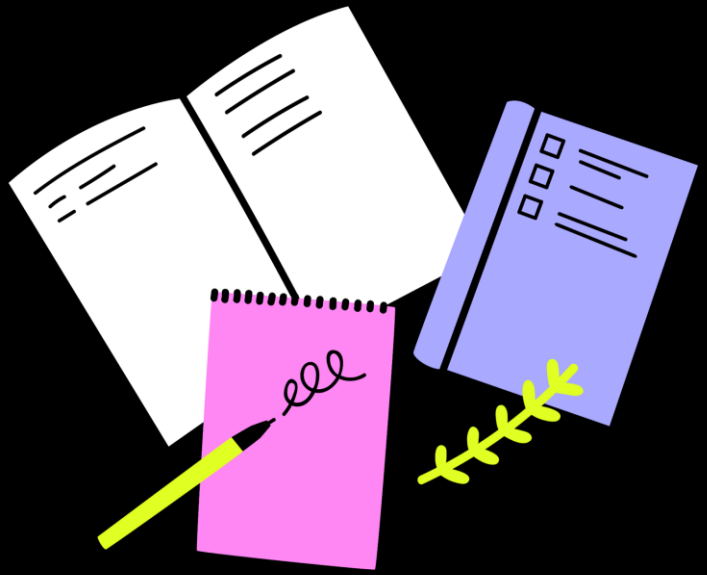


# Как подготовить данные

Как подготовить данные из первоисточника к последующему анализу и визуализации.

# Что будет на уроке сегодня

- Что такое Power Query
- Что является таблицей, принцип табличности
- Источники данных
- Импорт данных
- Универсальный порядок шагов в преобразовании данных
- Типы данных
- Формирование и объединение данных
- Практическая часть

# Что такое Power Query?

# Power Query - это алгоритм по загрузке данных

- Power Query - это инструмент, который позволяет подключиться к источнику данных и их преобразования
- Power Query превращает не табличные данные в табличные
- С помощью Power Query можно создавать ETL процесс (Extract, Transform, Load = «извлечение, преобразование, загрузка»)
- Доступен в Excel и Power BI
- Помогает автоматизировать рутинные функции по подготовке данных (Аналитики тратят до 80% времени на подготовку данных), например транспонирует таблицу, группирует, проставляет форматы данных для столбцов, удаляет пустые значения и ошибки, форматирует столбцы
- Есть возможность писать запросы на языке M

# Что умеет Power Query?

- Загружать данные для дальнейшей проработки
- Консолидировать данные из нескольких листов, папок, файлов
- Трансформировать загруженные данные: сортировать, фильтровать, группировать, сворачивать и так далее
- Объединять таблицы между собой
- Работать с разными форматами: зачистка, исправление регистра, удаление лишних пробелов и так далее. Например актуально для формата номеров телефона
- Выполнять простые вычисления с данными (включая логику ЕСЛИ: если содержит, если больше, если равно и прочее)

# Что не умеет Power Query

- Редактировать загруженные данные напрямую. Если менять данные, то только в источнике
- Производить сложные вычисления (математические и статистические) - это делает DAX: расчет суммы, средней, фильтрация значений и прочее
- Визуализировать. Power Query не про визуализацию, а про данные - это делает Power View

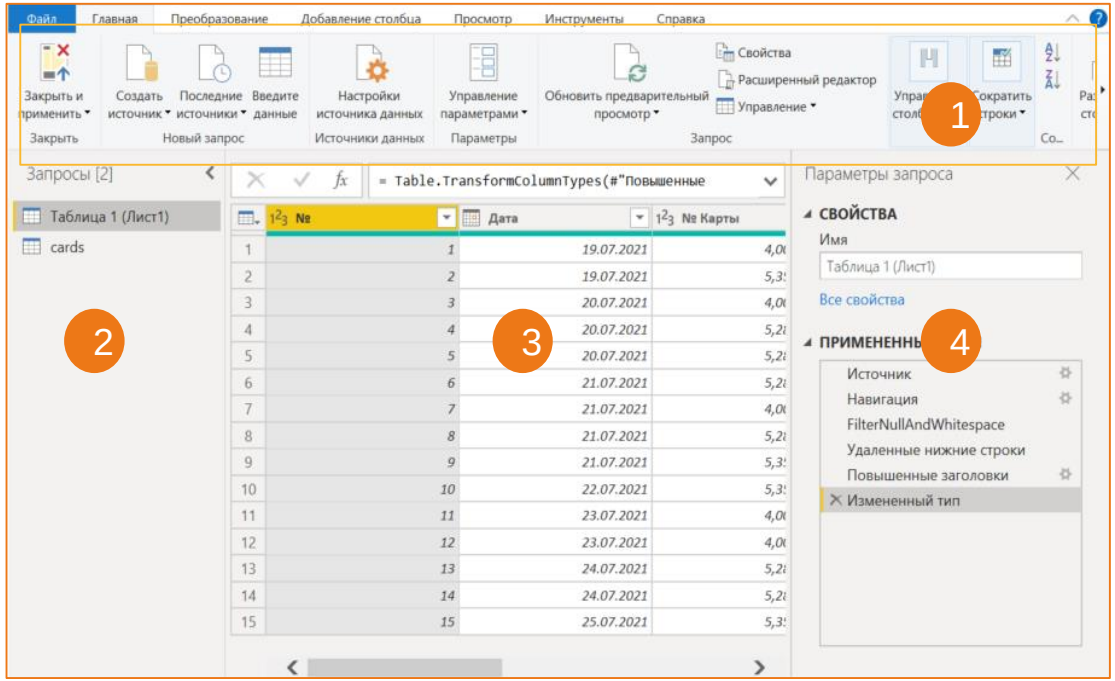
# Интерфейс

1 — кнопки на ленте, которые позволяют взаимодействовать с данными в запросе.

