

1. Сбор и извлечение данных из внешних источников.
2. Предобработка и структуризация данных.
3. Загрузка данных в долговременное хранилище данных.
4. Организация витрин данных.
5. Построение аналитических отченостей и моделирование.



Источники данных

Данные в поток обработки попадают из различных источников. Настроенный источник данных также называют **подключением**

Источники данных могут различаться как по способу доступа к информации, так и по возвращаемой структуре данных.

Среди источников выделяют:

1. программный интерфейс приложения (API);
2. SQL и NoSQL базы данных;
3. файлы данных;
4. потоковый сервис.

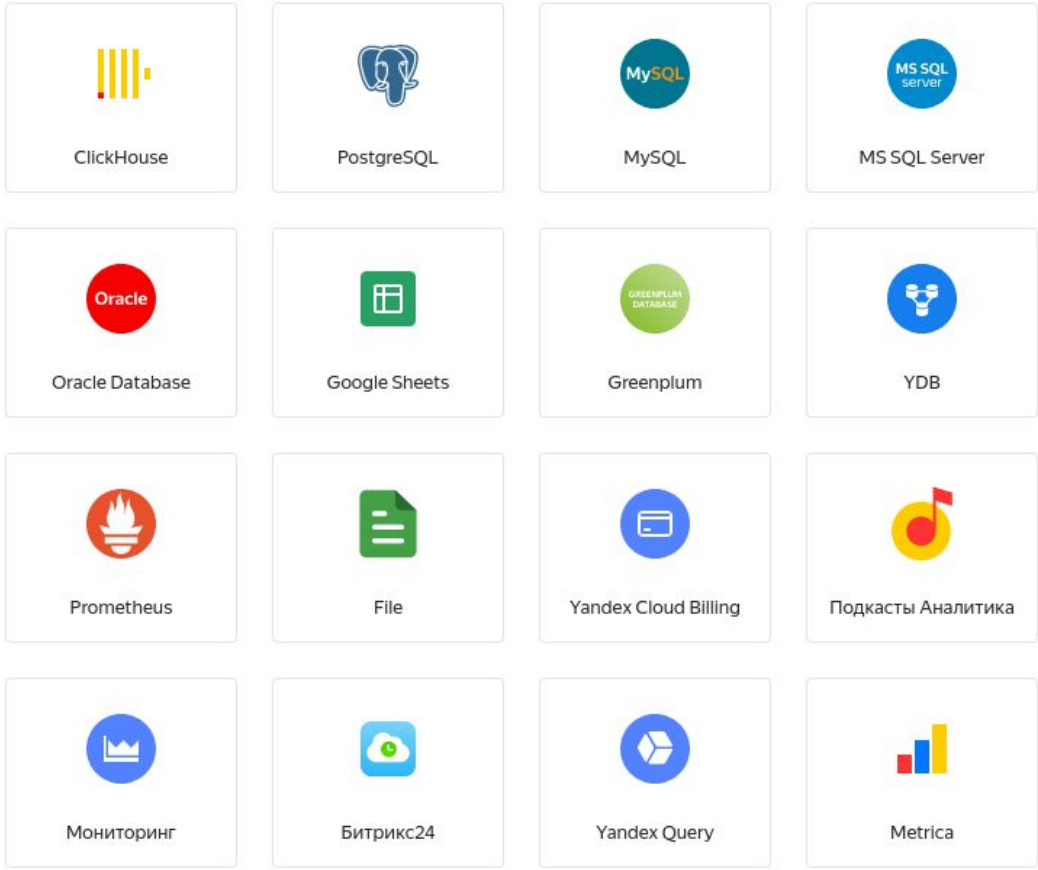


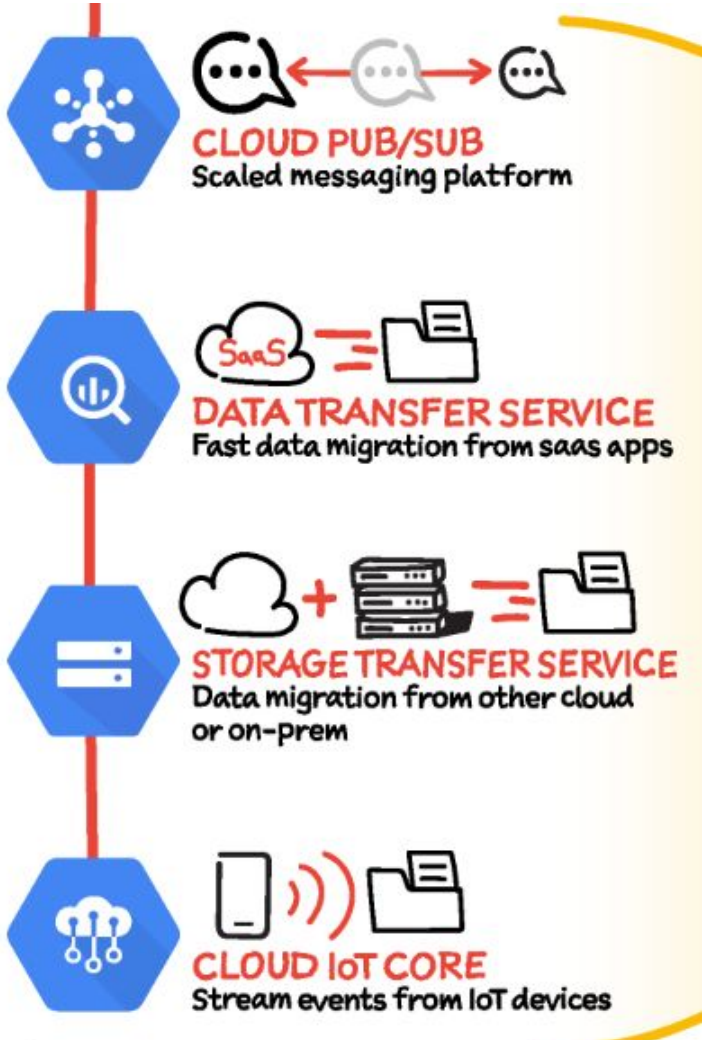
Рисунок. Источники данных в Yandex Data Lens

Стадия сбора/извлечения данных

Для переноса данных из подключения используют **инструменты миграции данных**.

Они позволяют осуществлять перенос из различных:

- местоположений (локальная сеть, облако и др.);
- форматов (xlsx, json, csv, mp3, mdb и др.);
- приложений (сторонние сервисы: YouTube, Google Ads, Amazon S3, Teradata, ResShift и др.; собственные приложения).



Стадия сбора/извлечения данных

На стадии **извлечения** и **сбора** данных ставится задача загрузки данных из нескольких внешних или внутренних источников организации в поток обработки данных.

Проблемы, возникающие на этапе сбора:

- количество внешних источников данных,
- разнородность формата и интерфейса предоставления данных,
- согласованность времени сбора данных,
- пропуски в данных и проблемы при интерпретации результатов,
- объем собираемых данных,
- консолидация данных.

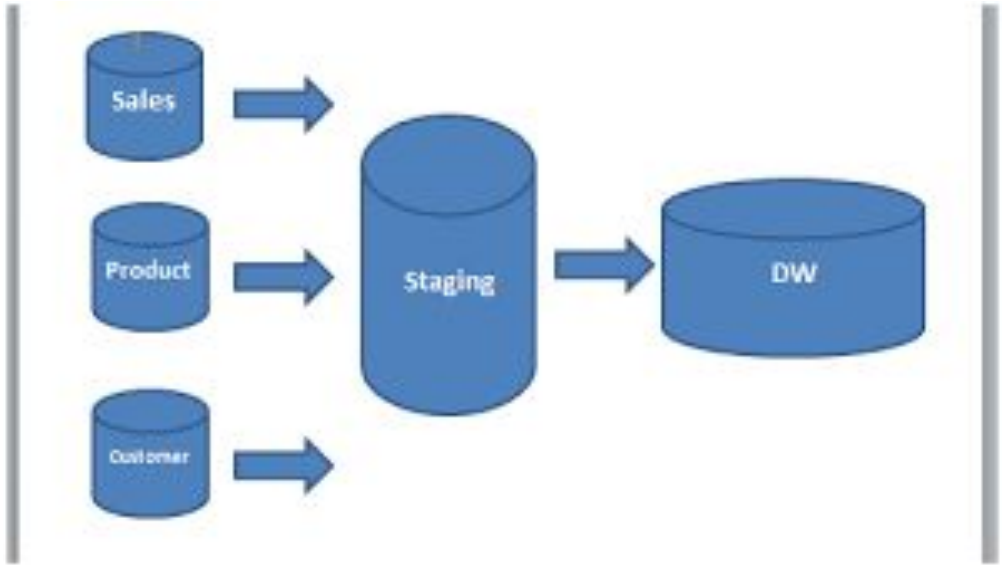
Стадия предобработки данных

Стадия предобработки данных решает задачу преобразования полученных данных в необходимую форму.

Среди направлений предобработки данных выделяют:

1. Этап очистки данных
2. Этап оптимизации данных

В результате данные представлены в формате структуры для хранения в долговременном хранилище и в форме, подходящей для анализа.



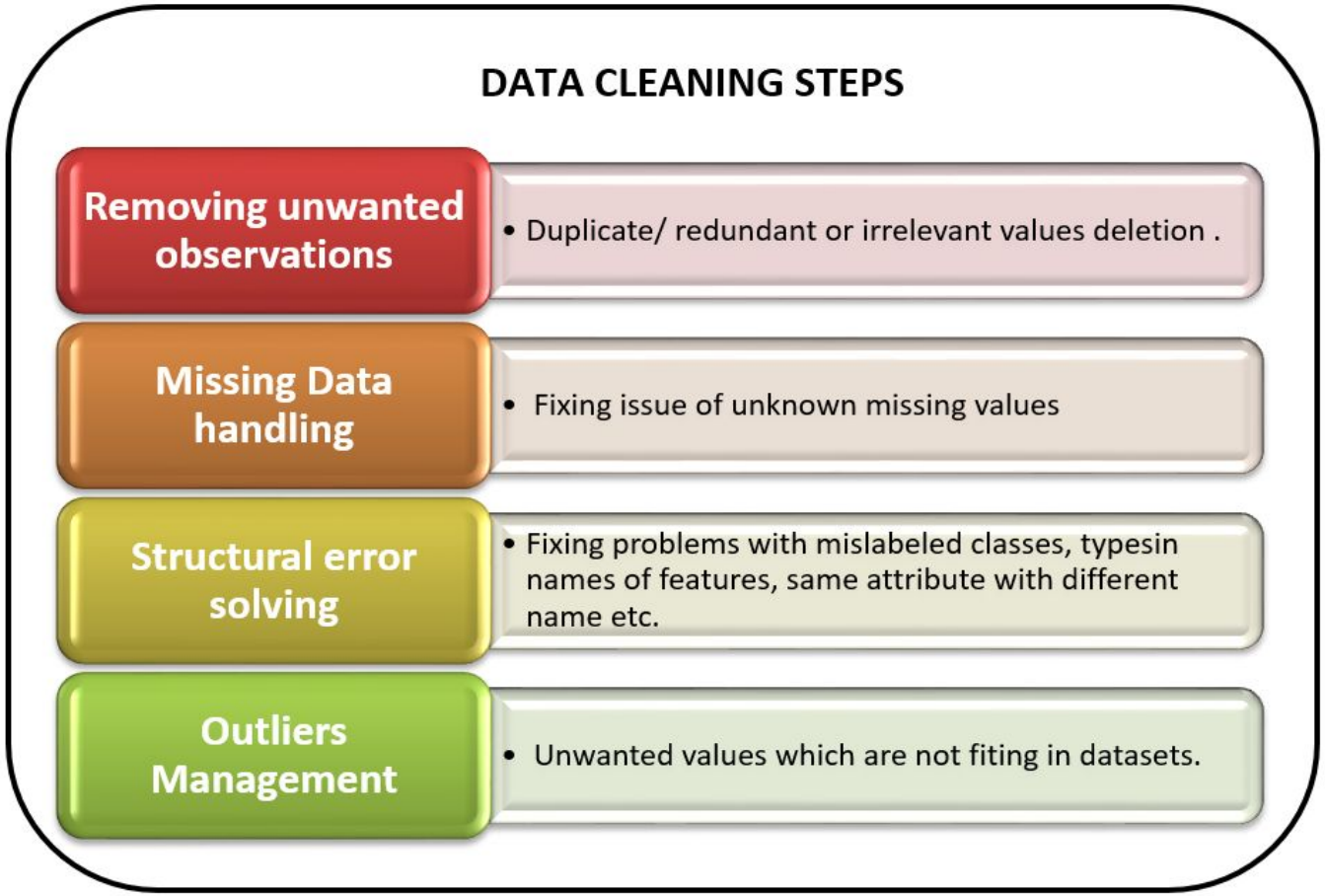
Очистка данных

Очистка данных — этап удаления нерелевантных значений показателей или записей данных с нетипичными значениями.

В очистке данных выделяют стадии:

- обработка пропусков,
- удаление дубликатов и противоречий,
- обработка выбросов (нетипичных значений)
- восстановление структуры данных,
- верификация целостности данных.

После этапа очистки данные готовы к загрузке в хранилище данных.

