

АНАЛИЗ ДАННЫХ

Введение в анализ данных

Вначале были данные

Есть база данных:

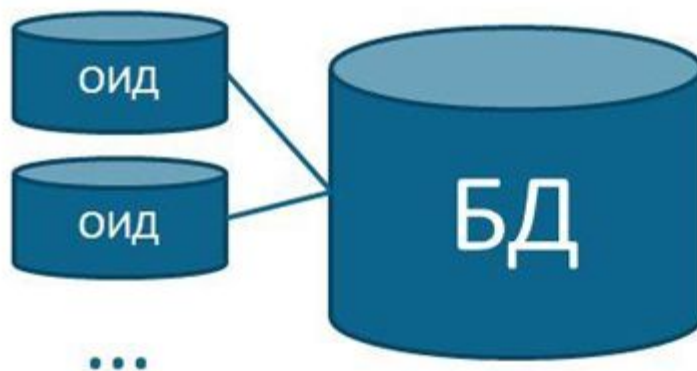


Например, БД банка.

Хранит персональные данные, счета, кредиты и т.д.

Источник данных

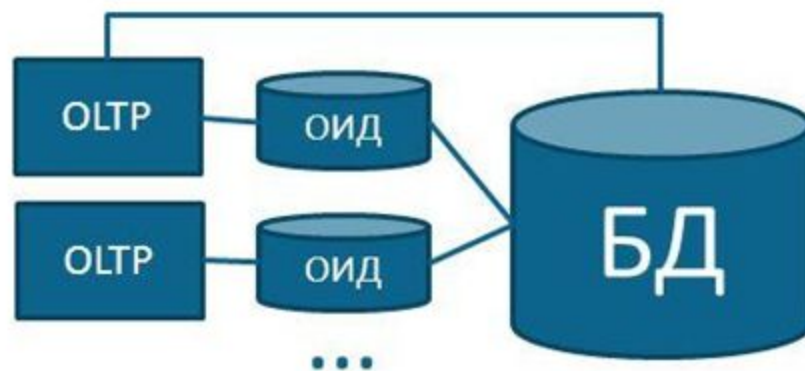
- Сама по себе БД пользы не несет
- Добавим **оперативный источник данных (ОИД)**



Для банка это терминалы, базы локальных отделений и т.д.

Работа с транзакциями

- Нужна обработка данных в БД
- Добавим **Online Transaction Processing (OLTP)**



Взаимодействует с ОИД и БД.

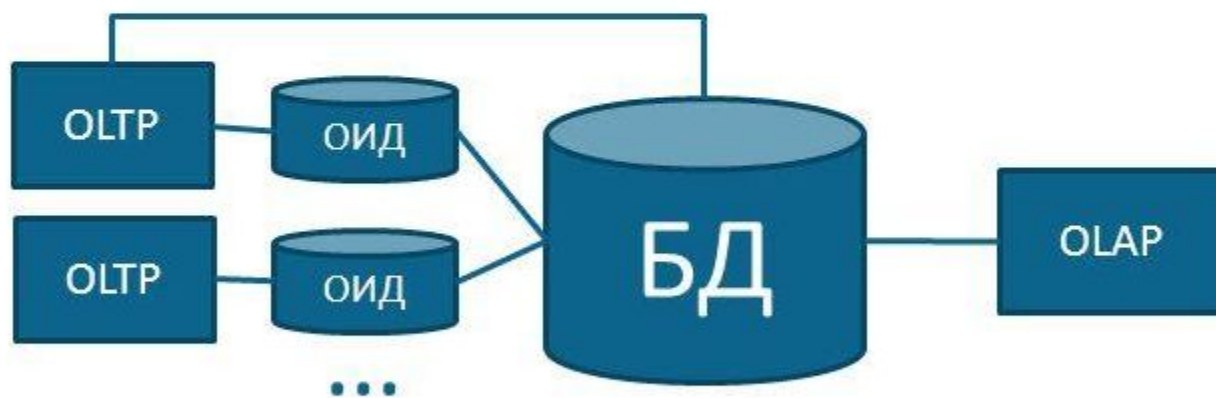
Примеры операция для банка: узнать счет, перевести деньги, пополнить баланс.

Особенности OLTP

- Важна скорость работы – результат максимум за пару секунд
- Простые операции с данными (около CRUD)
- Высокая частота вызовов => постоянная средняя загрузка процессора
- Работа только с оперативными данными

Аналитическая система

- Нужна более сложная система для работы с БД
- Добавим **Online Analytical Processing (OLAP)**



Взаимодействует с БД.