2 — список запросов (по одному для каждой таблицы или сущности). Их можно выбирать, просматривать и настраивать.

3 — рабочая область, где отображаются данные из выбранного запроса, которые можно настраивать.

4 — параметры запроса, где перечислены свойства запроса и применённые действия.



**Одна из основных ценностей Power Query в том, что он преобразовывает таблицы для дальнейшего анализа.**

Что такое таблица?

# Принцип табличности

Давайте определим что является таблицей, а что нет

# Какие есть характеристики у таблицы?

- Одна строка = одно событие или объект
- Столбцы содержат сравнимые показатели
- Для числовых данных в столбце осмысленная арифметическая операция: сумма, деление и прочее
- Нет объединений ячеек
- Таблица состоит из списка заголовков и списка списков значений

Далее мы изучим несколько примеров не таблиц.

# Таблица или не таблица

Давайте определим что является таблицей, а что нет на следующих слайдах и если не является, то что нужно сделать, чтобы преобразовать в таблицу

# Таблица или не таблица?

| Филиал | Класс | Выруч |
|---|---|---|
| Новосибирск | A | 188.72 |
| Новосибирск | B | 237.94 |
| Новосибирск | C | 380.13 |
| Новосибирск | D | 222.75 |
| Новосибирск Итого | | 680.67 |
| Екатеринбург | A | 765.28 |
| Екатеринбург | B | 919.50 |
| Екатеринбург | C | 819.53 |
| Екатеринбург | D | 309.24 |
| Екатеринбург Итого | | 424.64 |
| Ростов-на-Дону | A | 137.73 |
| Ростов-на-Дону | B | 897.55 |
| Ростов-на-Дону | C | 222.87 |
| Ростов-на-Дону | D | 244.51 |
| Ростов-на-Дону Итого | | 401.33 |

## Не таблица

| Филиал | Класс | Выруч |
|---|---|---|
| Новосибирск | A | 188.72 |
| Новосибирск | B | 237.94 |
| Новосибирск | C | 380.13 |
| Новосибирск | D | 222.75 |
| Новосибирск Итого | | 680.67 |
| Екатеринбург | A | 765.28 |
| Екатеринбург | B | 919.50 |
| Екатеринбург | C | 819.53 |
| Екатеринбург | D | 309.24 |
| Екатеринбург Итого | | 424.64 |
| Ростов-на-Дону | A | 137.73 |
| Ростов-на-Дону | B | 897.55 |
| Ростов-на-Дону | C | 222.87 |
| Ростов-на-Дону | D | 244.51 |
| Ростов-на-Дону Итого | | 401.33 |

**Нужно убрать итоговые значения**

# Таблица или не таблица?

| Ответственный | Клиент | Телефон |
|---|---|---|
| Зыков Эрик Юлианович | Галкин Игнатий Васильевич | 123456789 |
| | Сафонов Антон Тимурович | 123456790 |
| | Андреев Панкратий Иринеевич | 123456791 |
| | Гущин Игорь Якунович | 123456792 |
| | Веселов Аввакум Агафонович | 123456793 |
| | Орехов Мирослав Фролович | 123456794 |
| | Фёдоров Тихон Ростиславович | 123456795 |
| | Белозёров Лев Лаврентьевич | 123456796 |
| | Рыбаков Валентин Антонович | 123456797 |
| Фёдоров Варлам Святославович | Петров Фрол Созонович | 123456798 |
| | Авдеев Варлам Германнович | 123456799 |
| | Борисов Виктор Иринеевич | 123456800 |
| | Ермаков Виталий Богуславович | 123456801 |
| | Крылов Касьян Кириллович | 123456802 |
| | Федотов Исак Ильяович | 123456803 |
| | Сафонов Осип Парфеньевич | 123456804 |
| | Киселёв Евдоким Михаилович | 123456805 |
| | Быков Арнольд Митрофанович | 123456806 |
| | Белоусов Григорий Владимирович | 123456807 |
| | Костин Пантелей Станиславович | 123456808 |

# Не таблица

| Ответственный | Клиент | Телефон |
|---|---|---|
| Зыков Эрик Юлианович | Галкин Игнатий Васильевич | 123456789 |
| | Сафонов Антон Тимурович | 123456790 |
| | Андреев Панкратий Иринеевич | 123456791 |
| | Гущин Игорь Якунович | 123456792 |
| | Веселов Аввакум Агафонович | 123456793 |
| | Орехов Мирослав Фролович | 123456794 |
| | Фёдоров Тихон Ростиславович | 123456795 |
| | Белозёров Лев Лаврентьевич | 123456796 |
| | Рыбаков Валентин Антонович | 123456797 |
| Фёдоров Варлам Святославович | Петров Фрол Созонович | 123456798 |
| | Авдеев Варлам Германнович | 123456799 |
| | Борисов Виктор Иринеевич | 123456800 |
| | Ермаков Виталий Богуславович | 123456801 |
| | Крылов Касьян Кириллович | 123456802 |
| | Федотов Исак Ильяович | 123456803 |
| | Сафонов Осип Парфеньевич | 123456804 |
| | Киселёв Евдоким Михаилович | 123456805 |
| | Быков Арнольд Митрофанович | 123456806 |
| | Белоусов Григорий Владимирович | 123456807 |
| | Костин Пантелей Станиславович | 123456808 |