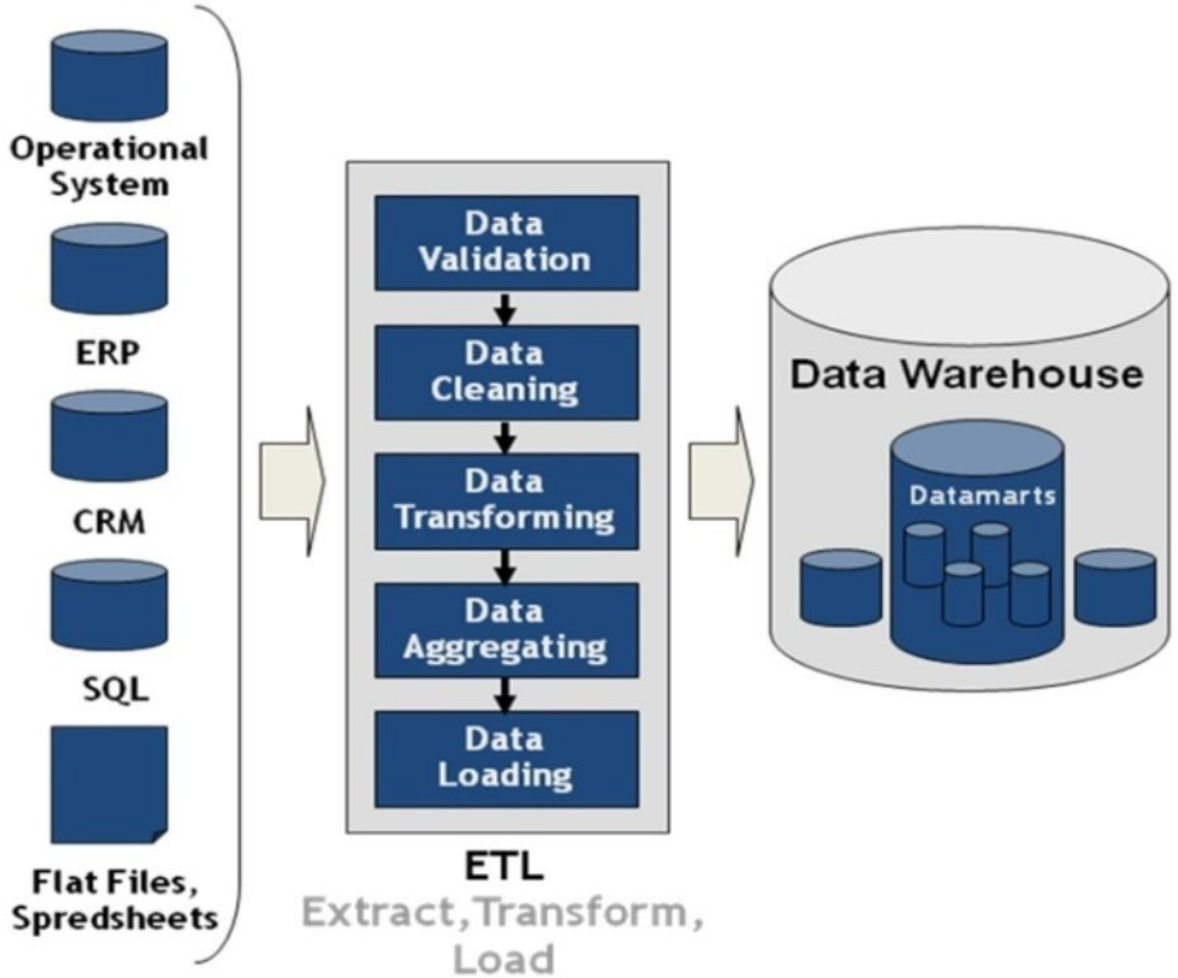


Стадия загрузки данных в хранилище

Процесс загрузки заключается в переносе данных из промежуточных таблиц в структуры хранилища данных (ХД).

После завершения загрузки выполняются дополнительные операции над данными (тестирование), только что загруженными в ХД. К ним относятся: переиндексация, верификация данных и т. д.

Если тестирование показало, что несоответствия, позволяющие заподозрить потерю или недостоверность данных, отсутствуют, то можно считать загрузку данных в ХД успешной.



Проблемы при загрузке данных в хранилище

Одной из основных проблем данного шага является то, что далеко не всегда данные загружаются полностью: в загрузке некоторых записей может быть отказано. Отклонение записей происходит по следующим причинам:

- на этапе преобразования данных не удалось исправить все критичные ошибки, которые блокируют загрузку записей в ХД;
- некорректный порядок загрузки данных;
- внутренние проблемы ХД, например недостаток места в нем;
- прерывание процесса загрузки или остановка его пользователем.

Понятие хранилища данных

Хранилище данных — это цифровая система хранения, которая выполняет объединение и согласование больших объемов данных из разных источников.

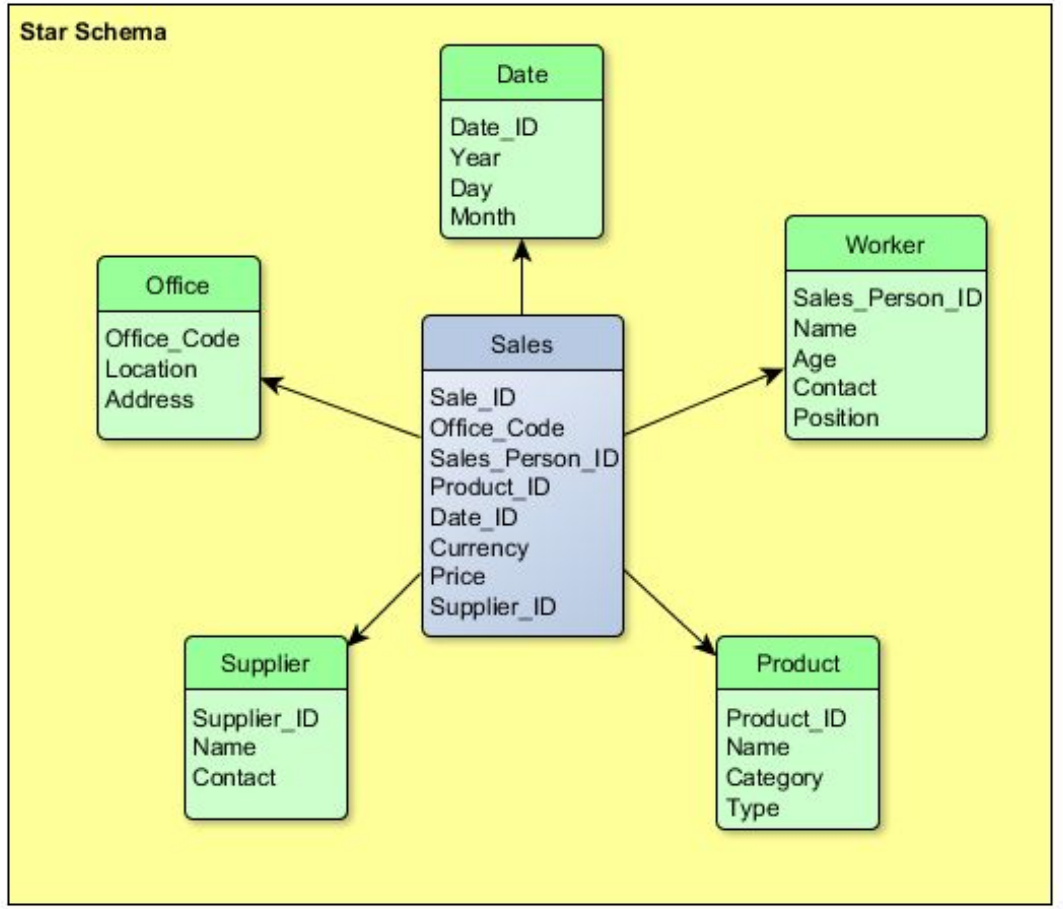
Оно предоставляет данные для бизнес-аналитики, отчетов и анализа, а также обеспечивает поддержку нормативных требований.

С помощью ХД компании превращают свои данные в ценную информацию и принимают взвешенные решения на основе данных.

Свойства хранилища данных

Хранилище данных должно обладать следующими свойствами:

- Предметная ориентированность – создается с ориентацией на решаемую проблему хранения и анализа;
- Консолидированность – данные объединены в схемы для удобной итеграции и решения задач
- Энергонезависимость – данные хранятся на энергонезависимых носителях
- Поддержка изменений во времени – настроенность на постоянную поддержку внесения новых данных

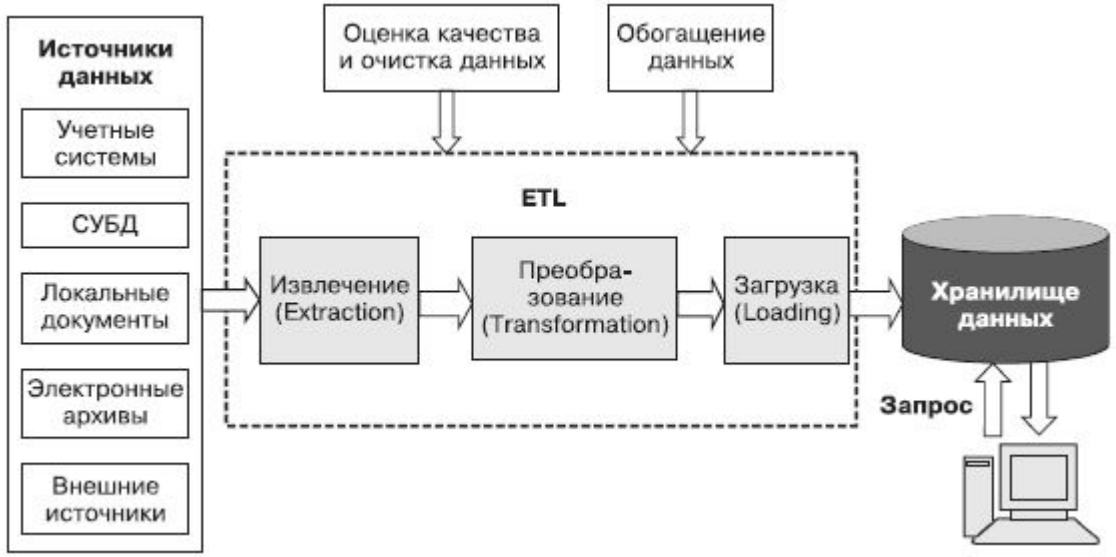


Основные задачи консолидации данных

Задача консолидации данных и заключается в соблюдении ранее упомянутых свойств.

В процессе консолидации данных решаются следующие задачи:

- выбор источников данных;
- разработка стратегии консолидации;
- оценка качества данных;
- обогащение;
- очистка;
- перенос в хранилище данных.



Задачи хранилища данных

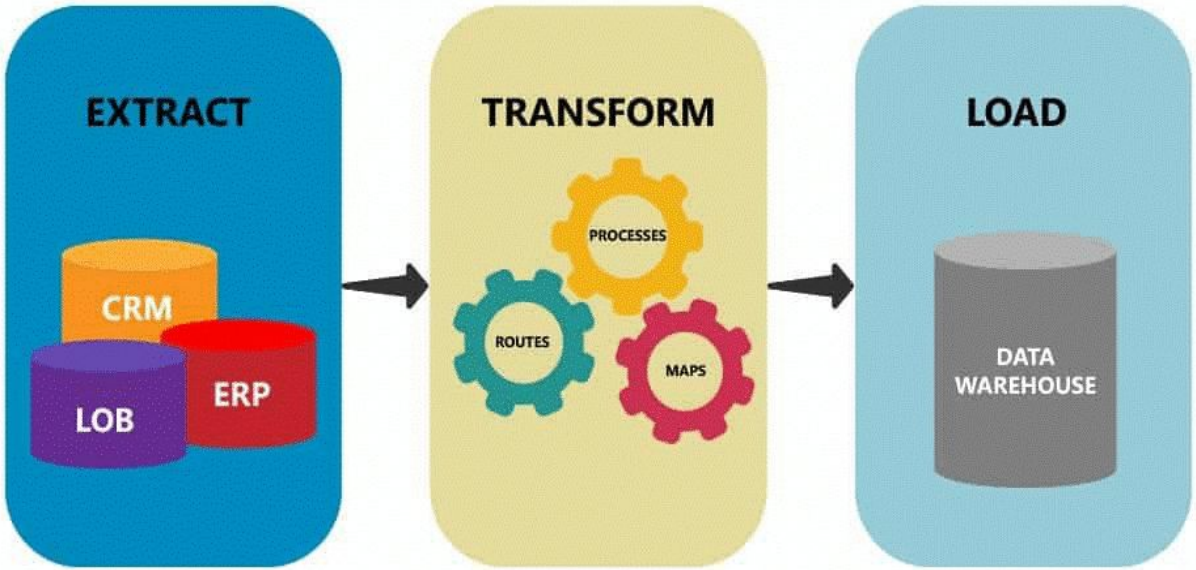
Хранилище данных решает ряд важных задач:

- предоставление оперативного доступа и хранение информации (структурированной и нет);
- расширение и масштабирование данных при растущем увеличении объема информации;
- функции безопасности (отражение локальных и сетевых атак, борьба с компроментацией данных);
- репликация (дублирование данных);
- виртуализация (распределение трафика между пользователями);
- сжатие данных.

ETL (Extract-Transform-Load)

Процесс **ETL** представляет собой комплекс операций, реализующих процесс переноса первичных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных.

ETL-приложения извлекают информацию из одного или нескольких источников, преобразуют ее в формат, поддерживаемый системой хранения и обработки, которая является получателем данных, а затем загружают в нее преобразованную информацию.



ETL - Extract, Transform, Load

ETL (Extract-Transform-Load)

Любая ETL-система должна обеспечивать выполнение трех основных этапов процесса переноса данных:

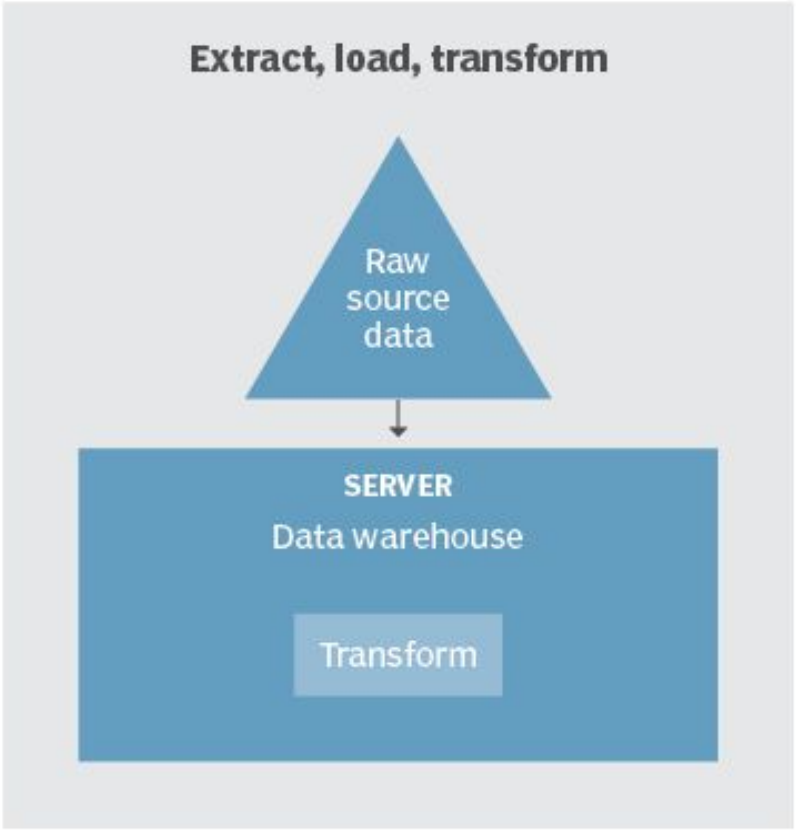
- **Извлечение данных** — на этом шаге данные извлекаются из одного или нескольких источников и подготавливаются к этапу преобразования.
- **Преобразование данных** — производится преобразование форматов и кодировки данных, а также их интеграция и очистка;
- **Загрузка данных** — запись преобразованных, интегрированных и очищенных данных в соответствующую систему хранения.