Примеры для банка: найти подозрительные переводы, определить кредитоспособность

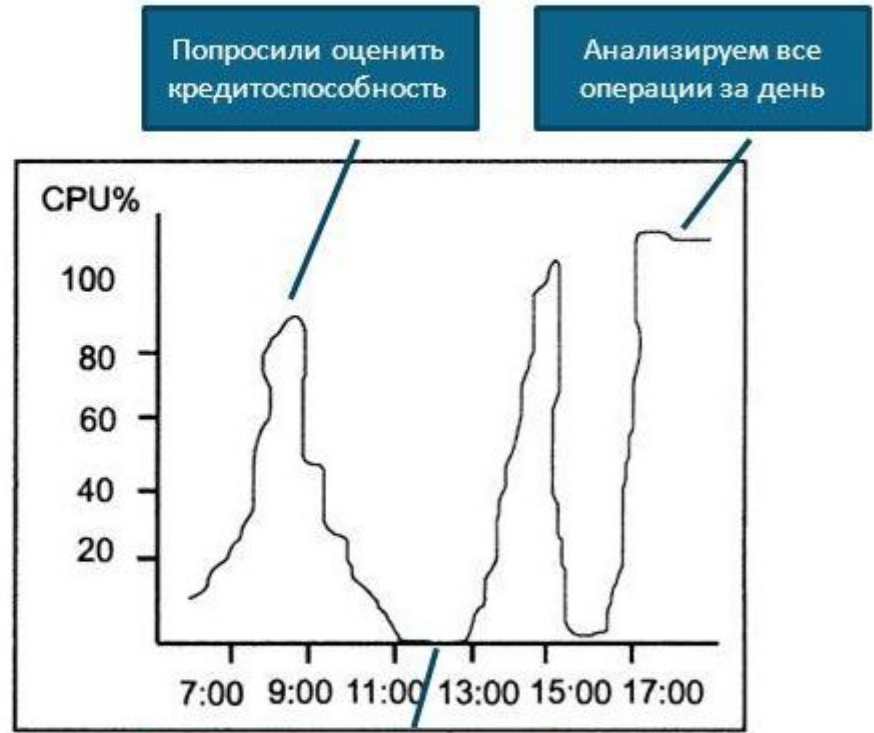
Особенности OLAP

- Важна точность анализа
- Сложные запросы, функции, процедуры
- Периодические вызовы чередуются с простоем. Нагрузка на процессор непостоянна.
- Работа с большой коллекцией данных

OLTP и OLAP во время работы



OLTP всегда готова выполнить запрос



Выдали результат, больше запросов нет, отключаемся

Возможности OLAP

Может:

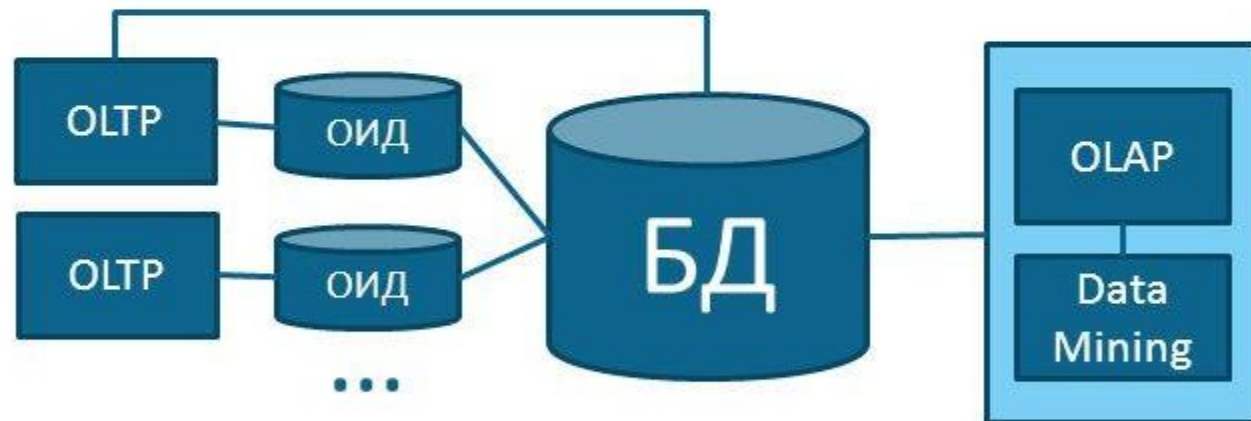
- Определить кредитоспособность по имеющимся правилам
- Прогнозировать прибыль банка на основе моделей и гипотез

Не может:

- Предложить правило оценки кредитоспособности
- Сгенерировать модель или гипотезу