**Нужно заполнить пустые значения**

# Таблица или не таблица?

| Период | Выручка |
|---|---|
| Янв | 88.52к |
| Фев, мар | 33.04к |
| Апр | 96.44к |
| Май | 38.56к |
| Июн, июль,авг | 44.35к |
| Сен | 22.45к |
| Окт | 41.31к |
| Ноя | 37.65к |
| дек | 80.95к |

## Не таблица

| Период | Выручка |
|---|---|
| Янв | 88.52к |
| Фев, мар | 33.04к |
| Апр | 96.44к |
| Май | 38.56к |
| Июн, июль,авг | 44.35к |
| Сен | 22.45к |
| Окт | 41.31к |
| Ноя | 37.65к |
| дек | 80.95к |

**Нужно перенести значение по марту, июлю и августу на отдельные строки**

# Таблица или не таблица?

| Продавец | Квартал1 2020 | Квартал2 2020 | Квартал3 2020 | Квартал4 2020 |
|---|---|---|---|---|
| Зыков Эрик Юлианович | 1693 | 1485 | 1452 | 1000 |
| Фёдоров Варлам Святославович | 1703 | 1598 | 1357 | 947 |
| Котов Антон Вениаминович | 1619 | 1624 | 1568 | 1054 |
| Михайлов Платон Дмитриевич | 1669 | 1489 | 1547 | 1167 |
| Тихонов Григорий Пётрович | 1700 | 1666 | 1331 | 1144 |
| Туров Соломон Всеволодович | 1635 | 1738 | 1539 | 1164 |
| Носков Филипп Владимирович | 1595 | 1713 | 1502 | 1084 |
| Гуляев Моисей Борисович | 1745 | 1774 | 1579 | 1117 |
| Соболев Фрол Иосифович | 1669 | 1595 | 1334 | 1487 |
| Воробьёв Бенедикт Николаевич | 1732 | 1745 | 1329 | 1452 |

# Таблица или не таблица?

| Продавец | Квартал1 2020 | Квартал2 2020 | Квартал3 2020 | Квартал4 2020 |
|---|---|---|---|---|
| Зыков Эрик Юлианович | 1693 | 1485 | 1452 | 1000 |
| Фёдоров Варлам Святославович | 1703 | 1598 | 1357 | 947 |
| Котов Антон Вениаминович | 1619 | 1624 | 1568 | 1054 |
| Михайлов Платон Дмитриевич | 1669 | 1489 | 1547 | 1167 |
| Тихонов Григорий Пётрович | 1700 | 1666 | 1331 | 1144 |
| Туров Соломон Всеволодович | 1635 | 1738 | 1539 | 1164 |
| Носков Филипп Владимирович | 1595 | 1713 | 1502 | 1084 |
| Гуляев Моисей Борисович | 1745 | 1774 | 1579 | 1117 |
| Соболев Фрол Иосифович | 1669 | 1595 | 1334 | 1487 |
| Воробьёв Бенедикт Николаевич | 1732 | 1745 | 1329 | 1452 |

**Формально таблица, но с плохой структурой данных**

# Таблица или не таблица?

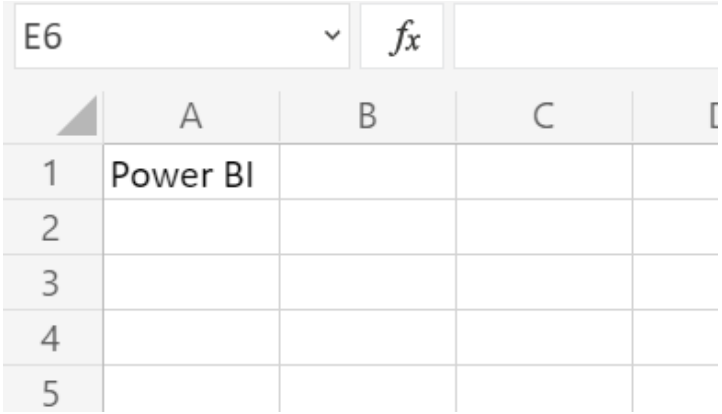
| | 2019 | | 2020 | |
|---|---|---|---|---|
| Филиал | Выручка | Количество | Выручка | Количество |
| Санкт-Петербург | 1000 | 68 | 1693 | 10 |
| Новосибирск | 947 | 50 | 1703 | 69 |
| Екатеринбург | 1054 | 22 | 1619 | 41 |
| Казань | 1167 | 35 | 1669 | 44 |
| Нижний Новгород | 1144 | 37 | 1700 | 35 |
| Челябинск | 1164 | 46 | 1635 | 10 |
| Самара | 1084 | 10 | 1595 | 24 |
| Омск | 1117 | 42 | 1745 | 59 |
| Ростов-на-Дону | 1334 | 90 | 1669 | 48 |
| Уфа | 1329 | 46 | 1732 | 19 |
| Красноярск | 1487 | 44 | 1784 | 18 |
| Воронеж | 1452 | 23 | 1485 | 54 |
| Пермь | 1357 | 68 | 1598 | 56 |
| Волгоград | 1568 | 44 | 1624 | 09 |
| Краснодар | 1547 | 39 | 1489 | 12 |
| Саратов | 1331 | 01 | 1666 | 56 |
| Тюмень | 1539 | 99 | 1738 | 89 |