Все операции над данными в процессе ETL производятся в так называемой промежуточной области, где для этого создаются временные таблицы.

ELT (Extract–Load–Transform)

ELT — это процесс переноса данных из разнородных источников в хранилище данных с целью их дальнейшего анализа. В целом, процесс ELT выполняет те же функции, что и ETL с той только разницей, что этапы загрузки и преобразования меняются местами.

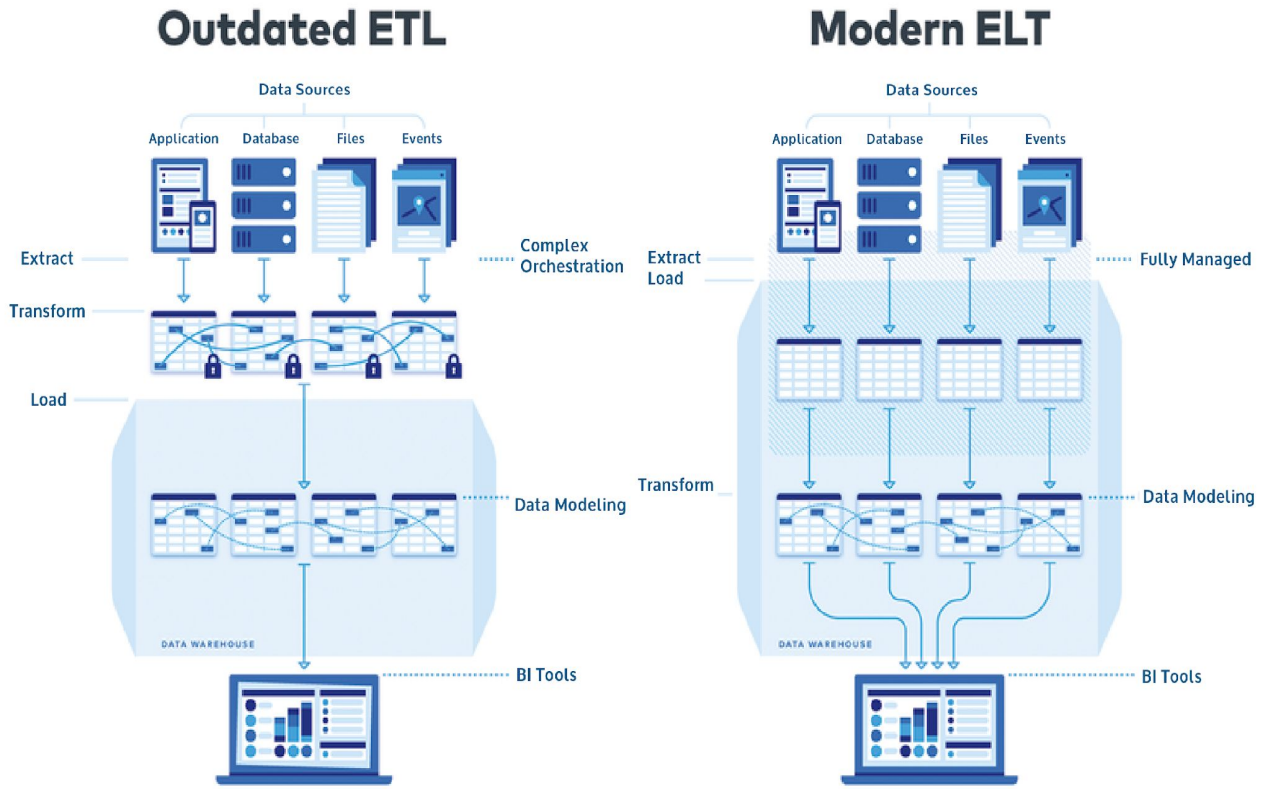
ELT извлекает данные из исходных местоположений, но вместо перемещения их в промежуточную область для преобразования, загружает необработанные данные непосредственно в приемник, где их можно преобразовать по мере необходимости в соответствии с конкретными целями и задачами анализа.



Основные различия ETL и ELT

Помимо порядка проведения операций, между процессами ETL и ELT встречаются следующие различия:

- технология ELT новее ETL;
- ELT позволяет работать с неструктурированной, слабоструктурированной и структурированной информацией. ETL только со структурированной;
- в ELT данные загружаются в хранилище сразу после извлечения. Их преобразование производится по мере необходимости. Это экономит время ожидания загрузки данных.



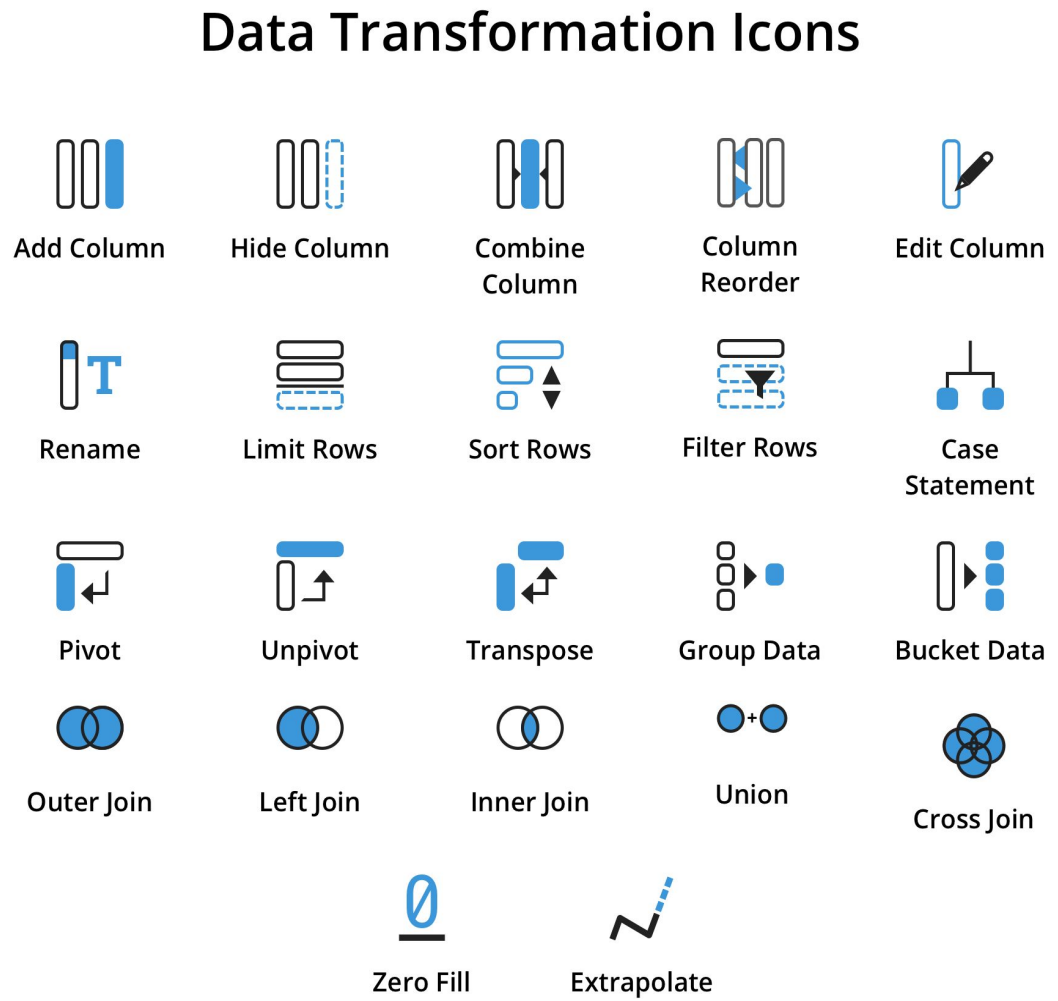
Оптимизация данных

Оптимизация данных — этап преобразования данных в формат, удобный для анализа.

В оптимизации данных выделяют:

- выбор релевантных признаков (столбцов),
- снижение размерности данных,
- агрегация данных,
- выделение новых признаков (инженерия признаков).

После этапа оптимизации данные готовы к обработке или загружаются в витрины данных.

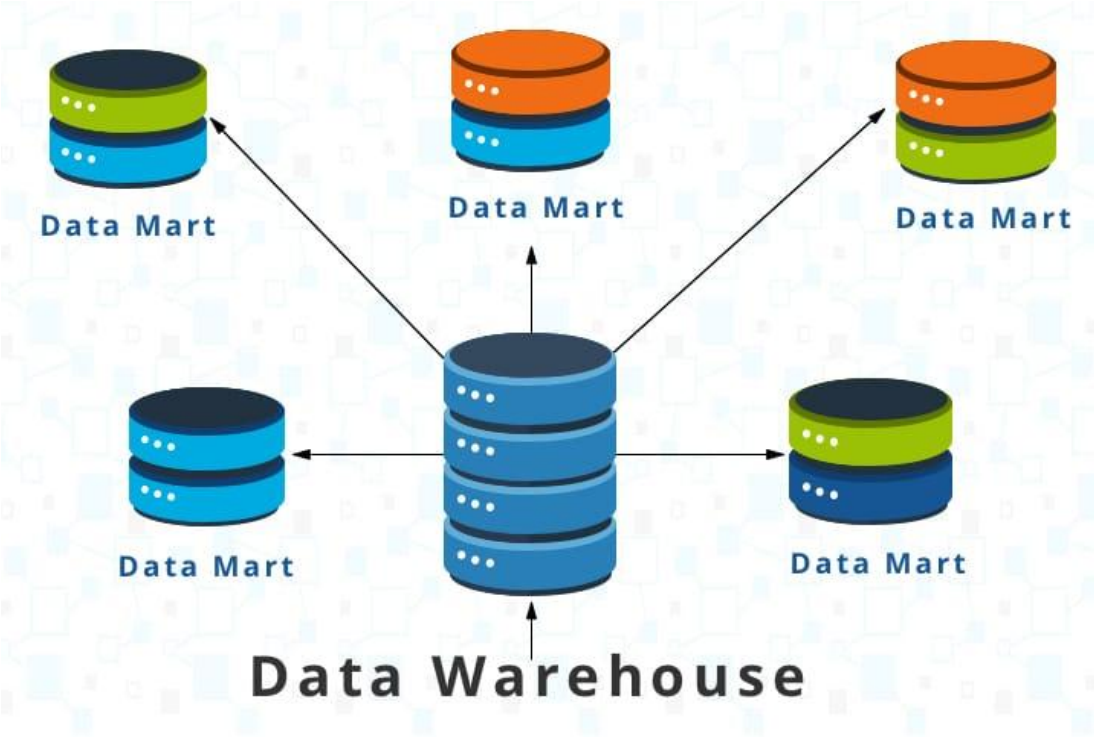


Витрины данных

Витрина данных — это часть хранилища данных, секционированная для отделов или направлений бизнеса (например, продажи, маркетинг или финансы). Витрина строится из данных, которые запрашиваются чаще других или нужны для выполнения бизнес-задач.

В одном хранилище данных часто развертывается несколько витрин.

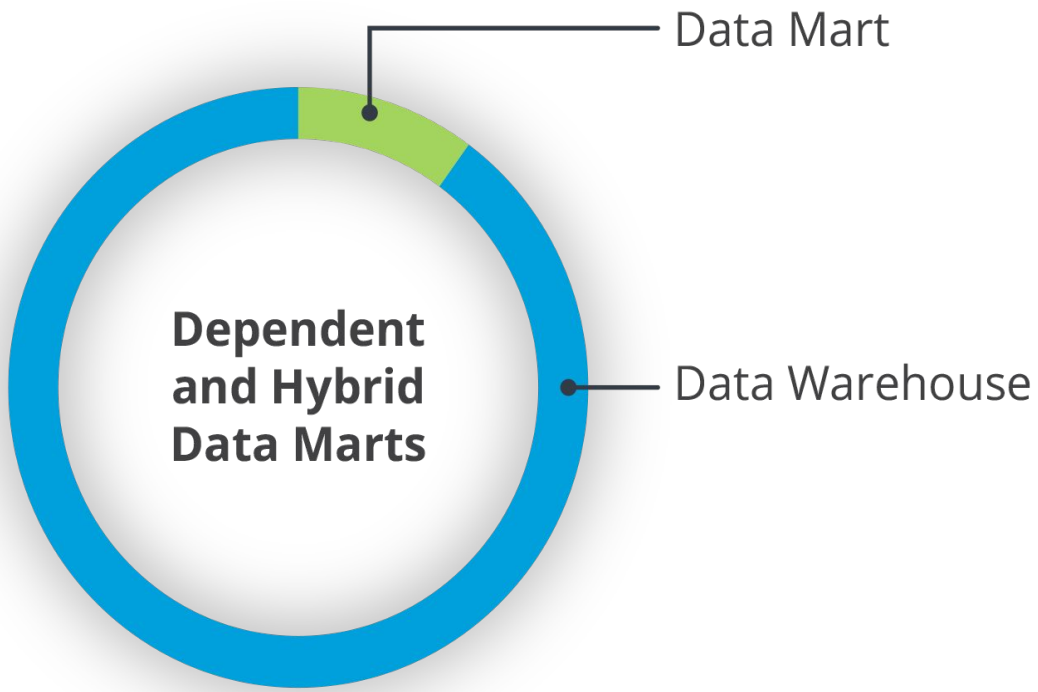
Витрины данных хранятся в виде многомерной схемы, которая служит основой для анализа пользователями хранилища данных. Две основные архитектуры витрин данных — это звезда, и снежинка. Встречается также реализация в виде многомерного OLAP-куба.



Типы витрин данных

Существует три основных типа витрин данных. Разница между ними определяется их отношением к хранилищу данных и источникам данных, использованным для их создания. К ним относятся:

- **зависимые** — зависят от информации, извлеченной из корпоративных хранилищ данных;
- **независимые** — не связаны с хранилищем данных. Данные извлекаются из внутренних или внешних источников и загружаются в независимый репозиторий витрин данных.
- **гибридные** — объединяют источники первичных данных из существующего хранилища данных и других внешних источников данных..

