Analytic
Workspace

Цикл работы с данными

Диана Нечунаева
Аналитик внедрения BI-систем

Analytic Workspace

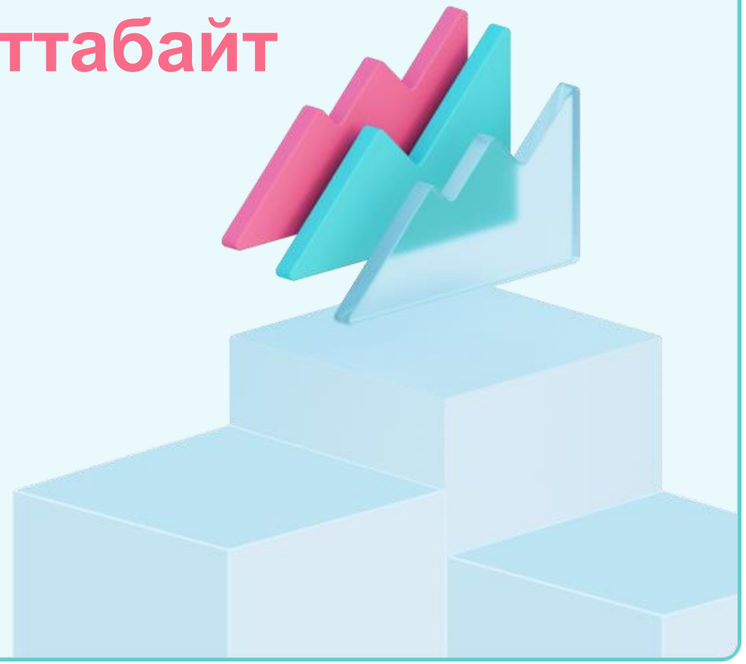
План занятия

- Ценность данных
- Методология исследования данных CRISP
- Основные аспекты этапов подготовки и моделирования данных
- Сбор данных
- Хранение данных
- Обработка данных
- Визуализация данных

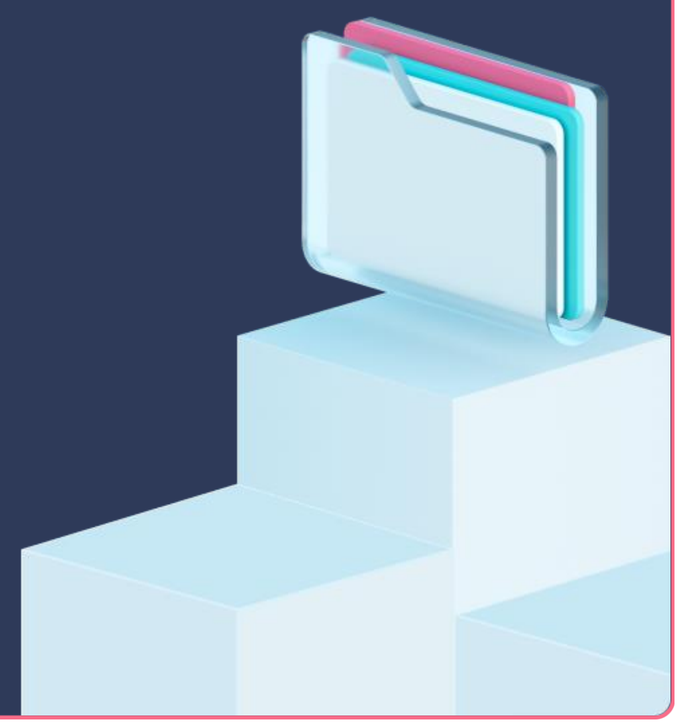
Analytic Workspace

Ценность данных

К 2025 году объемы всех данных, накопленных человечеством, возрастут до 180 зеттабайт



Не все собранные данные имеют ценность



Analytic
Workspace

Ценность данных

Данные имеют ценность, только если они принесут пользу

80% компаний не используют все данные, которые собирают, так как:

- объем информации увеличивается слишком быстро
- у сотрудников недостаточно технических навыков
- инфраструктура компании не позволяет получать максимум выгоды из собранных данных