## Не таблица

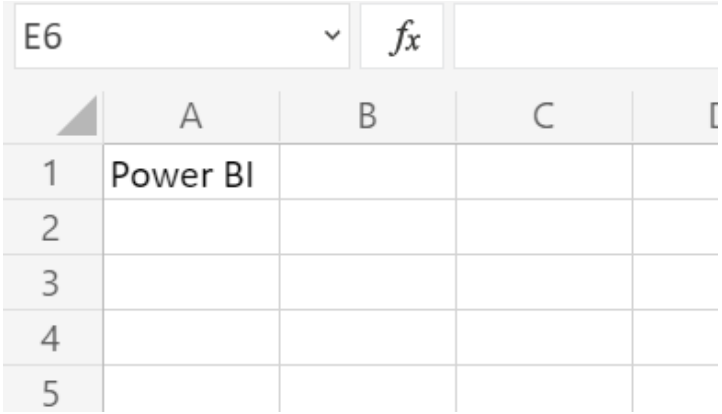
| | 2019 | | 2020 | |
|---|---|---|---|---|
| Филиал | Выручка | Количество | Выручка | Количество |
| Санкт-Петербург | 1000 | 68 | 1693 | 10 |
| Новосибирск | 947 | 50 | 1703 | 69 |
| Екатеринбург | 1054 | 22 | 1619 | 41 |
| Казань | 1167 | 35 | 1669 | 44 |
| Нижний Новгород | 1144 | 37 | 1700 | 35 |
| Челябинск | 1164 | 46 | 1635 | 10 |
| Самара | 1084 | 10 | 1595 | 24 |
| Омск | 1117 | 42 | 1745 | 59 |
| Ростов-на-Дону | 1334 | 90 | 1669 | 48 |
| Уфа | 1329 | 46 | 1732 | 19 |
| Красноярск | 1487 | 44 | 1784 | 18 |
| Воронеж | 1452 | 23 | 1485 | 54 |
| Пермь | 1357 | 68 | 1598 | 56 |
| Волгоград | 1568 | 44 | 1624 | 09 |
| Краснодар | 1547 | 39 | 1489 | 12 |
| Саратов | 1331 | 01 | 1666 | 56 |
| Тюмень | 1539 | 99 | 1738 | 89 |

**Есть группировка по годам - их нужно вынести отдельным признаком в столбец**

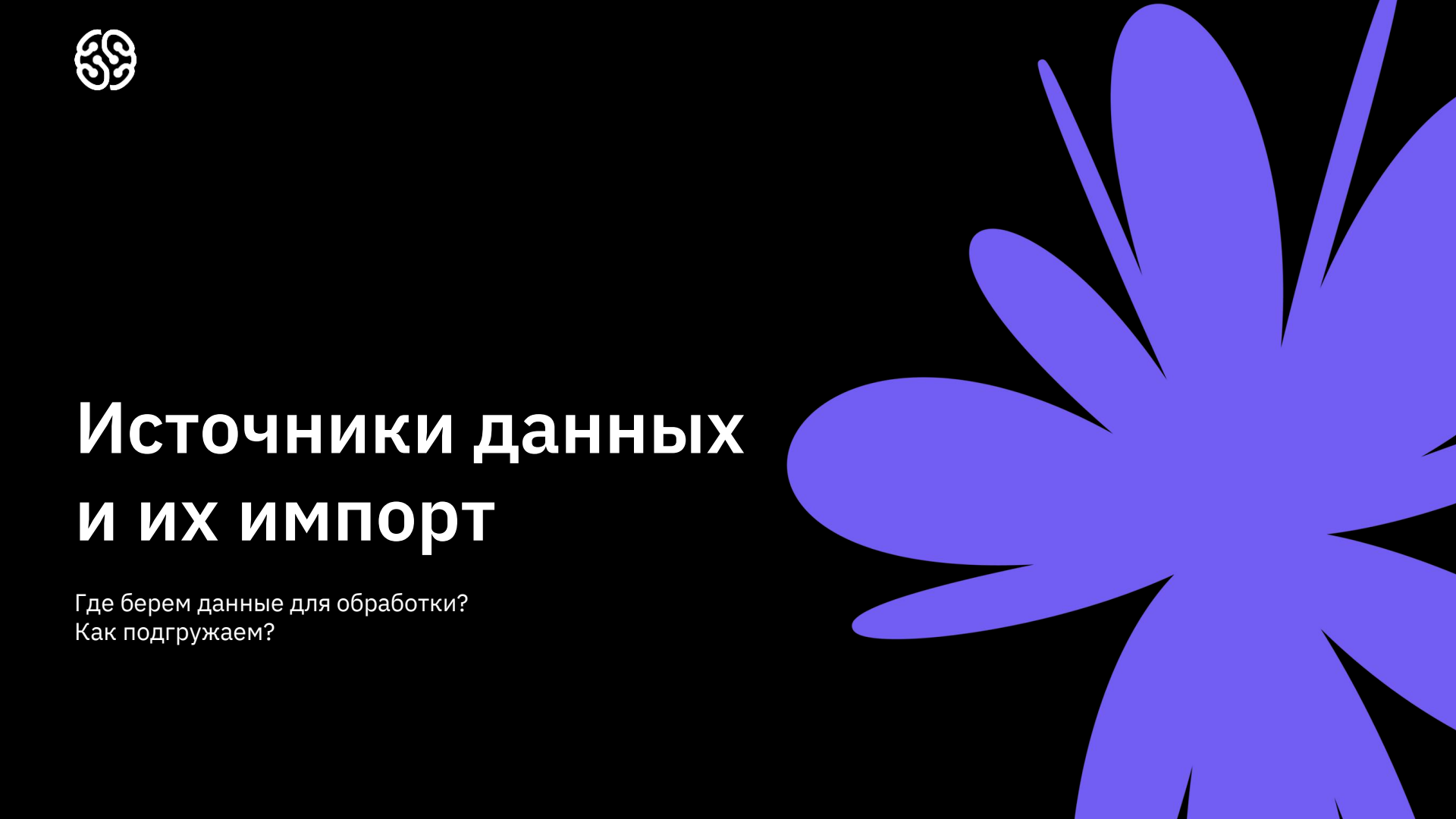
# Таблица или не таблица, если “Power BI” - это заголовок столбца?