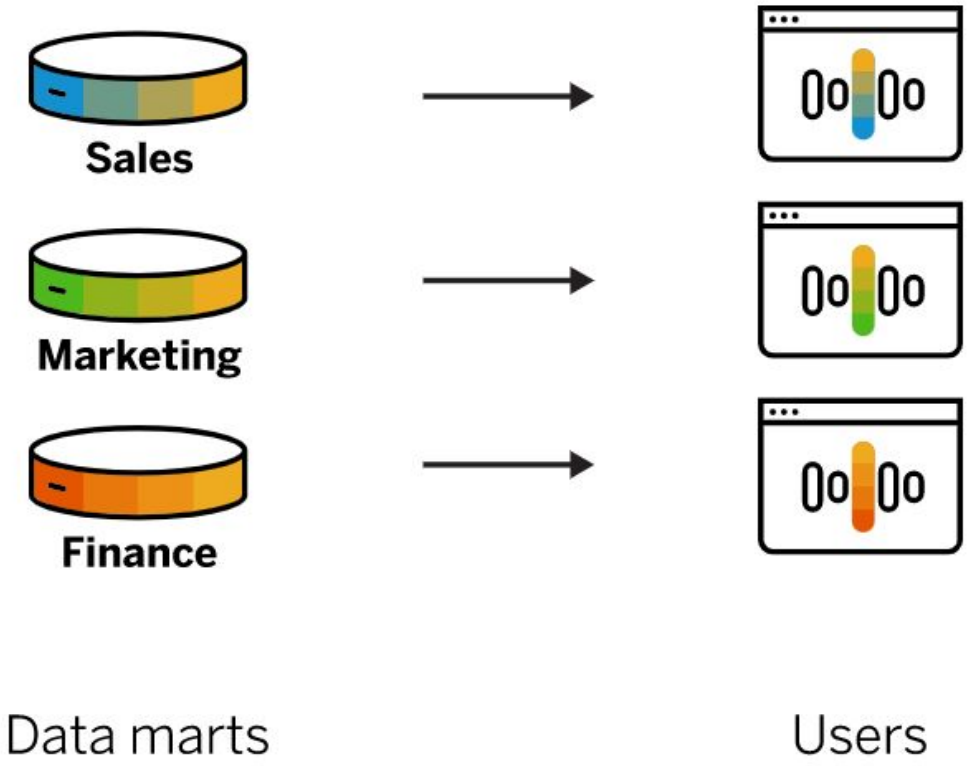


Стадия формирования витрин данных

После загрузки данных в хранилище формируем витрины данных по следующему алгоритму:

1. Создаем представления данных (временные таблицы на осове существующих) из баз данных хранилища;
2. Определяем архитектуру схемы витрин;
3. Сегментируем данные по отделам (тематикам);
4. Запрещаем доступ к хранилищу данных для всех пользователей по умолчанию. И далее предоставляем пользователям доступ только к тематическим витринам.

На этом шаге могут быть исключения — некоторым пользователям будет необходимо иметь права доступа как к хранилищу, так и к витринам.

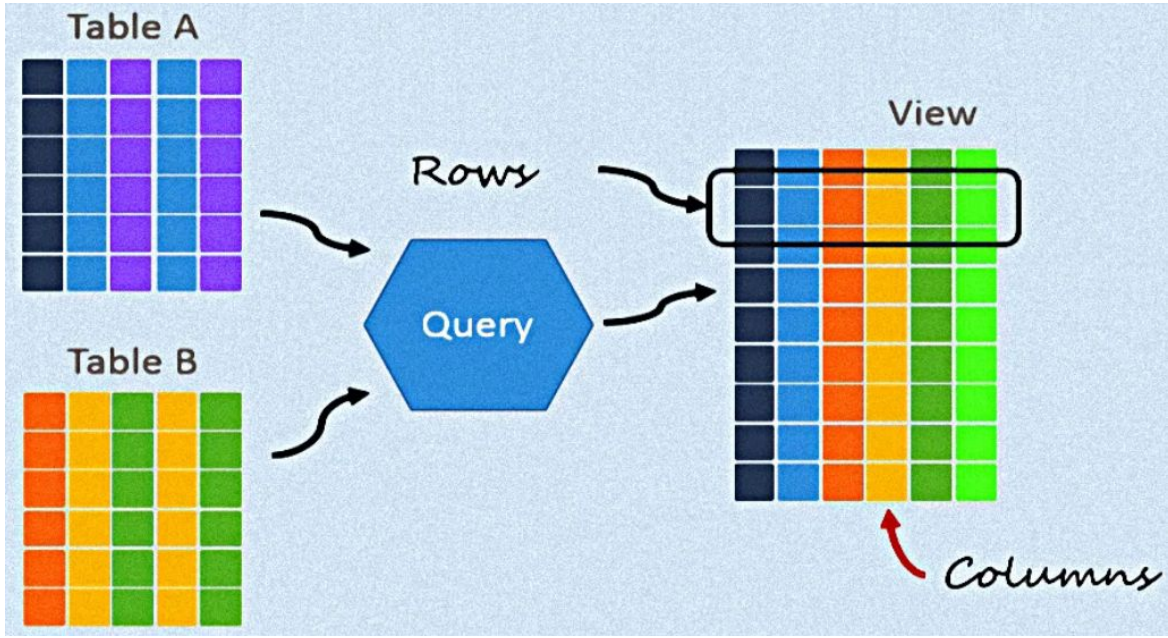


Стадия формирования витрин данных

Представления данных (VIEW) — специальные конструкции в реляционных СУБД, позволяющие хранить предметно-ориентированные таблицы, составленные из исходной схемы базы данных с использованием запроса (SELECT).

Данные представления могут как храниться в виде запроса и исполняться «на лету», так и храниться в виде отдельной таблицы, связанной с исходными.

Вставка данных в представления может сопровождаться триггером на вставку данных в базу данных таблиц.

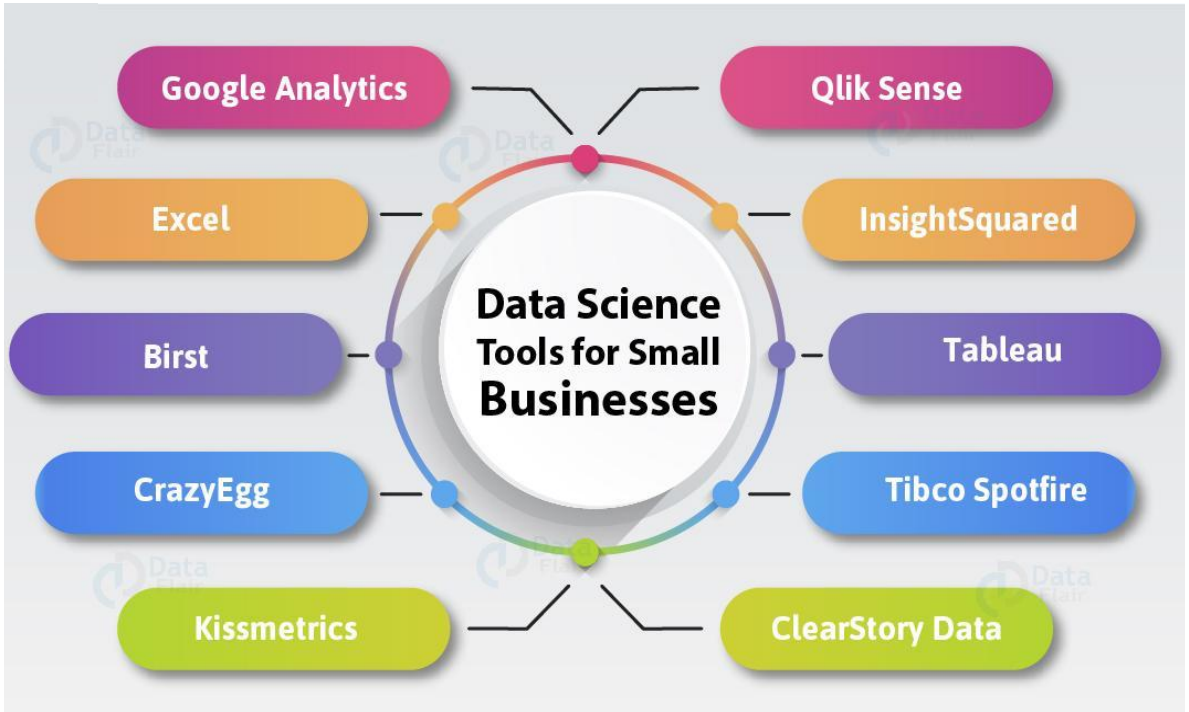


Стадия формирования аналитических отчетностей

Цель ведения аналитической отчетности — обнаружить проблему или возможность и объяснить, как это влияет на организацию, и как организация должна реагировать. Во многих случаях также ожидают рекомендацию, основанную на проведенном анализе.

Примеры аналитической обработки данных:

- инвестиционный анализ;
- финансовый анализ показателей деятельности;
- анализ вероятности банкротства;
- ABC -анализ;
- сегментный анализ;
- факторный анализ.



Часть 2. Технологии анализа данных. Аналитика данных

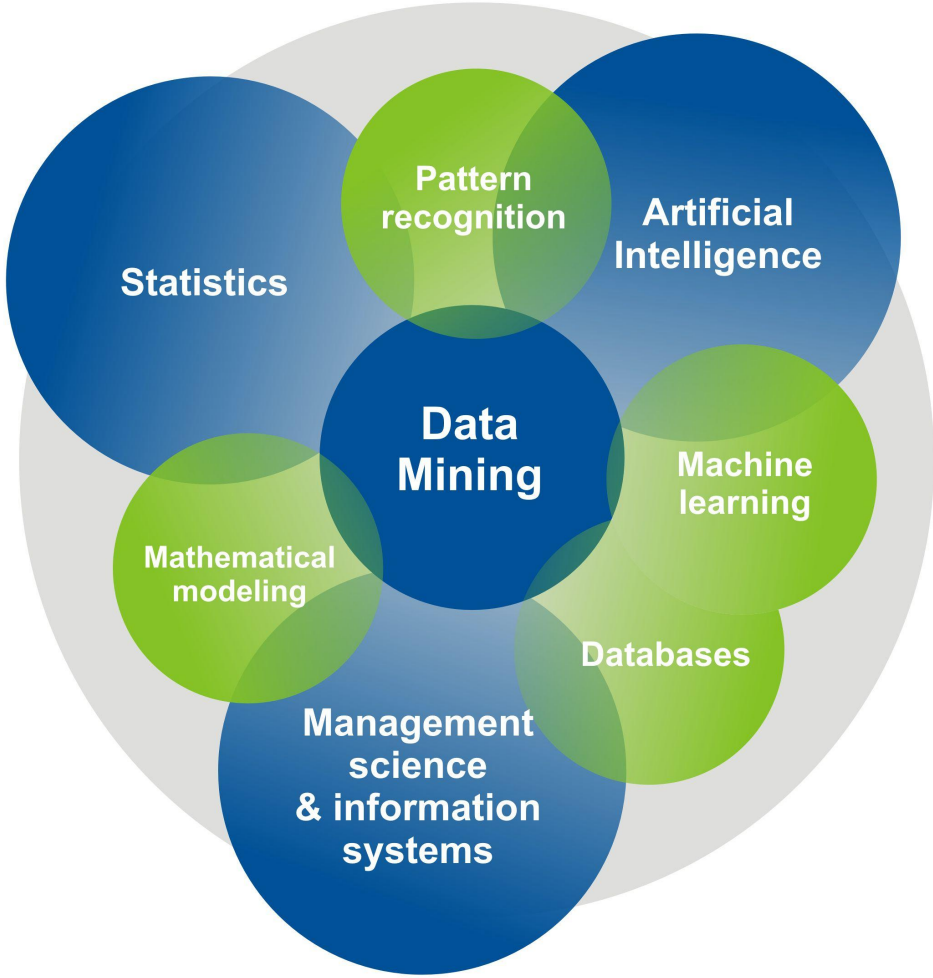
Понятие аналитики данных

Аналитика данных — область занимающаяся преобразованием «сырых» данных в практические выводы.

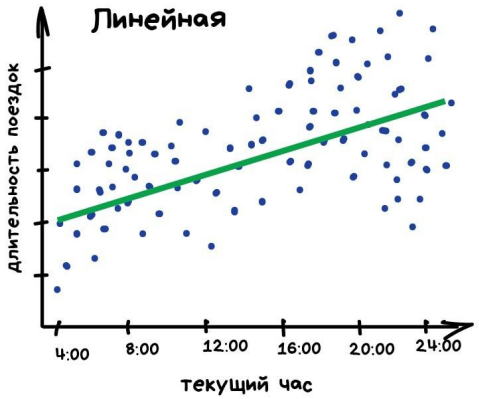
Использует определенный набор процессов и технологий для поиска тенденций и закономерностей в данных

Задача аналитики данных — получение выгоды из данных путем их преобразований, визуализации и построения описательных и предиктивных моделей, а также их оценки.