Analytic
Workspace

Цикл обработки данных

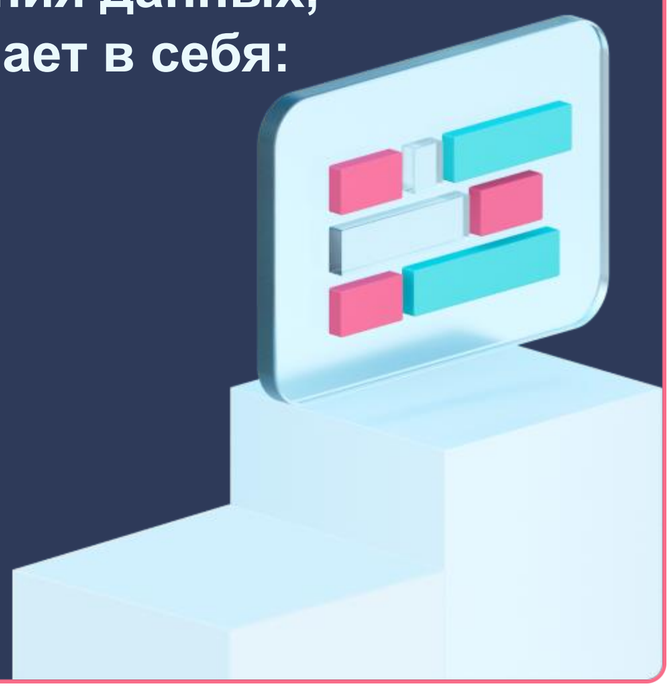
Чем качественнее и чем чаще мы будем собирать всю необходимую информацию для принятия решений, тем эффективнее будет наше управление бизнесом