# Таблица или не таблица, если Power BI - это заголовок столбца?



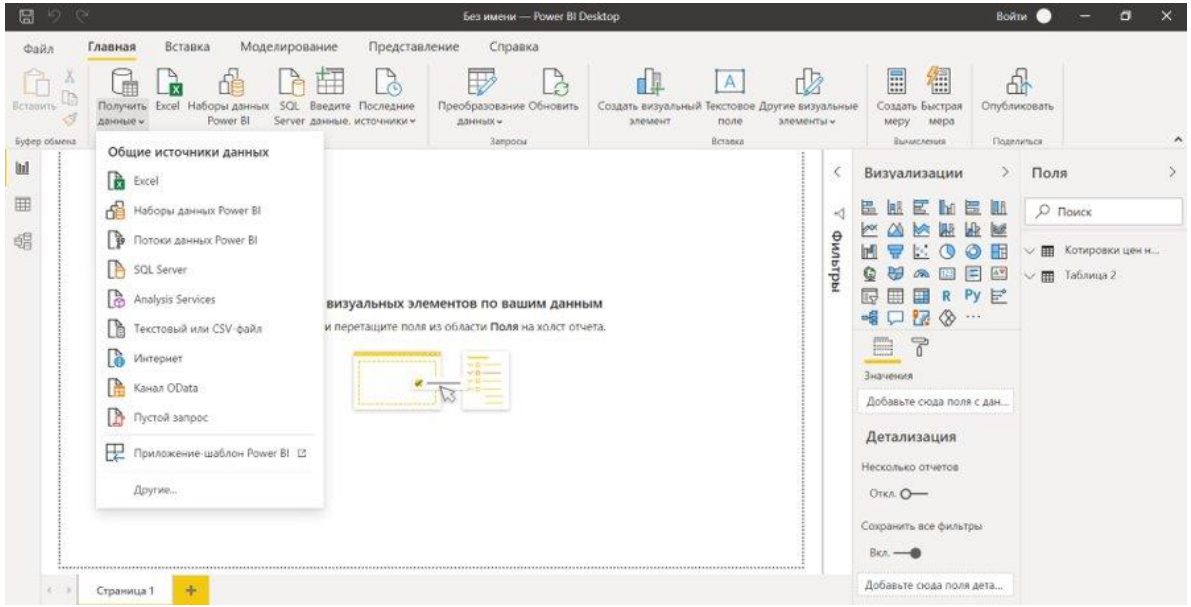

**Формально будет таблицей**



# Источники данных и их импорт

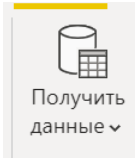
Где берем данные для обработки?
Как подгружаем?

# В Power BI есть коннекторы к десяткам систем

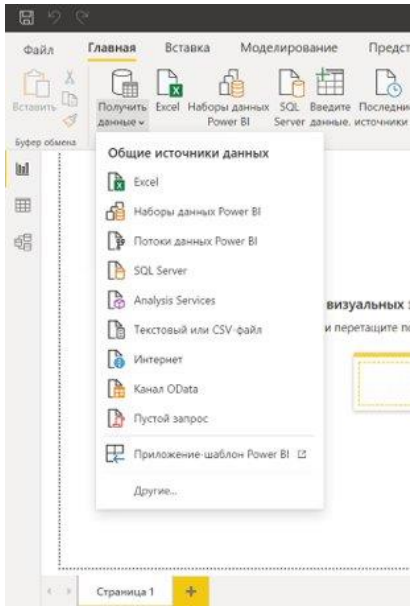


# Загрузка данных

- Выберите Получить данные на вкладке «Главная» Ленты.