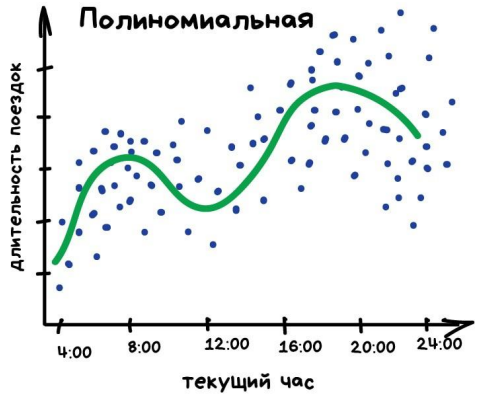
Интеллектуальный анализ данных



Применение


Линейная
длительность поездок
4:00
8:00
12:00
16:00
20:00
24:00
текущий час

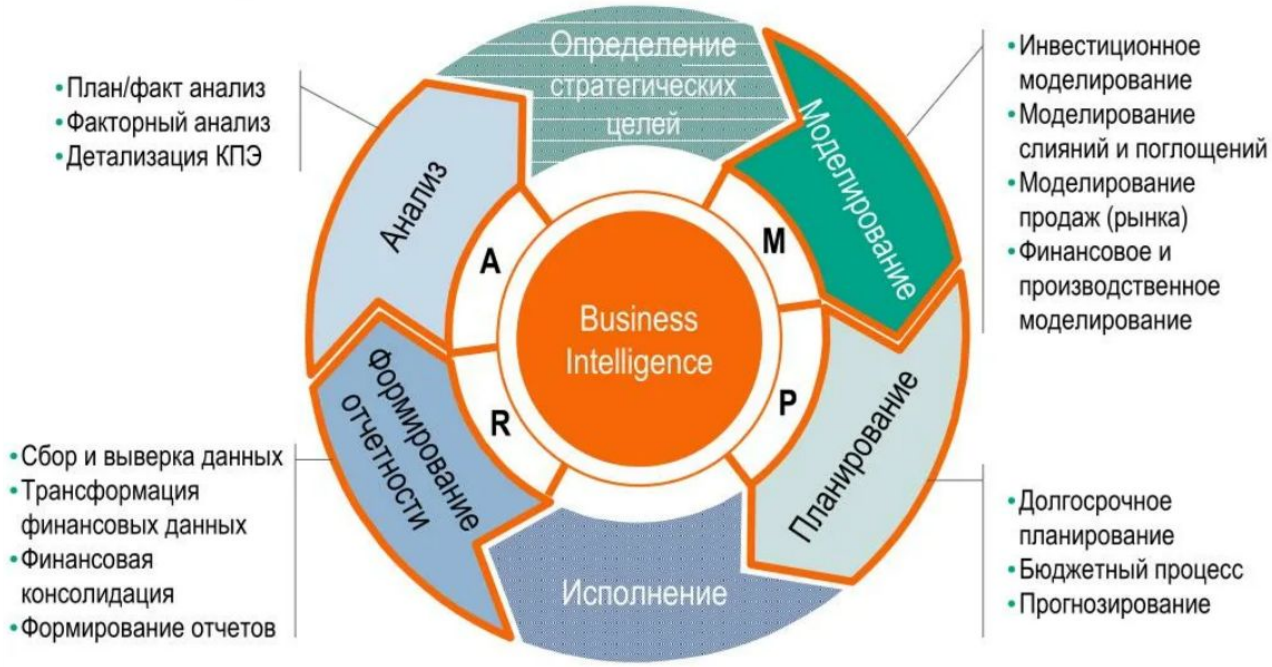

Возраст автовладельца > 40
Да
Нет
Место эксплуатации
Езда без аварий
Город
Сельская местность
Да
Нет
Стаж более 10 лет
Страховать
Страховать
Тип автомобиля
Да
Нет
Минивэн
Спорткар
Страховать
Не страховать
Страховать
Не страховать


Полиномиальная
длительность поездок
4:00
8:00
12:00
16:00
20:00
24:00
текущий час


1. Постановка задачи
2. Сбор данных
3. Подготовка данных
4. Изучение данных
5. Построение моделей
6. Обучение моделей на данных
7. Исследование и валидация моделей
8. Визуализация
9. Развертывание и обновление моделей
SQL

Задачи бизнес-аналитики

- Изучение и формализация предметной области клиента
- Оптимизация бизнес-процессов
- Разработка характеристик IT продукта
- Внедрение новых характеристик продукта
- Расчет метрик качества принятия решений
- Аналитическая отчетность и визуализация

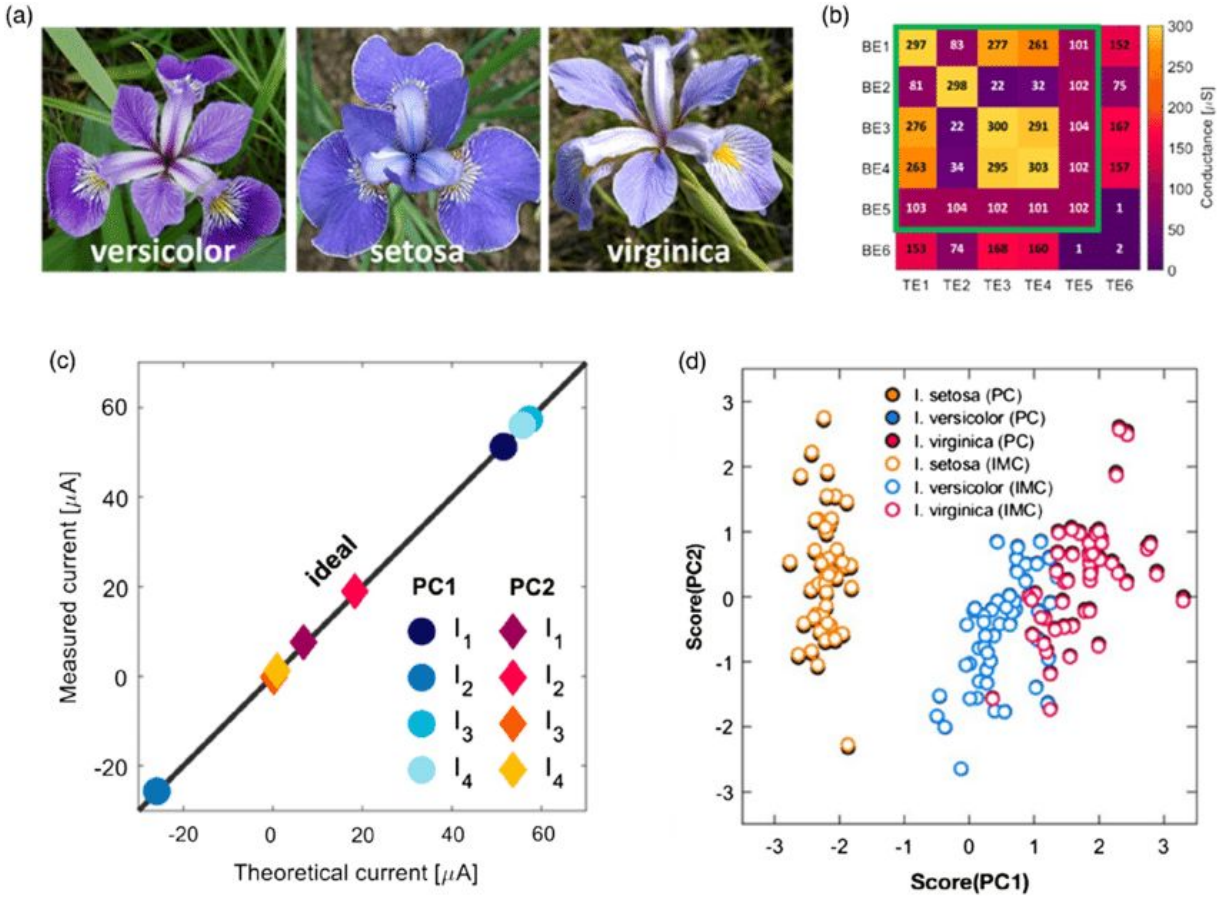


Задачи машинного обучения

Задачи машинного обучения заключаются в получении прогноза или вывода, восстанавливая закономерность исходных данных.

Виды задач:

- Обучение с учителем (в данных присутствует истинный ответ на пример)
- Без учителя (ответа на пример нет, закономерности ищутся в данных)



Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

Классификация

«Разложи носки по цвету»

Предсказать значение

Регрессия

«Разложи галстуки по длине»

Без учителя

Разделить по схожести

Кластеризация

«Разложи похожие вещи по кучкам»

Найти зависимости

Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

Выявить последовательности

Ассоциация

«Найди какие вещи я часто ношу вместе»

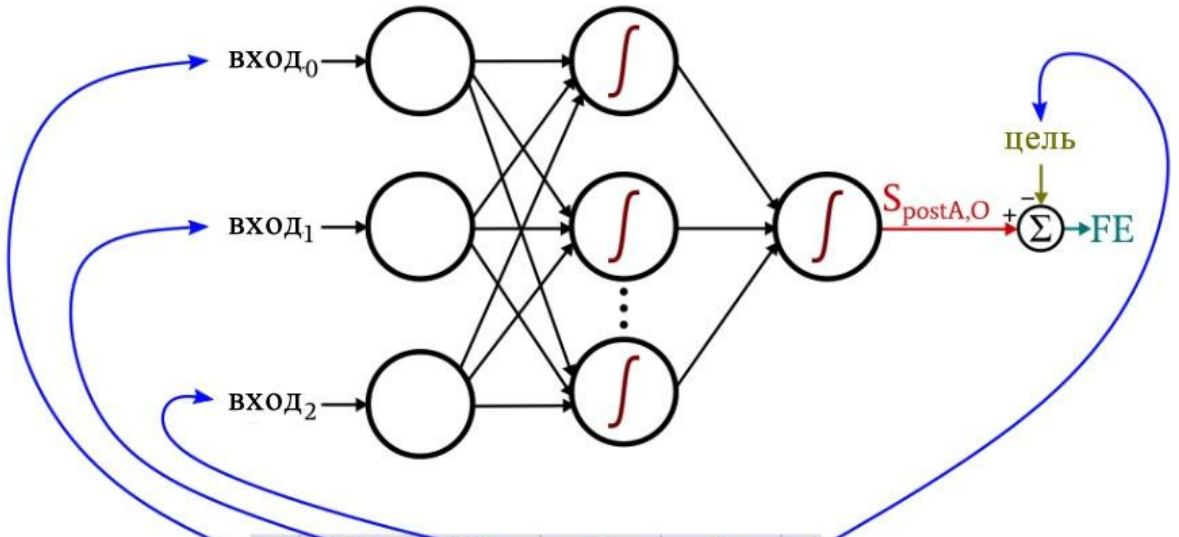
Обучение с учителем

Алгоритмы машинного обучения, настраивающие свои параметры по методу минимизации ошибок между предсказаниями на данных и истинными метками

Обучение на основании множества примеров вида «известный вход – известный выход»

Примеры задач:

- Регрессия
- Классификация
- Ранжирование



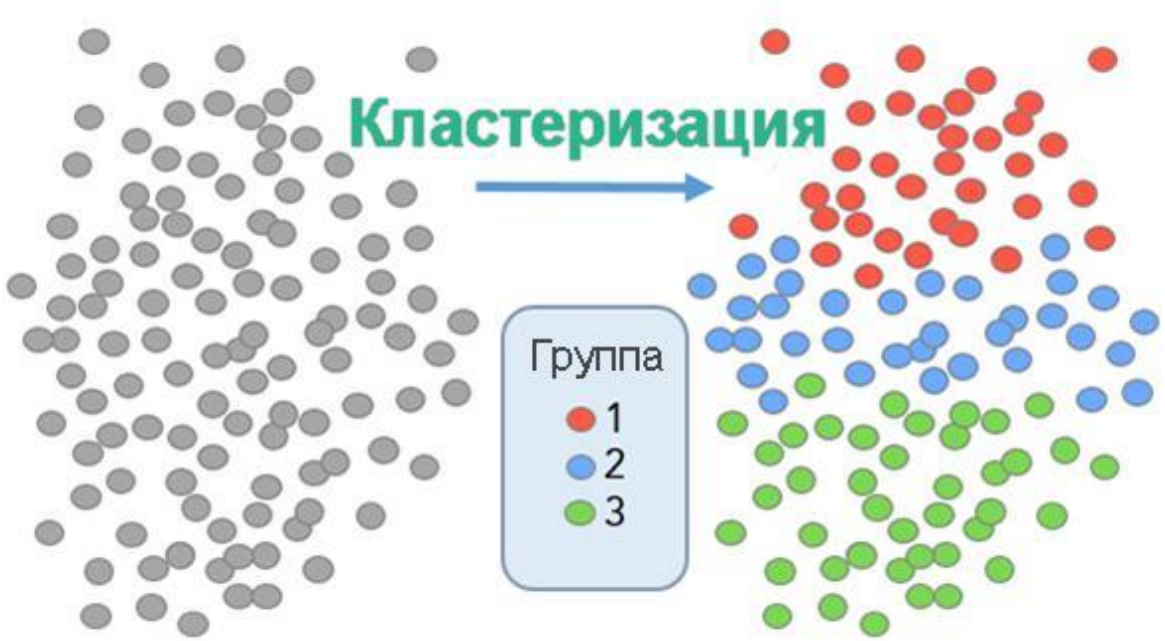
	A	B	C	D
1	input_0	input_1	input_2	output
2	-4.5	4.5	-1	0
3	-4.5	-1.5	-5	0
4	4	4	-0.5	1
5	2.5	4	-2.5	1
6	-3	2.5	-5	0
7	5	-1.5	5	0

Обучение без учителя

Обучение без учителя – процесс при котором система учится находить закономерности в данных без правильных ответов (истинных меток)

Примеры задач:

- Кластеризация
- Обобщение
- Обнаружение аномалий

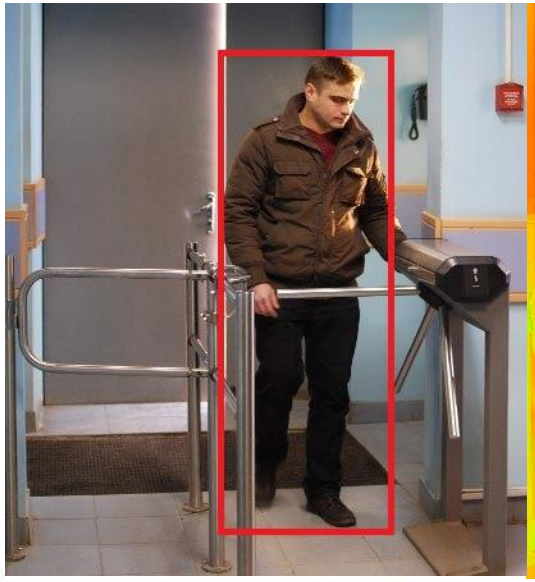


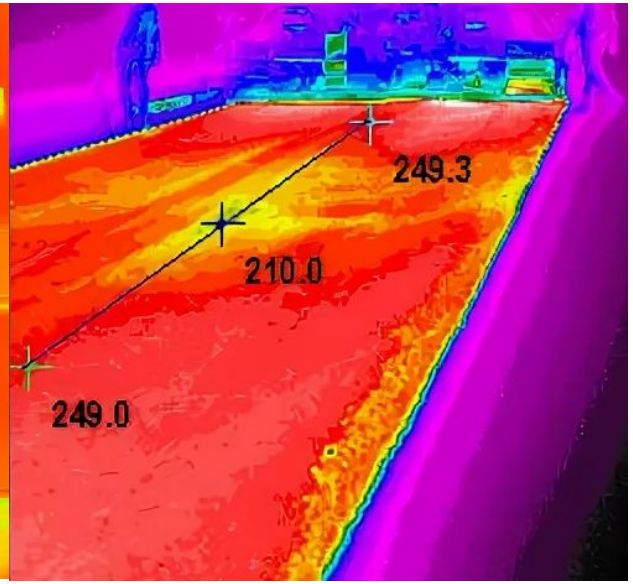
Задачи глубокого обучения

Подкласс задач машинного обучения с учителем при решении которых используются нейронные сети.