Анализ данных

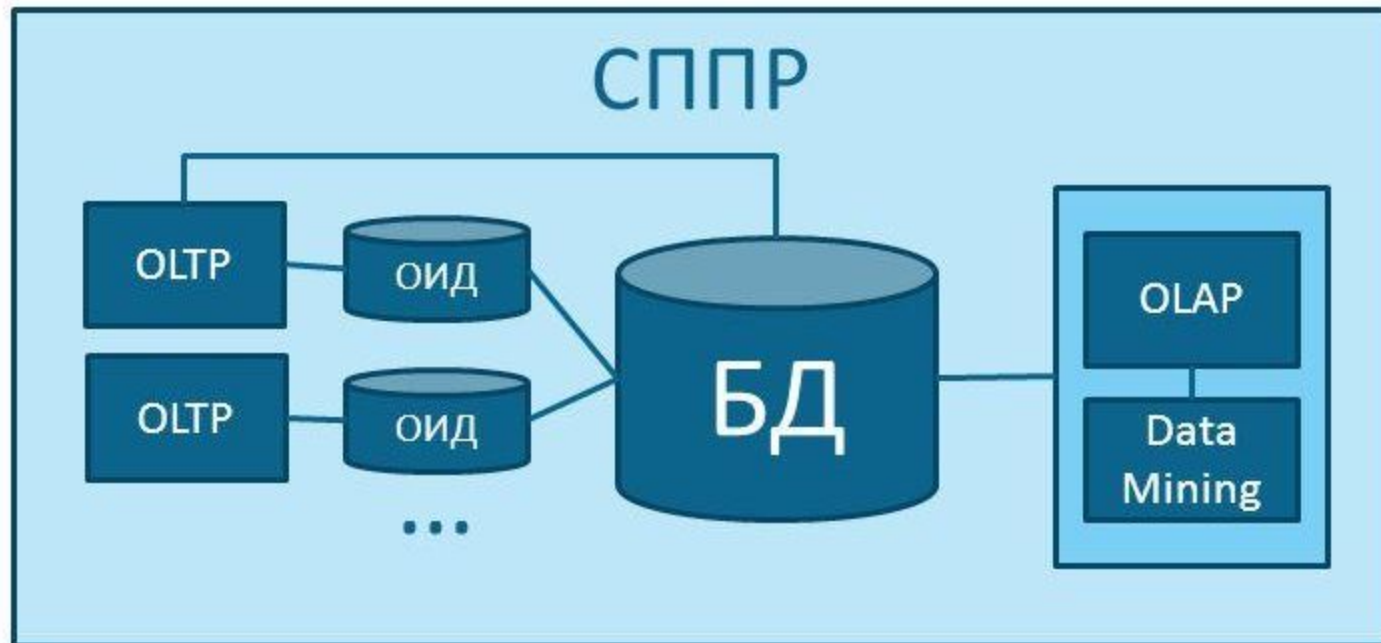
- Нужна система генерации гипотез
- Вот и **Анализ Данных** (Data Mining)



Примеры для банка:

- Выяснить зависимость кредитоспособности человека от наличия высшего образования
- Каковы признаки подозрительных переводов?

Система поддержки принятия решений



Комплекс для **сбора, хранения и анализа** информации

Определение

Анализ Данных - это процесс обнаружения в «сырых» данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Примеры для самопроверки

Какие правила являются успешным результатом анализа данных:

- Если фигура - четырехугольник, то сумма его углов равна 360 градусам
- Во время беспорядков повышается спрос на бейсбольные биты
- Если что-то выглядит как утка и крякает как утка, то это вероятно это и есть утка
- Вместе с хлебом люди часто покупают молоко
- Люди старше 60 не ищут ночные клубы в Москве

Примеры для самопроверки

Какие правила являются успешным результатом анализа данных:

- Если фигура - четырехугольник, то сумма его углов равна 360 градусам
- уже известная информация
- Во время беспорядков повышается спрос на бейсбольные биты
+ полезно на будущее
- Если что-то выглядит как утка и крякает как утка, то это вероятно это и есть утка
- тривиальная информация
- Вместе с хлебом люди часто покупают молоко
+ можно продавать комплектом
- Люди старше 60 не ищут ночные клубы в Москве
- практически бесполезное знание

Типичные задачи АД

- Классификация (Classification)
- Кластеризация (Clustering)
- Ассоциация (Associations)
- Визуализация (Visualization, Graph Mining)
- Последовательность (Sequence)
- Прогнозирование (Forecasting)
- Определение отклонений (Deviation Detection)
- Анализ связей (Link Analysis)

Примеры задач классификации

- Определение DDoS-атак
- Спам-фильтры
- Привлечение выгодных клиентов (определение целевой аудитории продукта)
- Определение профиля ДНК