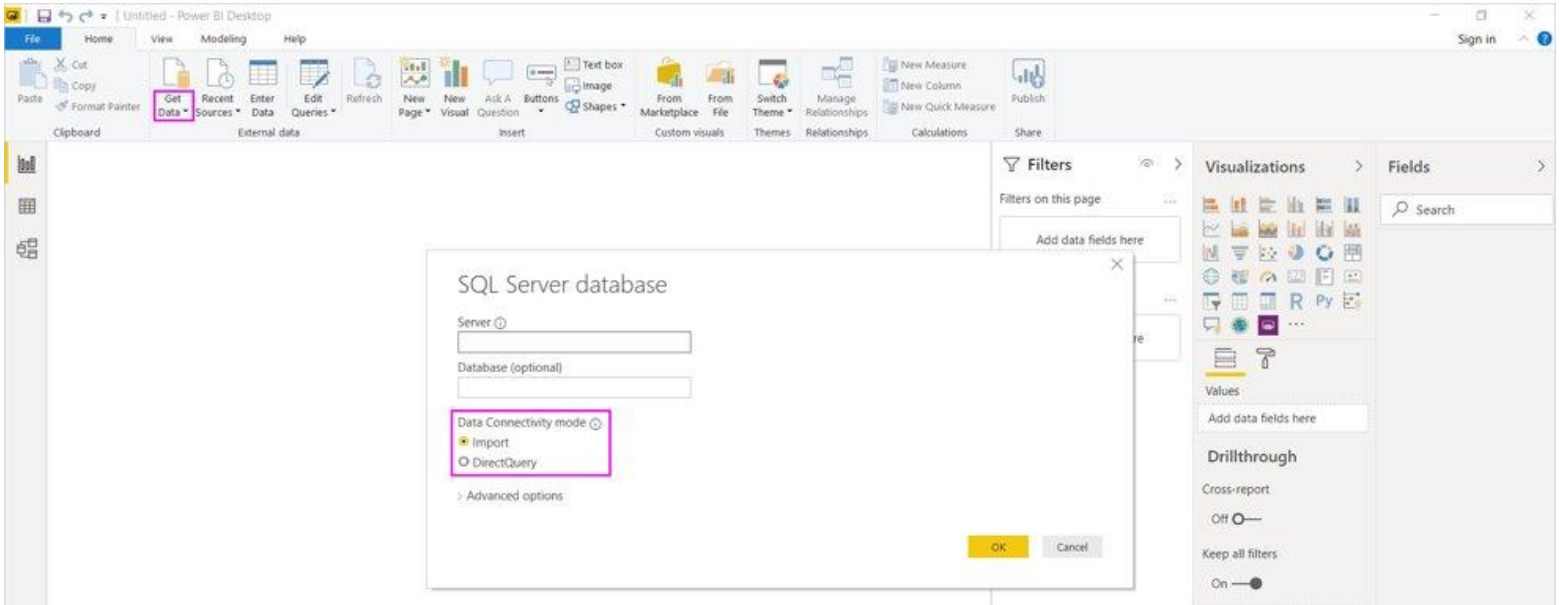



- Выберите нужный тип файла из основного или дополненного списка источников данных.
- Выберите нужный файл в открывшемся окне файлового менеджера.
- Нажмите «Открыть».



# Популярные источники данных

# Метод загрузки данных в Power BI



# Метода загрузки данных

**Import**

- Загружает все данные из базу внутрь Power BI
- Больше весит файл pbix
- Подходит для облачных верси
- Подходит для работы с данными внутри Power BI
- Лучше по скорости работы с данными
- Нет ограничений по работе в DAX

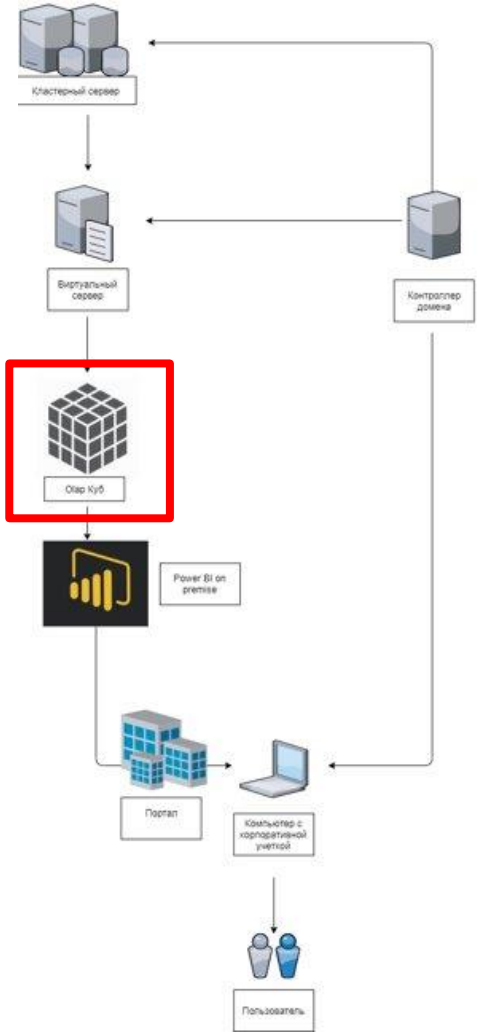
**Direct Query**

- Запрашивает данные напрямую из источника
- Файл pbix весит мало
- Подходит для версий on premise
- Вся работа с данными должна быть проделана до Power BI
- Real-time обновление данных из источника

По умолчанию Import лучше для полноты функционала Power BI, мы будем практиковаться именно с ним.