Примеры задач:

- Распознавание речи
- Компьютерное зрение
- Обработка естественного языка





TEXT PROMPT an illustration of a baby daikon radish in a tutu walking a dog

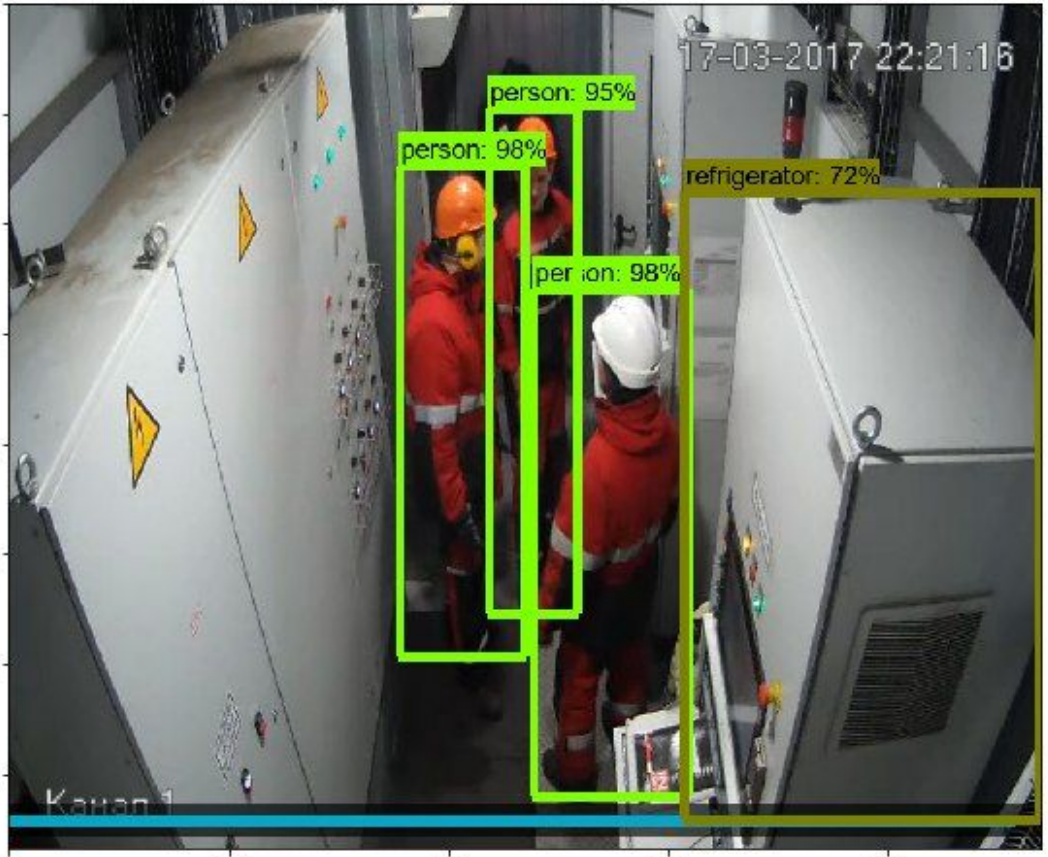
AI-GENERATED IMAGES

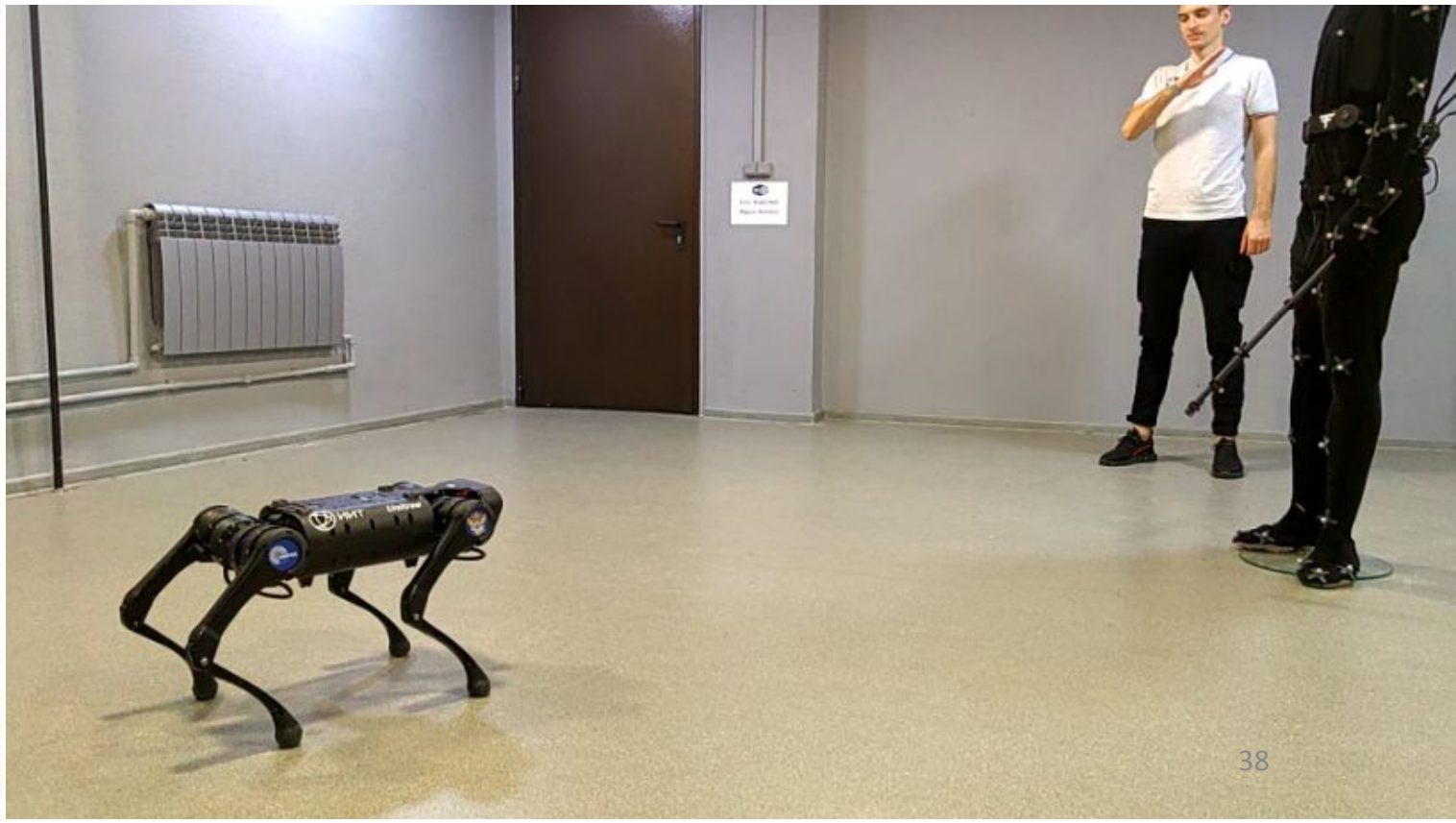
Edit prompt or view more images↓

TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓





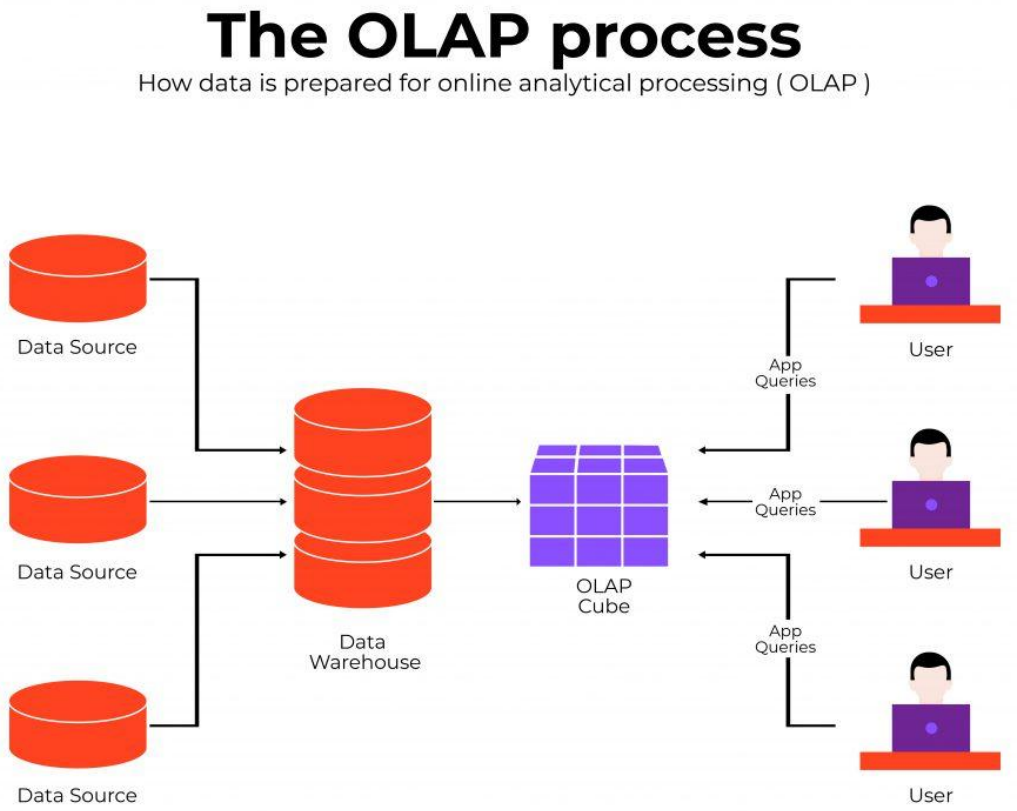
Часть 3. OLAP системы

OLAP системы

OLAP (Online Analytical Processing) – это система аналитической обработки данных. Она предназначена для подготовки отчетов, построения прогностических сценариев и выполнения статистических расчетов.

Является классическим применением витрин данных и является продолжением конвейера обработки данных.

OLAP системы основаны на **аналитических базах данных**, которые оптимизируют запросы на доступ к данным с целью ускорения таких операций



Скорость доступа к данным в OLAP системах

Аналитические базы данных – специализированные колоночные РСУБД, оптимизированные для быстрой выборки данных из витрин.

Основаны на системах массивных параллельных вычислений (MPP) и базах данных, поддерживающих такой режим работы.

Предназначены такие БД для обработки данных на стадии работы непосредственно с витринами больших данных.

Greenplum

snowflake

Amazon Redshift

Колоночные СУБД

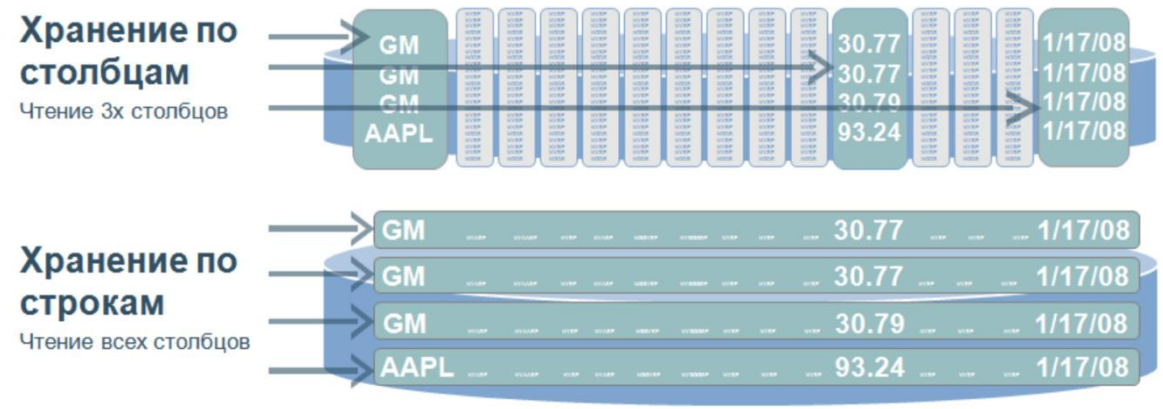
Колоночные СУБД — системы управления базами данных в которых данные хранятся и индексируются столбцами.

Преимущества:

- Быстрая выборка данных
- Более качественное сжатие данных в колонках ввиду однородности измерений
- Гибкость схемы данных

Недостатки:

- Сложности при внесении новых данных
- Не подходят для транзакционных систем

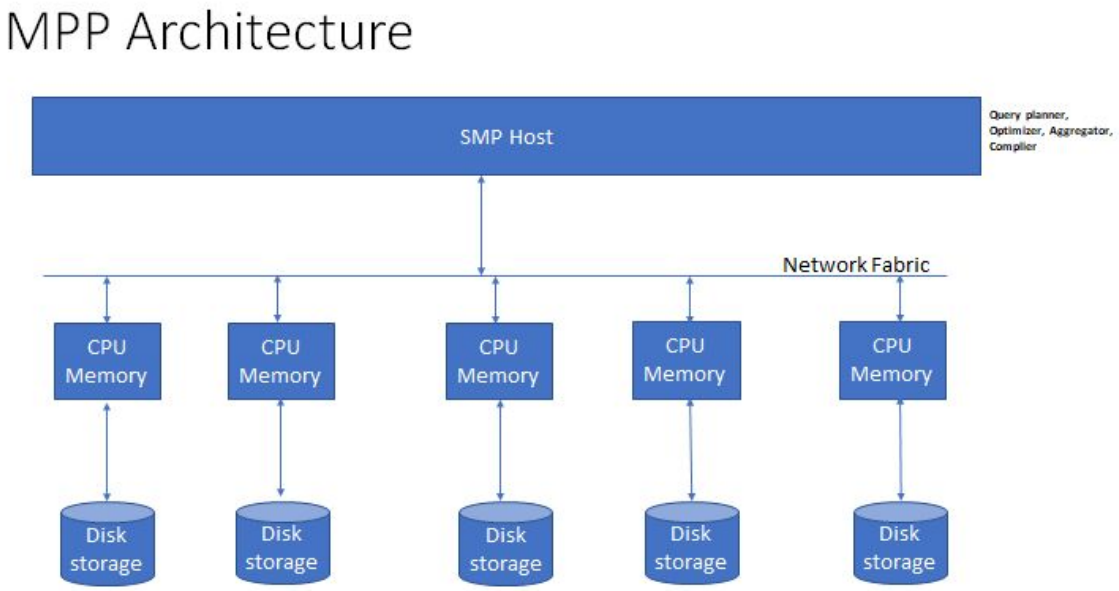


MPP-системы

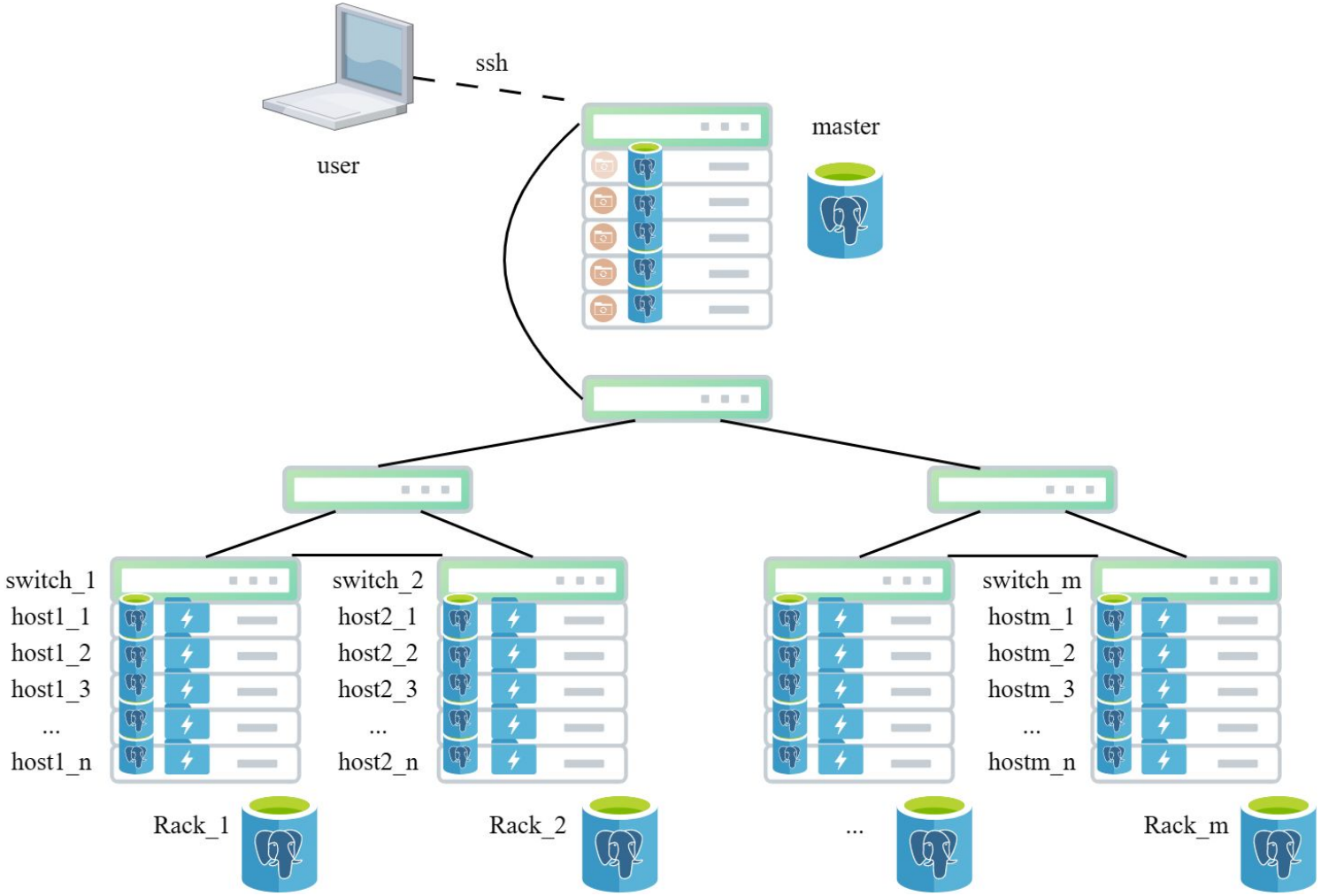
MPP – архитектура параллельных вычислений, при которой память физически разделена.

Система строится из отдельных узлов, содержащих процессор и выделенную оперативную память

В такой системе скорость обработки запросов напрямую зависит от количества задействованных узлов.



Аппаратная масштабируемость



Принципы работы с MPP

Используются, если:

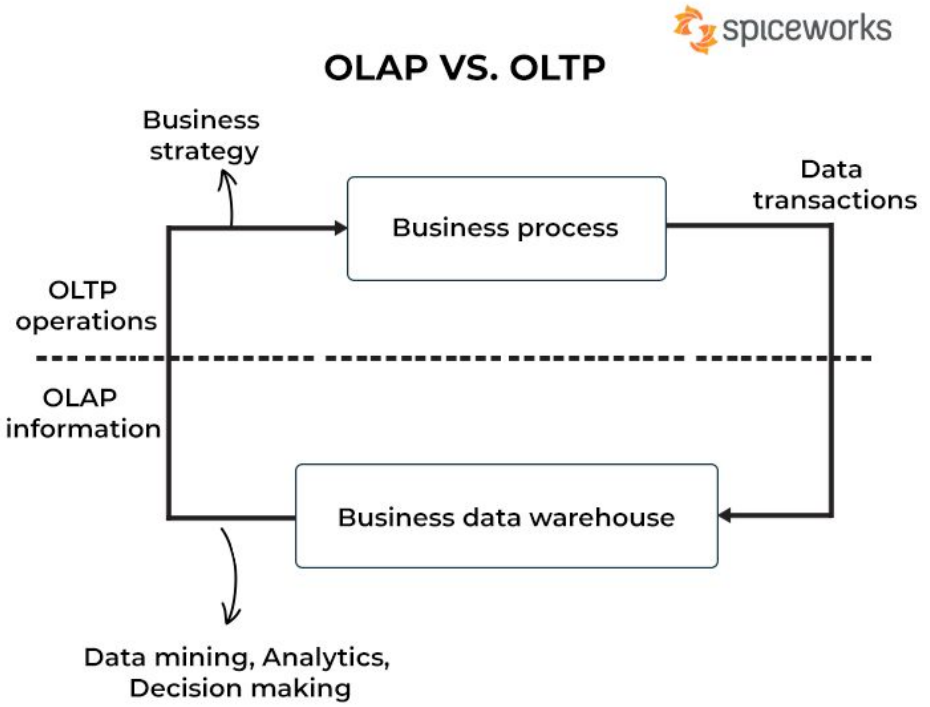
- объемы данных слишком большие для классической СУБД
- когда есть готовое хранилище но отчеты генерируются недостаточно быстро

По источникам данных получаются инкрементальные данные, из которых строятся витрины

OLAP против OLTP

OLTP — системы хранения оперативных данных с высокой скоростью записи данных и фиксации изменений.

- OLTP БД придерживаются принципа нормализации данных
- OLAP придерживается принципа денормализации
- Нормализация данных – хранение информации максимально просто и не избыточно



Часть 4. Способы визуализации данных

Разведочный анализ данных (EDA)

Разведочный анализ данных — анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей на основе визуализаций.

Принципы визуализации данных:

1. Логика
2. Простота
3. Цвет

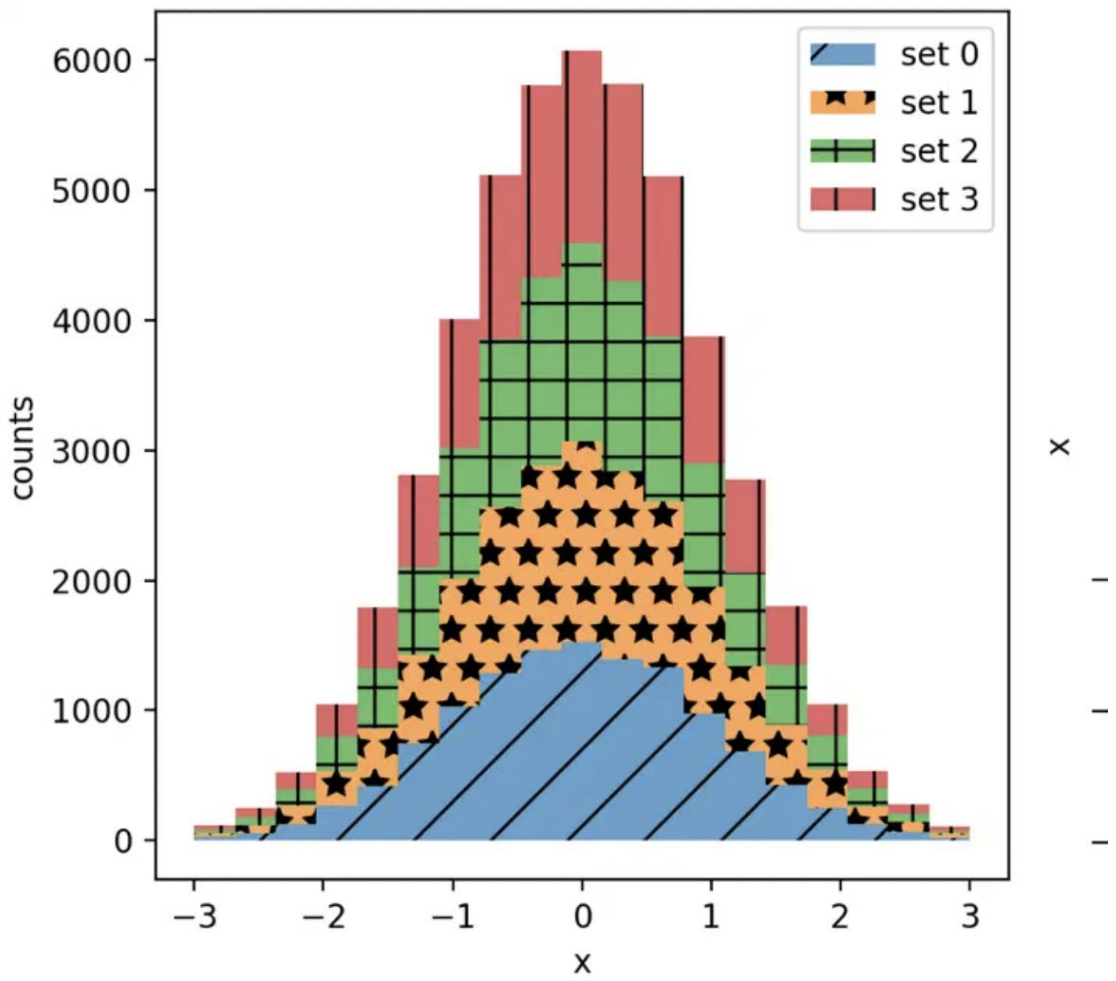


График разброса

График разброса — это средство для показа взаимоотношений между двумя переменными.

Строит визуализацию точек-строк на пересечении координат-столбцов.

Принципы построения:

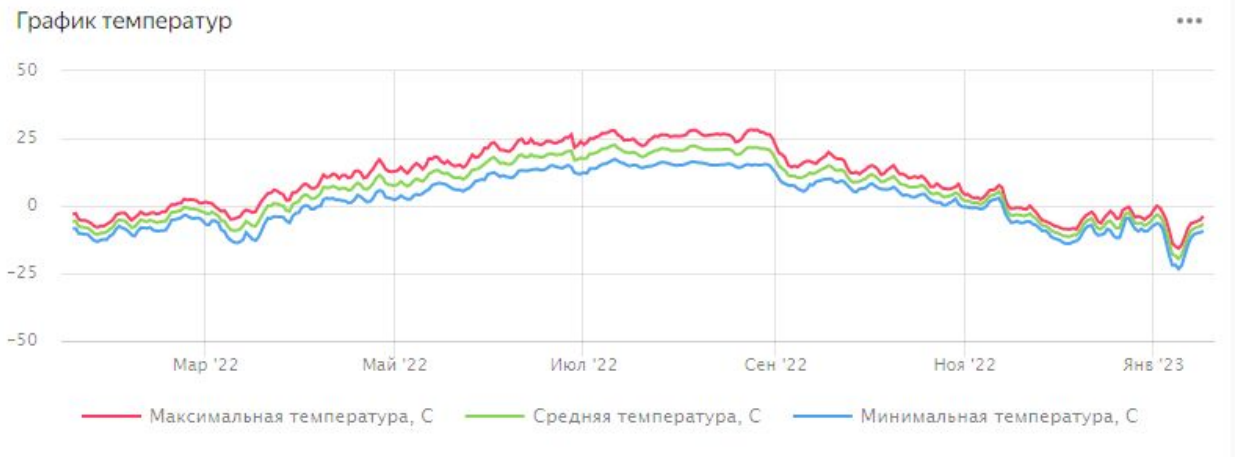
- Полностью заполненная информацией область,
- Описание осей или зависимости
- Цветом выделяем третье измерение

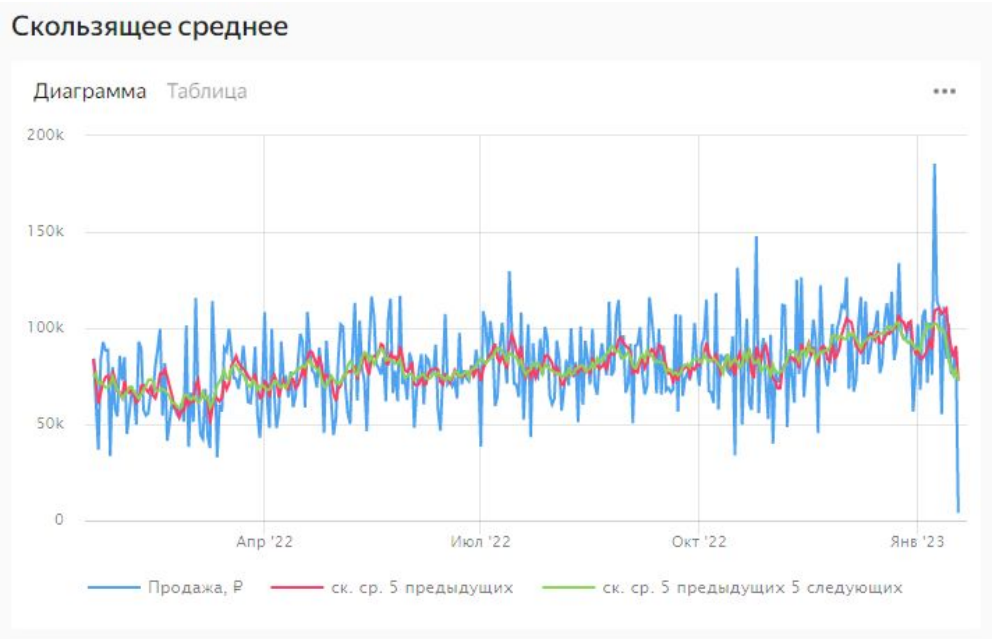
Визуализация помогает понять взаимосвязь двух атрибутов по их корреляции.