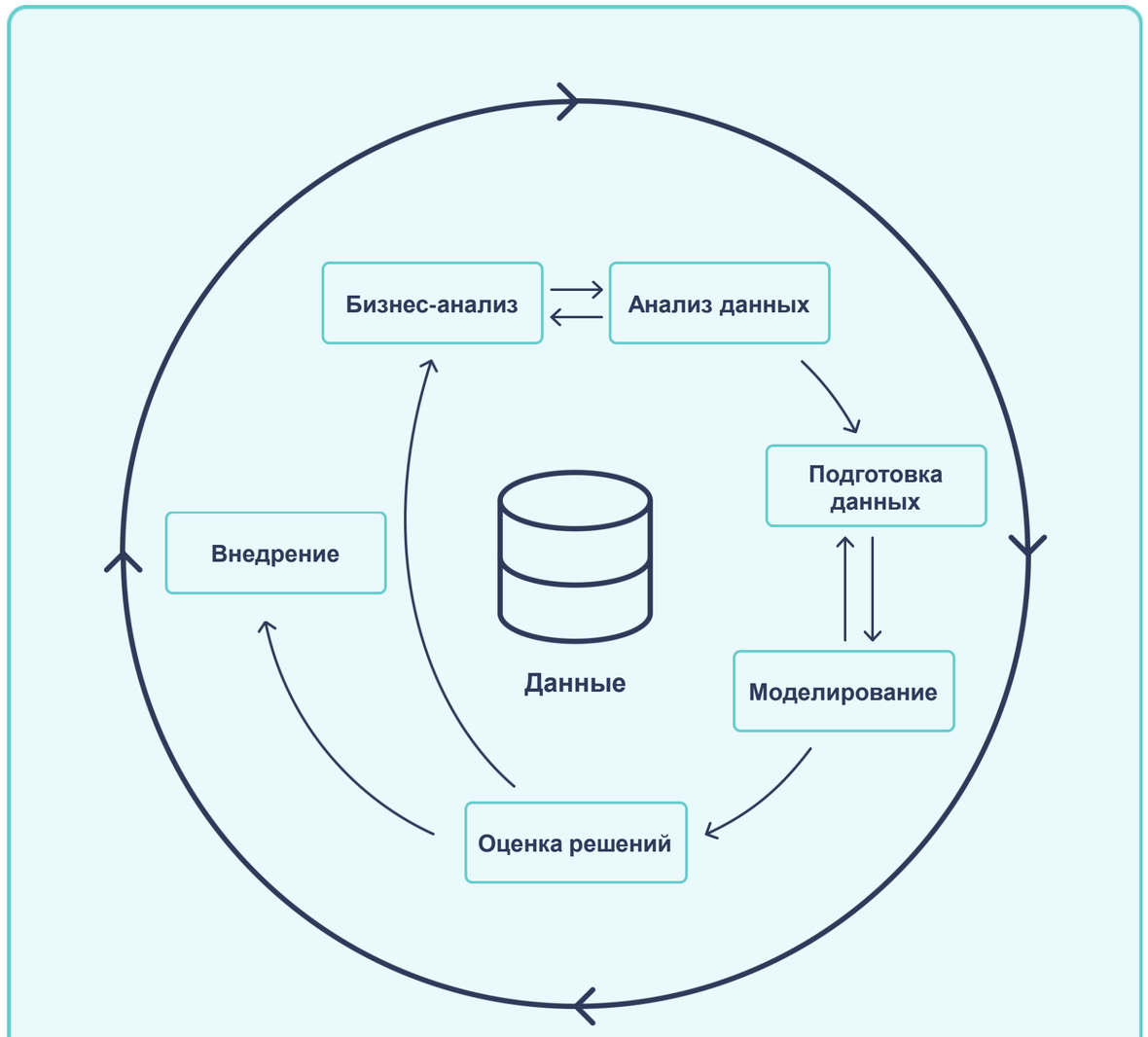
Analytic Workspace

Цикл обработки данных

Методология исследования данных CRISP – межотраслевой стандартный процесс для исследования данных, который включает в себя:

- Бизнес-анализ
- Анализ данных
- Подготовку данных
- Моделирование
- Оценку решения
- Внедрение





Analytic
Workspace

Бизнес-анализ
Задачи

01

Определить цели
организации

02

Определить цели
анализа данных

03

Оценить текущую
ситуацию

04

Составить план
проекта

Analytic Workspace

Анализ данных
Задачи

Определить источники

01

Описать структуру

02

Исследовать данные на качество

03

Analytic Workspace

Подготовка данных
Задачи

01

Отобрать данные (таблицы, записи и атрибуты)

02

Очистить данные, в т.ч. выполнить их конвертацию и подготовку к моделированию

03

Сделать расчёт производных данных

04

Объединить данные

05

Привести данные в нужный формат

Analytic Workspace

Моделирование
Задачи

01

Выбрать методику моделирования

02

Построить модель

03

Протестировать модель данных

04

Оценить модель

Analytic
Workspace

Оценка решения
Задачи

Оценить
результаты

01

Сделать ревью
процесса

02

Определить следующие шаги

03

Analytic Workspace

Внедрение
Задачи

01

Запланировать развертывание

02

Сделать финальный отчет

03

Запланировать поддержку и мониторинг развернутого решения

04

Сделать ревью проекта

Analytic Workspace

Аспекты этапов подготовки и моделирования данных. Сбор данных

Поиск данных осуществляется по схеме:

01 → Формулировка запроса – что ищем

02 → Запрос консультаций с целью помощи в поиске источников

04 → Самостоятельный поиск

05 Запрос и получение данных

Результат этапа — **список источников данных**

Analytic Workspace

Аспекты этапов подготовки и моделирования данных. Сбор данных

Источник данных –
это место, где собирается информация

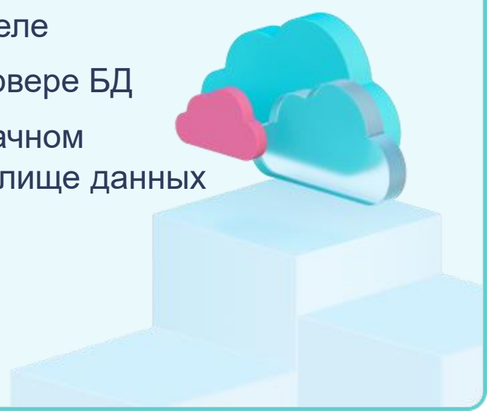
Моделирование данных –
это создание визуального представления обо всей информационной системе либо ее части

Analytic Workspace

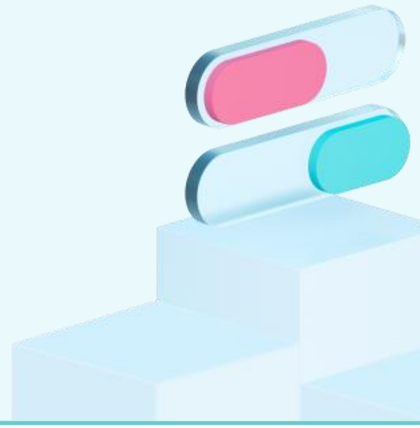
Аспекты этапов подготовки и моделирования данных. Хранение данных

Данные можно хранить разными способами:

- на твердотельном съемном или несъемном носителе
- на сервере БД
- в облачном хранилище данных