## Пример архитектуры с Direct Query

**Ключевое отличие - это этап преобразования данных в Olap кубах, для Direct query все преобразования должны быть сделаны до Power BI, Olap кубы или многомерные кубы - это один из вариантов**



# Универсальный подход к обработке данных

Принцип универсального подхода

# Алгоритм создания набора данных

1. **Выбор данных**

Подключаемся к источнику данных

1. **Трансформация (очистка) данных**

Приводим данные к необходимому виду

1. **Загрузка данных**

Определяем, какие именно данные в табличном виде понадобятся к загрузке

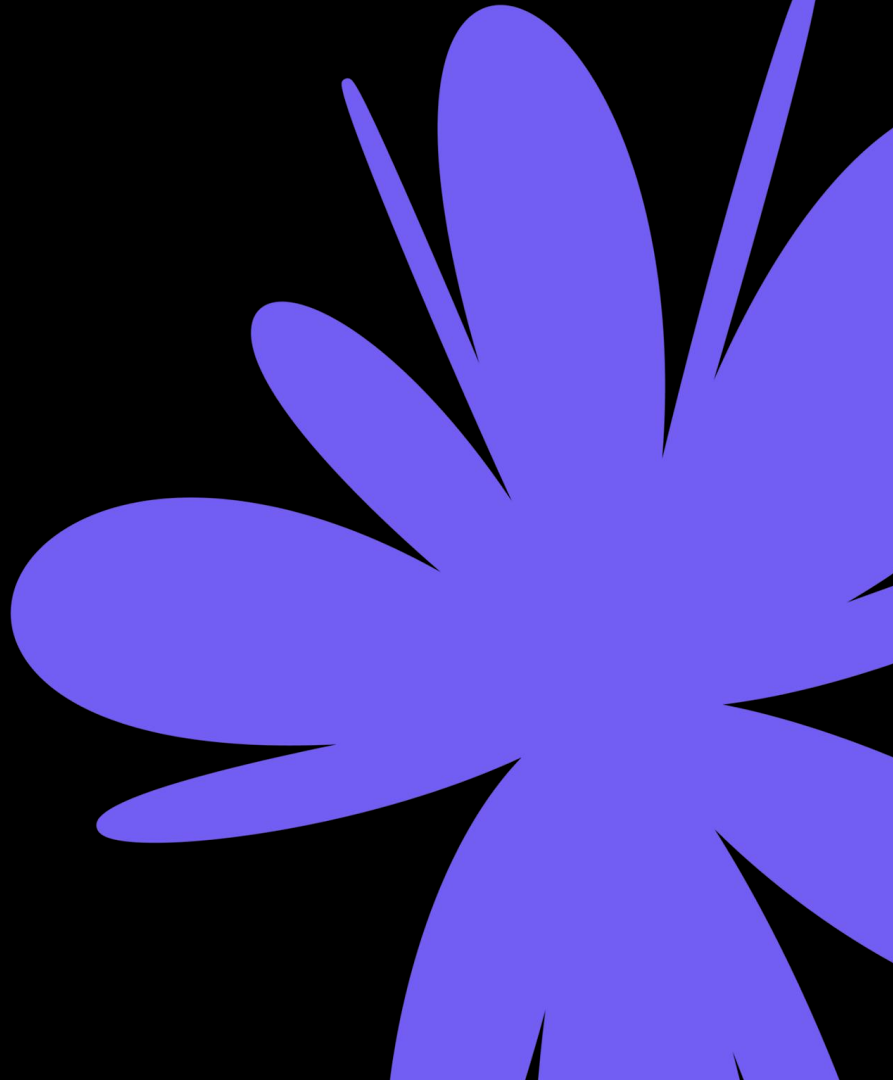
**Важно!** На первом этапе не идет речь про настройку обновления данных.

# Практика: перейдем в интерфейс Power BI и произведем загрузку данных с сайта ЦБ РФ

# Порядок обработки данных

1. Запрос к источнику данных состоит из 4х этапом: Загрузка данных + Обработка/Трансформация + Выгрузка данных + **Обновление данных (!)**
2. Работа с шагами запроса должна быть максимально универсальной
3. Мы должны стремится к таким решениям, чтобы при изменении исходных данных не менять запрос, например учесть все варианты написания телефона, даже если какие-то примеры отсутствуют в текущих данных: номер телефона может начинаться с +7, 8 или сразу с 9, разделять цифры может пробел или “-”
4. Для того, чтобы обновление работало структура данных не должна меняться: порядок столбцов, названия столбцов, название файла или витрины

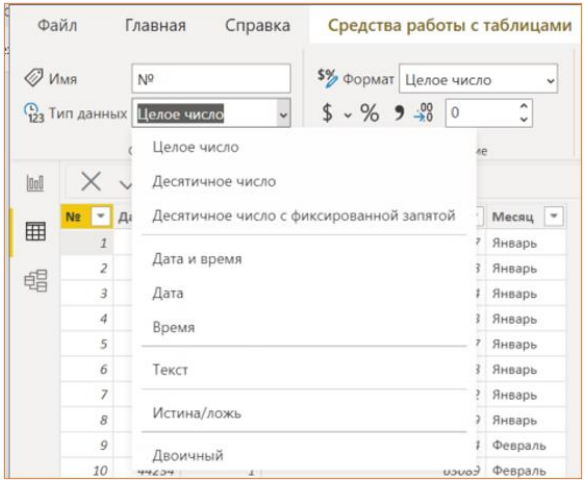
**Важно!** Добавление этапа обновления данных меняет весь подход.