Примеры задач кластеризации

- Группировка документов по темам
- Идентификация людей на записях с камер видеонаблюдения
- Кластеризация тикетов
- Кластеризация структуры фондового рынка

Примеры задач поиска ассоциаций и последовательностей

- Рекомендации товаров
- Обнаружение скрытых факторов влияния

Примеры задач визуализации

- Связи в социальных сетях
- Пробки на дорогах
- Инфографика

Семейство направлений АД

- **Web Mining** – специализируется на анализе страниц в интернете (определение ТИЦ, выделение модульной сетки сайта)
- **Opinion Mining** – специализируется на анализе отношений пользователей к различным объектам (whatdoestheinternetthink.net)
- **Information Retrieval** – поиск неструктурированной информации в текстовых документах (поисковые системы)

Основные определения

- **Данные** – необработанный материал, используемый для формирования информации на основе данных.



Например, текст документа

Основные определения

- **Объект** - описывается как набор атрибутов



Сам текстовый документ

- **Атрибут** - свойство, характеризующее объект.



Имя: Текстовый документ (42).txt

Размер документа: 5кб

Дата создания: 05.09.2011

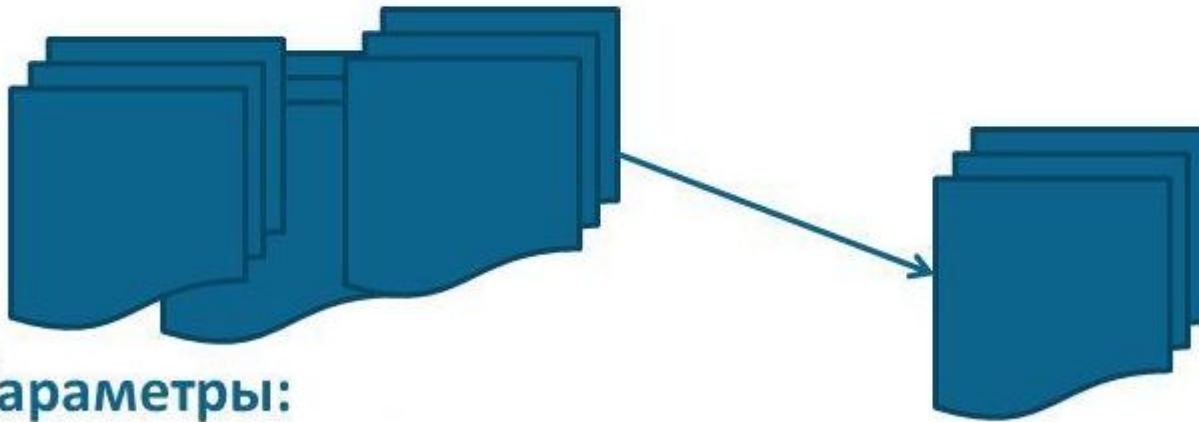
Основные определения

- **Генеральная совокупность** (population) - вся совокупность изучаемых объектов, интересующая исследователя.
- **Выборка** (sample) - часть генеральной совокупности



Основные определения

- **Параметры** - числовые характеристики генеральной совокупности.
- **Статистики** - числовые характеристики выборки.



Параметры:

Научные статьи: 30%

Анекдоты: 70%

Статистика:

Научные статьи: 32%

Анекдоты: 68%

Основные определения

- **Гипотеза** - частично обоснованная закономерность знаний, служащая либо для связи между различными эмпирическими фактами, либо для объяснения факта или группы фактов.



Гипотеза:

Если файл < 5кб, то он скорее всего содержит анекдоты

Основные определения

- **Измерение** - процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу.



Количество слов:

количество пробельных символов + 1

Шкалы измерений

Номинальная шкала (nominal scale) -содержит только категории

- Нельзя упорядочить
- Доступные операции: ==, !=

Пример: месяцы, царства животного мира, категории статей

Дихотомическая шкала (dichotomous scale) - шкала, содержащая только две категории.

Пример: бит, пол

Шкалы измерений

Порядковая шкала (ordinal scale) - шкала, в которой числа присваивают объектам для обозначения относительной позиции объектов, но не величины различий между ними.

- Доступные операции: $=$, \neq , $>$, $<$

Пример: место в рейтинге

Шкалы измерений

Интервальная шкала (interval scale) - шкала, разности между значениями которой могут быть вычислены, однако их отношения не имеют смысла.

- Доступные операции: $=$, \neq , $>$, $<$, $+$, $-$

Пример: температура

Шкалы измерений

Относительная шкала (ratio scale) - шкала, в которой есть определенная точка отсчета и возможны отношения между значениями шкалы.

- Доступные операции: $=$, \neq , $>$, $<$, $+$, $-$, $*$, $/$

Пример: вес и размеры предметов