Выбор способа хранения данных зависит от многих критериев



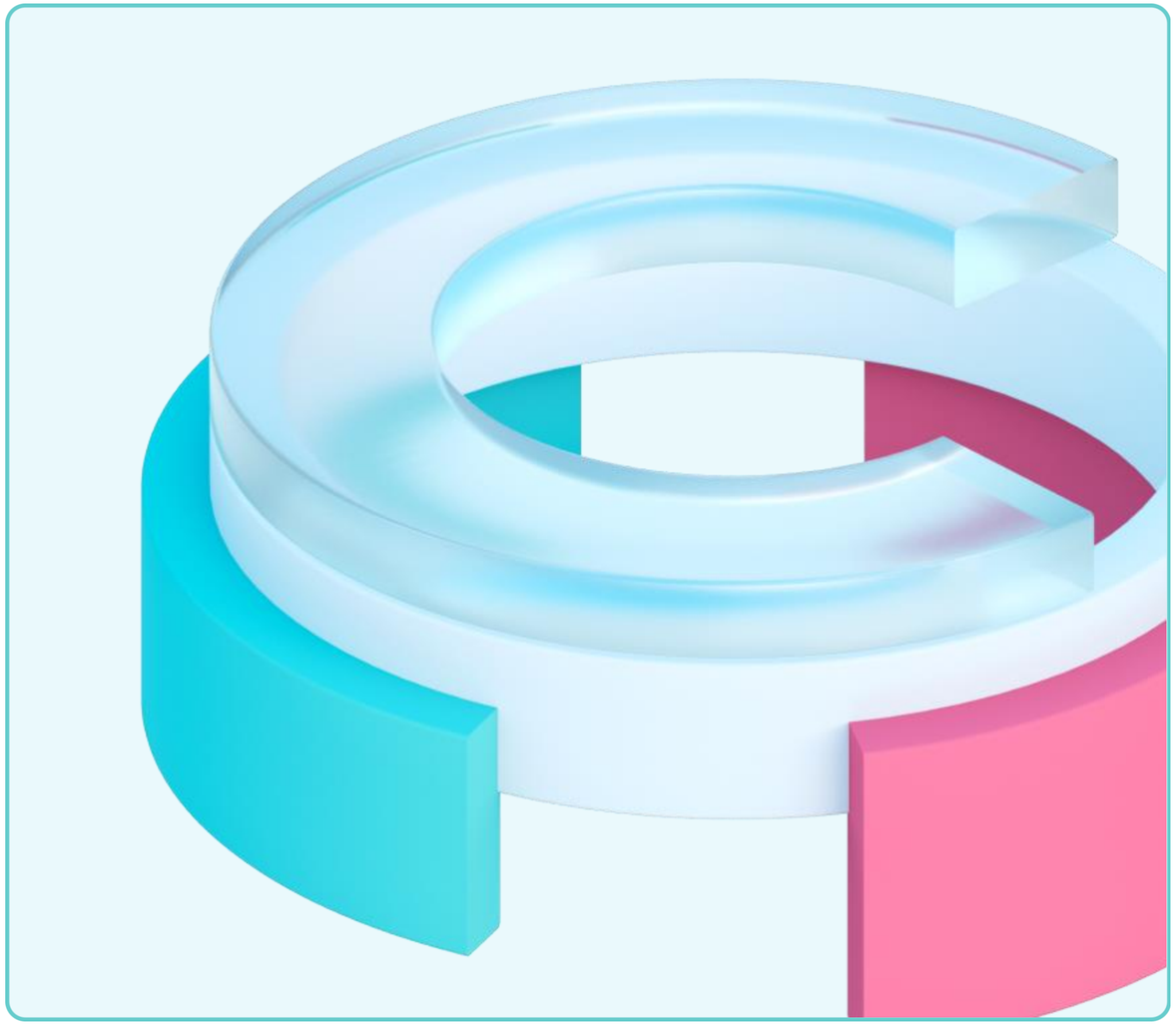
Основной формой хранения данных является база данных

Analytic Workspace

Аспекты этапов подготовки и моделирования данных. Обработка данных

Обработка данных —

определенная последовательность операций с данными



Основные процедуры обработки данных:

- сортировка (упорядочение)
- выборка
- слияние
- поиск
- корректировка
- сжатие

Analytic Workspace

Аспекты этапов подготовки и моделирования данных. Обработка данных

«Грязные данные» – это нормально

Универсальных решений для очистки данных от всех ошибок не существует

Возможные проблемы при обработке данных:

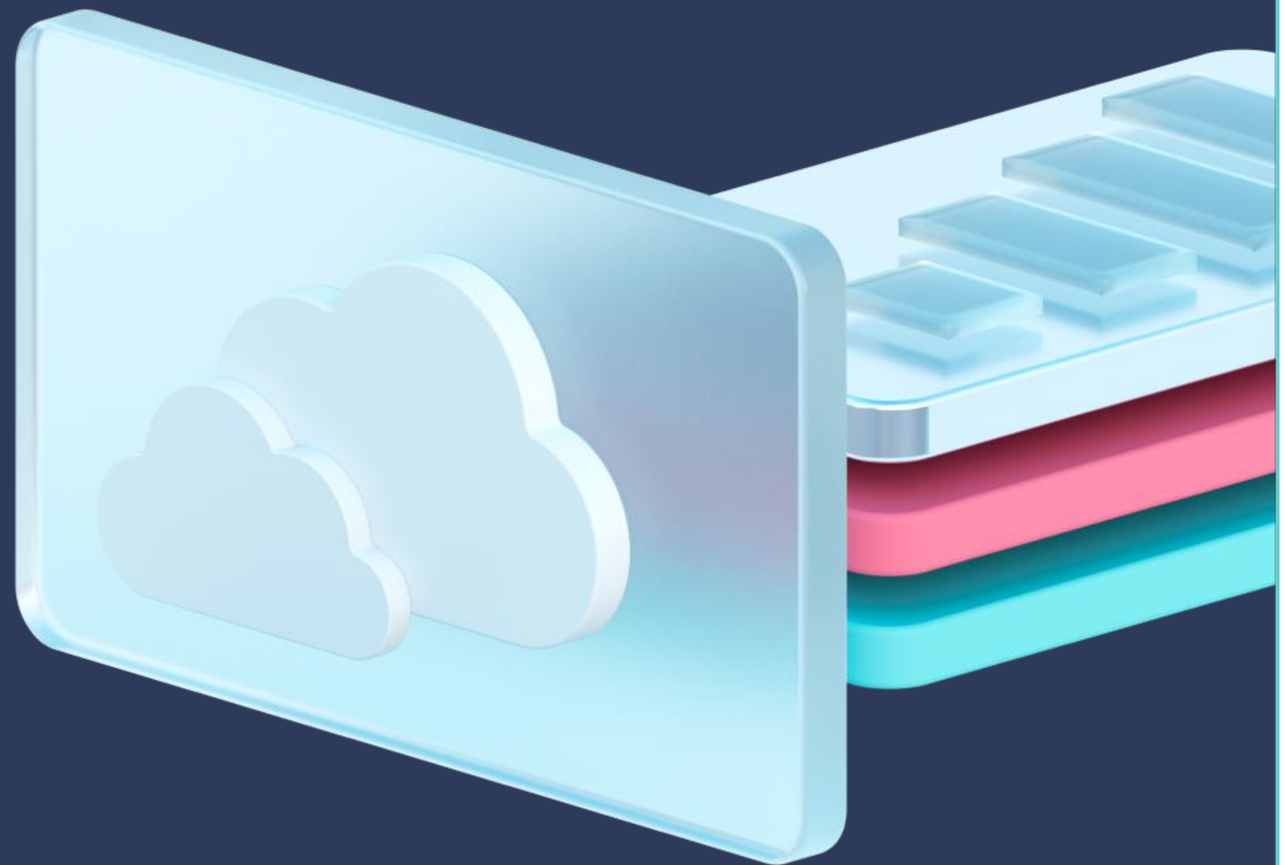
- Дублирование записей
- Неуникальные значения
- Противоречивые записи
- Отсутствующие значения
- Недопустимые значения
- Орфографические ошибки и опечатки
- Аномальные значения
- Многозначность
- и другие

Analytic Workspace

Аспекты этапов подготовки и моделирования данных. Обработка данных

Трансформация данных

- Фильтрация по признакам, ограниченная выборка
- Транспонирование
- Обогащение данных



Аспекты этапов подготовки и моделирования данных. Визуализация данных

Аналитический отчет является очень важным документом, который поможет проанализировать рынок и собственную деятельность на его фоне

Analytic Workspace

Аспекты этапов подготовки и моделирования данных. Визуализация данных

Критерии, которым должен соответствовать аналитический отчет

Analytic
Workspace

Аспекты этапов подготовки и моделирования данных. Визуализация данных

Визуализация данных происходит
на следующих уровнях:

Стратегический
для ТОП менеджеров

Оперативный
для управленцев

Аналитический
для бизнес-аналитиков

Визуализация данных — процесс представления данных в агрегированном, понятном для восприятия человеком виде

Analytic Workspace

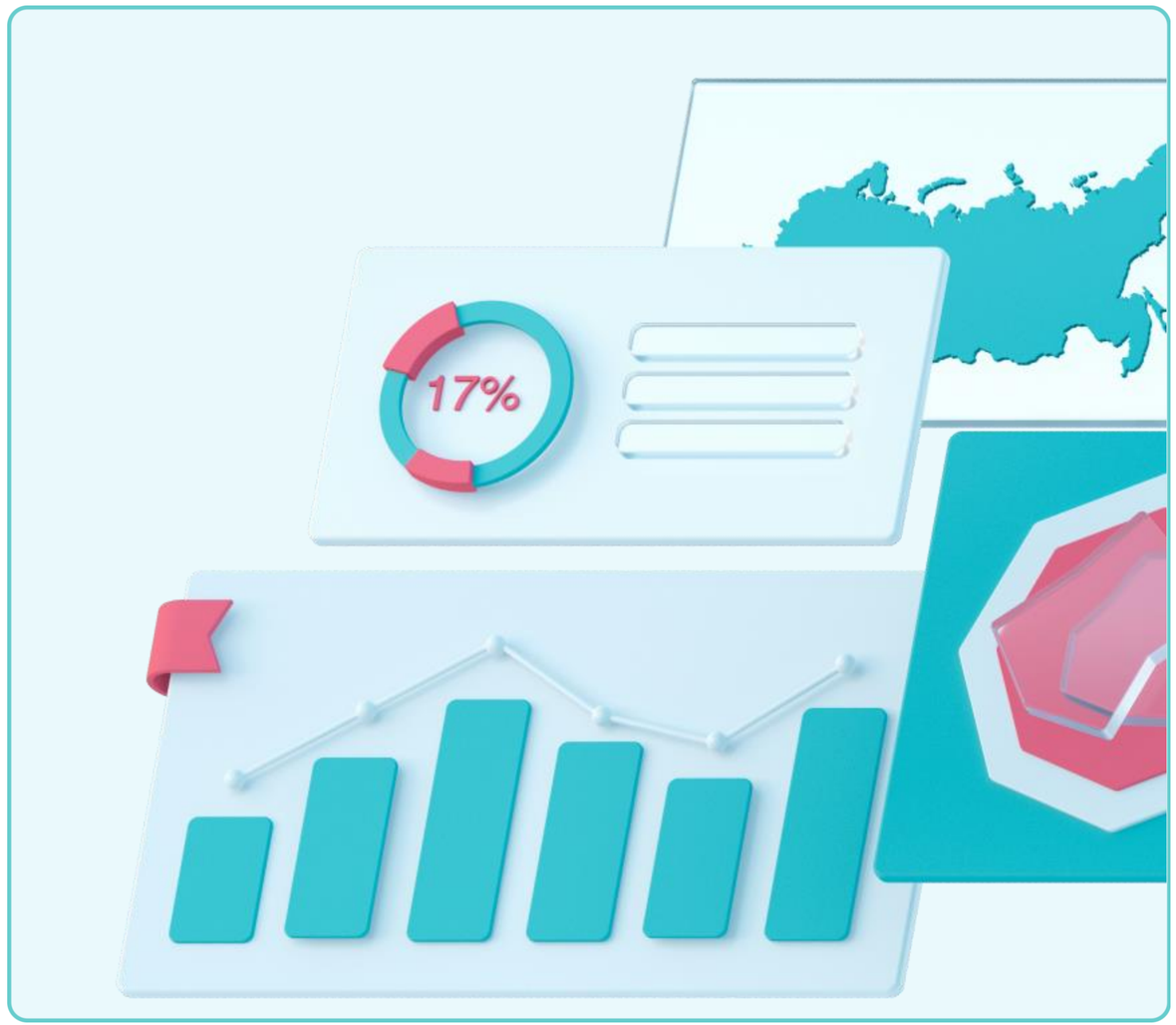
Аспекты этапов подготовки и моделирования данных. Визуализация данных

Визуализация —

это необходимый способ оформления данных в понятный человеку вид

Вид визуализации данных:

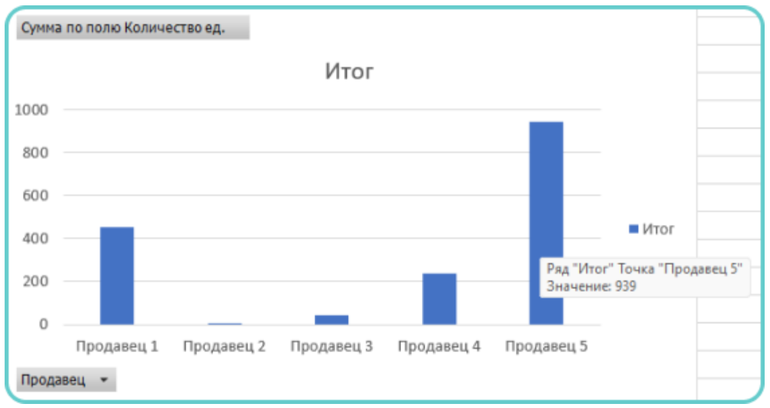
- Графики
- Диаграммы
- Блок-схемы
- Таблицы и матрицы
- Инфографика



Analytic Workspace

Аспекты этапов подготовки и моделирования данных. Визуализация данных. Пример

A	B	C	D	E	F	G
Категория	Артикул	Наименование	Месяц продаж	Продавец	Количество ед	Итоговая сумма
Товарная категория 1	Артикул 1	Наименование 1	Март	Продавец 1	5	2568
Товарная категория 1	Артикул 2	Наименование 2	Март	Продавец 2	6	21145
Товарная категория 1	Артикул 3	Наименование 3	Март	Продавец 4	14	663
Товарная категория 1	Артикул 4	Наименование 4	Март	Продавец 1	26	4458
Товарная категория 2	Артикул 5	Наименование 5	Март	Продавец 5	78	2484
Товарная категория 2	Артикул 6	Наименование 6	Март	Продавец 1	2	26541
Товарная категория 2	Артикул 7	Наименование 7	Март	Продавец 1	3	9952
Товарная категория 2	Артикул 8	Наименование 8	Март	Продавец 3	45	696
Товарная категория 2	Артикул 9	Наименование 9	Март	Продавец 1	87	145
Товарная категория 2	Артикул 10	Наименование 10	Март	Продавец 4	6	154154
Товарная категория 3	Артикул 11	Наименование 11	Март	Продавец 1	3	151
Товарная категория 3	Артикул 12	Наименование 12	Март	Продавец 5	852	5441
Товарная категория 3	Артикул 13	Наименование 13	Март	Продавец 1	45	65
Товарная категория 3	Артикул 14	Наименование 14	Март	Продавец 1	25	121
Товарная категория 3	Артикул 15	Наименование 15	Март	Продавец 1	9	1785
Товарная категория 3	Артикул 16	Наименование 16	Март	Продавец 4	6	96521
Товарная категория 3	Артикул 17	Наименование 17	Март	Продавец 1	4	4441
Товарная категория 3	Артикул 18	Наименование 18	Март	Продавец 1	7	6412
Товарная категория 3	Артикул 19	Наименование 19	Март	Продавец 5	2	65877
Товарная категория 3	Артикул 20	Наименование 20	Март	Продавец 1	14	2112
Товарная категория 4	Артикул 21	Наименование 21	Март	Продавец 1	78	5512
Товарная категория 4	Артикул 22	Наименование 22	Март	Продавец 1	36	5962
Товарная категория 4	Артикул 23	Наименование 23	Март	Продавец 1	25	442
Товарная категория 4	Артикул 24	Наименование 24	Март	Продавец 1	23	5523