График линий

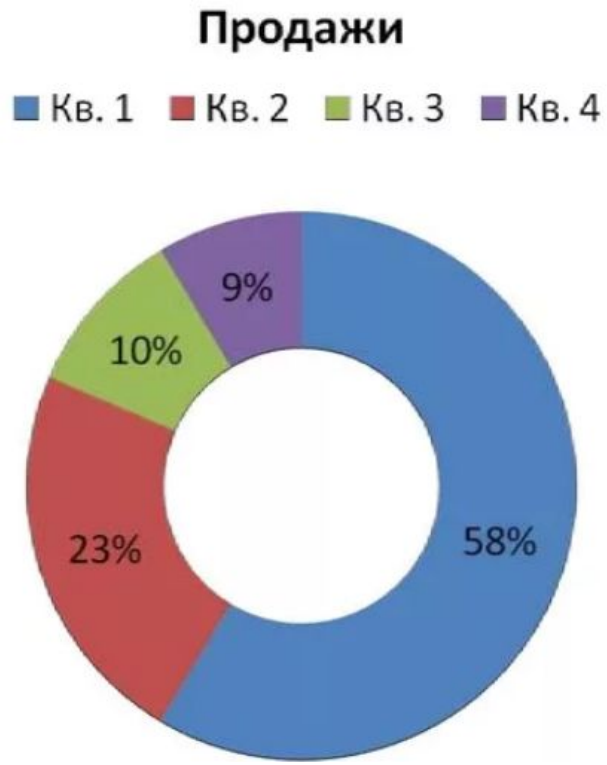
График линий отображает динамику развития процесса во времени или измерении.

Основная цель — отследить зависимость от упорядоченного фактора, который вносит не только прямую зависимость от величины, но и от порядка измерения.





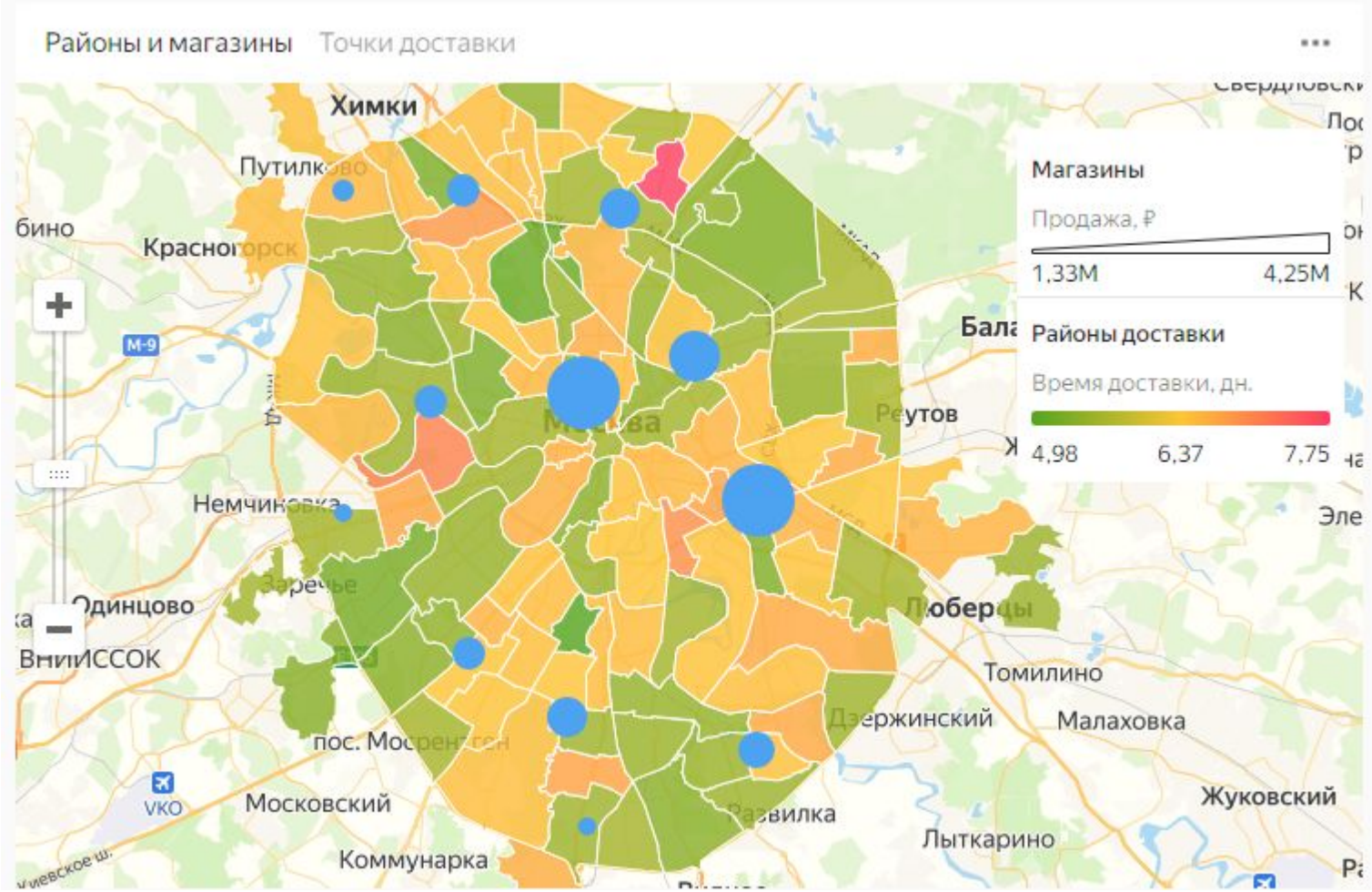
Столбчатая диаграмма



Карты

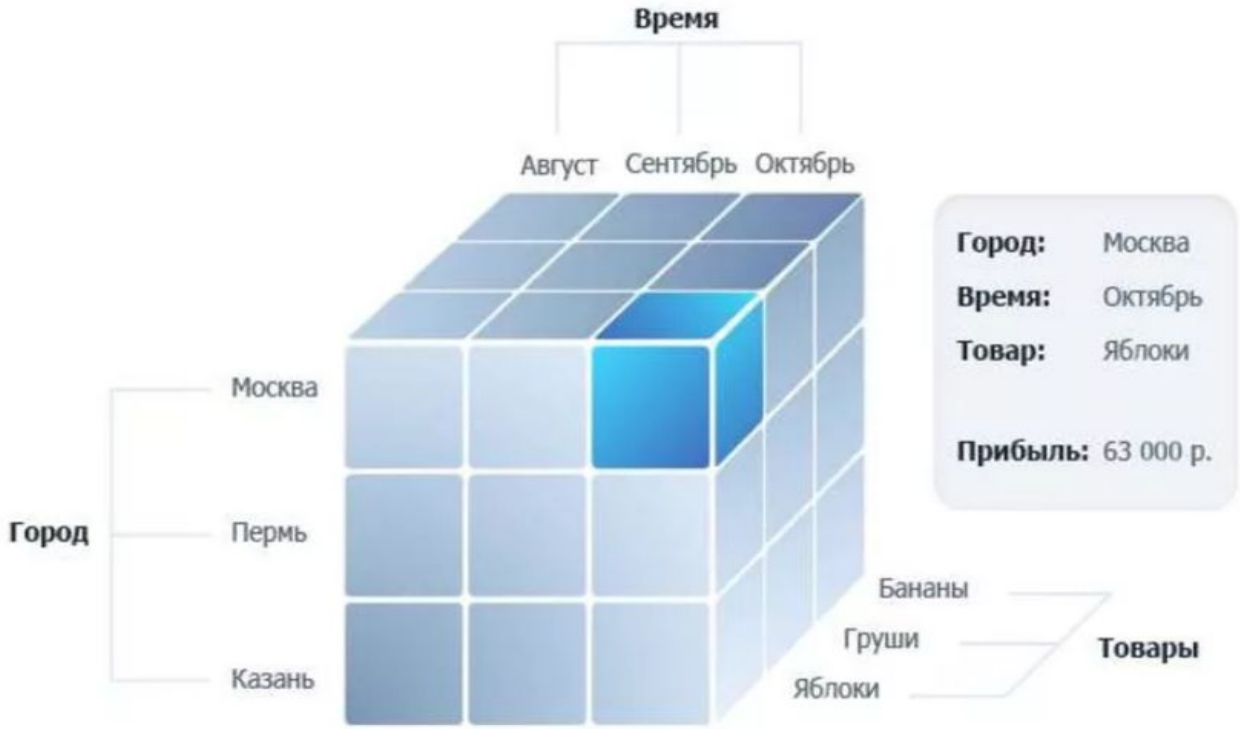
Карты помогают отследить распределение спроса на реальных географических данных.

Визуализация позволяет сконцентрироваться на популярных точках распространения продукции и услуг, или позволяет принимать логистические решения в различных сценариях.



OLAP куб

OLAP куб – предназначен для визуализации многомерных массивов данных.



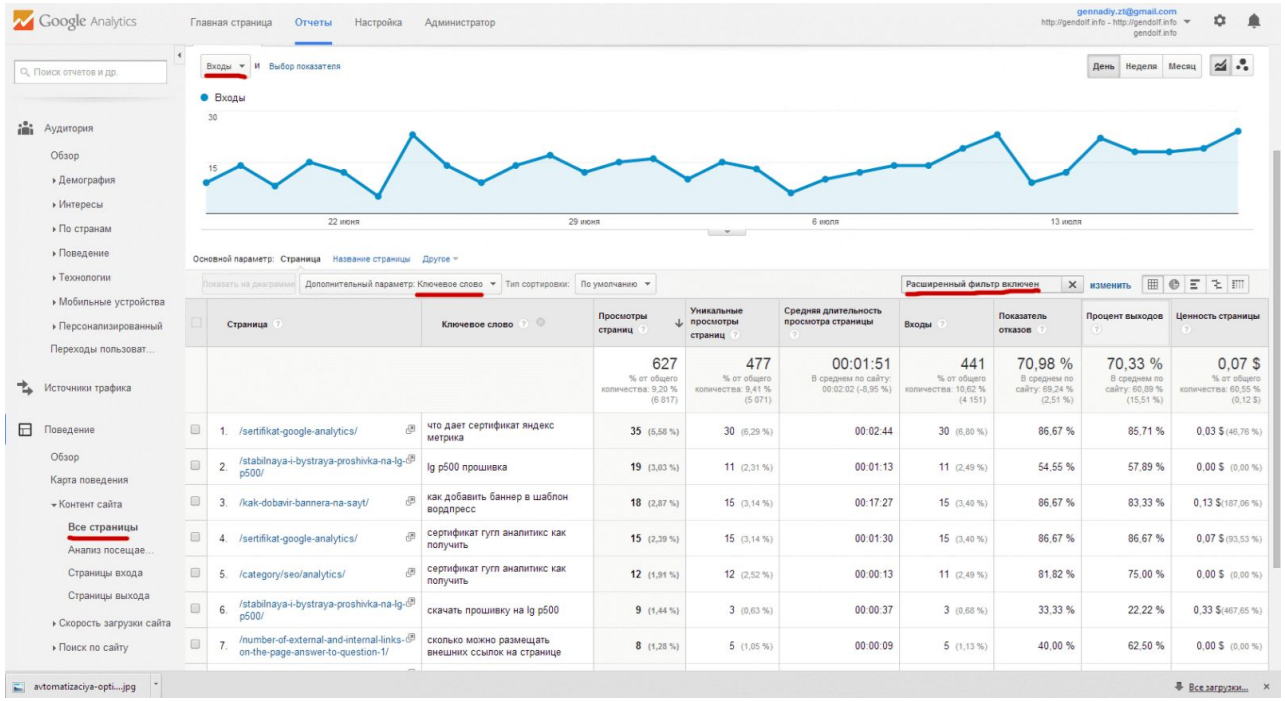
Построение динамических отчетов

- Упрощение
- Сравнение
- Сопровождение
- Взгляд иначе
- Вопрос “почему?”
- Скептицизм
- Отклик

Отчеты в аналитике

Отчеты используются для работы с определенными наборами данных, например, создания ежедневных отчетов о количестве заказов или количестве доставленных продуктов.

- Понять цель создания дашборда
- Подключить источники данных
- Обработать данные
- Связать данные между собой
- Сделать необходимые расчеты
- Визуализировать
- Проверить корректность

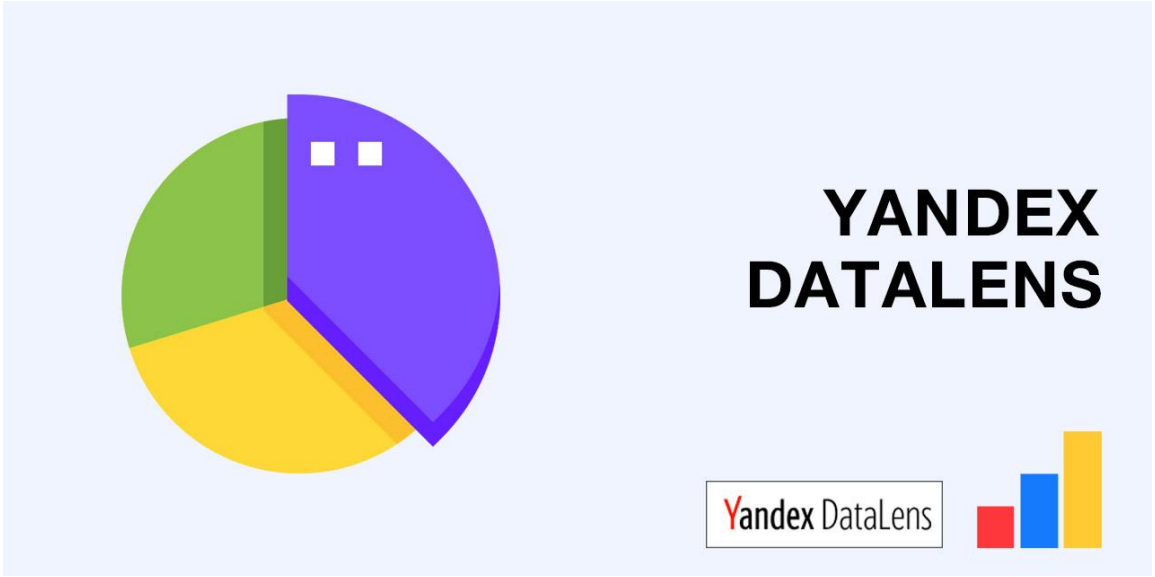


Дашборды

Дашборд — это интерактивная аналитическая панель, графический интерфейс. Смысл в том, что на одном экране расположены все ключевые метрики, показатели цели или процессов. С помощью этих метрик можно выявить и проанализировать тренды и изменения.

Для визуального анализа хорошо использовать визуализацию данных, но еще лучше применять динамические и интерактивные диаграммы и графики, чтобы использовать эффективный инструмент.

Системы построения дашбордов



Существует много систем построения дашбордов: Apache Superset, Preset, Metabase, Redash, Power BI и т.д.

Современной системой является **Yandex DataLens.**

Сервис позволяет подключаться к разным источникам данных, собирать дашборды, строить визуализации и делиться полученными результатами.