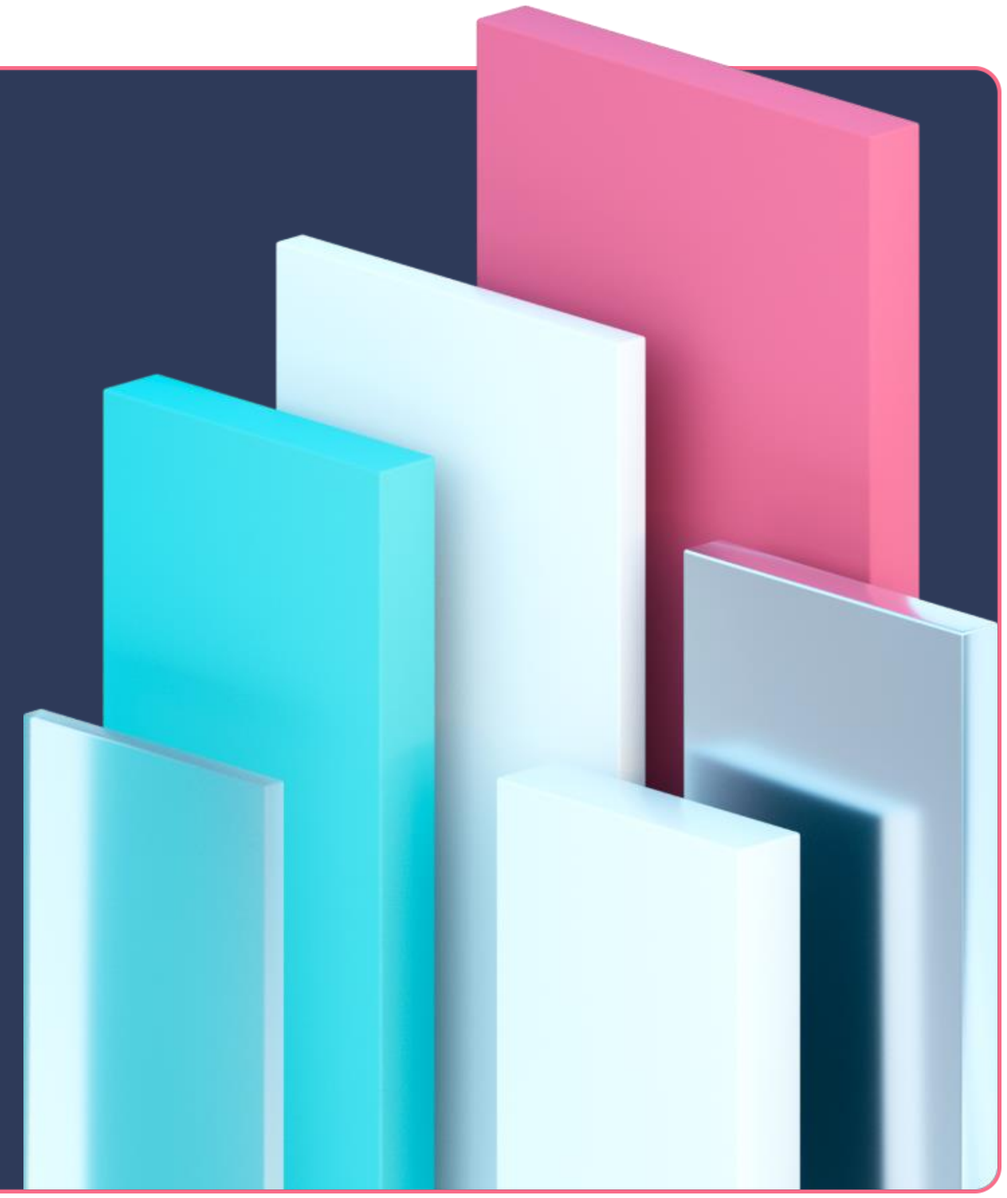
Analytic Workspace

Итоги

- Методология исследования данных CRISP включает шесть этапов: понимание бизнеса, начальное изучение данных, подготовка данных, моделирование, оценка решения и внедрение.

- Поиск данных включает четыре этапа: формулировка запроса, консультации, самостоятельный поиск, запрос и получение данных

- Хранить данные можно на твердых носителях, серверах или в облачных хранилищах.



Analytic
Workspace

Итоги

- Обработка данных включает в себя: первичную обработку и очистку, выделение общих признаков, уплотнение данных, выбор модели для анализа.

- Анализ данных – совокупность действий исследователя, направленных на получение определенных представлений о характере явления, описываемых этими данными.

- Визуализация данных – процесс представления данных в агрегированном, понятном для восприятия человеком виде.

Analytic Workspace

Список литературы

1. Жернакова М.Б., Шестакова И.М. Обоснование управленческих решений на основе данных управленческого учета. Управление. 2016;(4):45-51. https://doi.org/10.12737/22788
2. Коточигов Константин CRISP-DM: проверенная методология для Data Scientist-ов. Хабр. 2017 https://habr.com/ru/company/lanit/blog/328858/
3. Цикл работы с данными по методологии CRISP. CDTOwiki: база знаний по цифровой трансформации. 2020 https://cdto.wiki/Цикл_работы_с_данными
4. Ценность - данные. Большая энциклопедия нефти и газа.2022 https://www.ngpedia.ru/id583341p1.html

Analytic
Workspace

aw-bi.ru

Спасибо
за внимание!