# Типы данных

# Какие есть типы данных?

В большинстве импортируемых наборов данных содержится более одного типа данных. Глобально у нас 3 группы данных: качественные, числовые и дата.




Power BI, как и Excel, поддерживает различные типы данных:

1. целое число
2. десятичное число
3. десятичное число с фиксированной запятой
4. дата и время
5. дата
6. время
7. текст
8. истина/ложь
9. двоичный

От выбранного типа зависит набор операций, которые можно будет выполнить с этими данными.

Например, для числовых данных доступны суммирование, определение среднего арифметического, отображение максимального и минимального значений, определение количества пустых и уникальных значений, стандартное отклонение и др.

# Формирование данных

Join’ы,  типы объединения данных

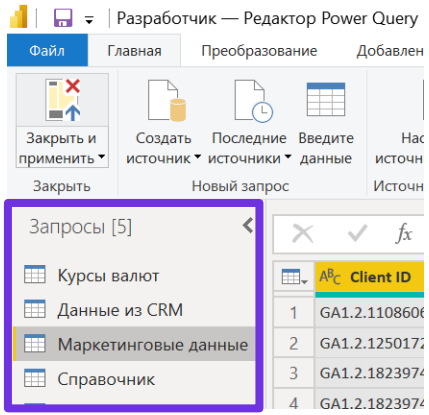
# Формирование и объединение данных

Формировать и объединять данные в Power BI можно с помощью встроенного редактора Power Query.

**Формирование данных** — это преобразование данных, например переименование столбцов или таблиц, замена текста числами, удаление строк, установка первой строки в качестве заголовков и т. д.

**Объединение данных** — это подключение к нескольким источникам данных, формирование данных в соответствии с потребностями и их последующее объединение в один удобный запрос.

В Power Query отдельные таблицы с данными называются запросами.

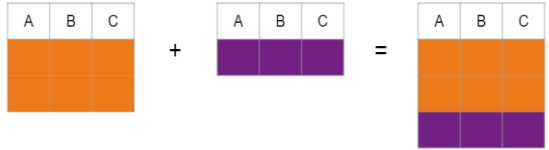


# Способы объединения данных

С помощью Power Query объединить несколько источников данных (например, несколько таблиц MS Excel) можно двумя способами:

## Добавление таблицы под таблицу

Такой подход возможен, если таблицы имеют одинаковую структуру и заголовки (шапки)




## Слияние таблиц

Подход применяется в случае, если содержание одной таблицы дополняет содержание другой, а таблицы имеют разные заголовки (шапки).

# Типы объединения данных

Для объединения данных нужно, чтобы 2 столбца в двух таблицах имели одинаковые значения

Power Query есть шесть способов объединения таблиц:

1. Внешнее соединение слева

   Все из первой таблицы, совпадающие по выбранному столбцу — из второй.

1. Внешнее соединение справа

   Все из второй таблицы, совпадающие — из первой.

1. Полное внешнее

   Все строки из обеих таблиц.

1. Внутреннее

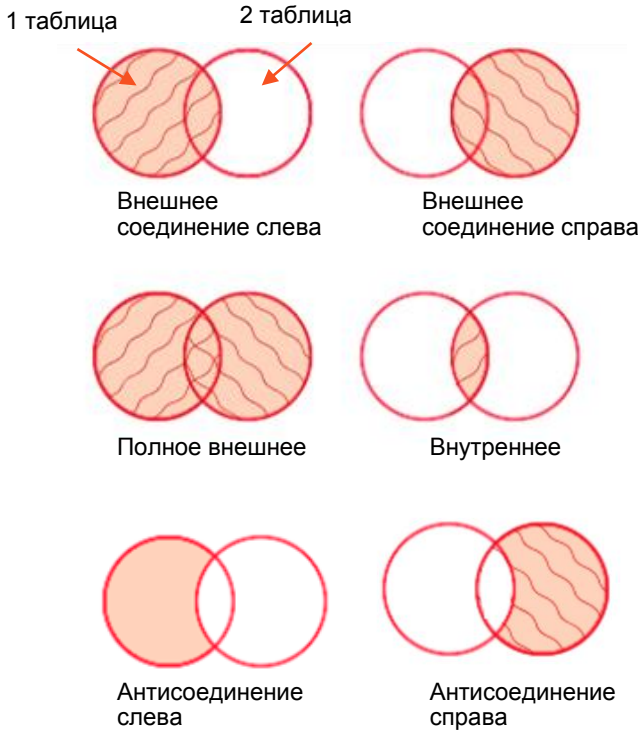
   Только совпадающие строки

1. Антисоединение слева

   Строки, уникальные только для первой таблицы

1. Антисоединение справа

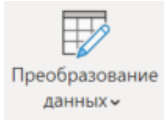
   Строки, уникальные только для второй таблицы




На схеме круг иллюстрирует отдельный источник данных (таблицу), закрашенная область — это те данные, которые попадут в финальный запрос.

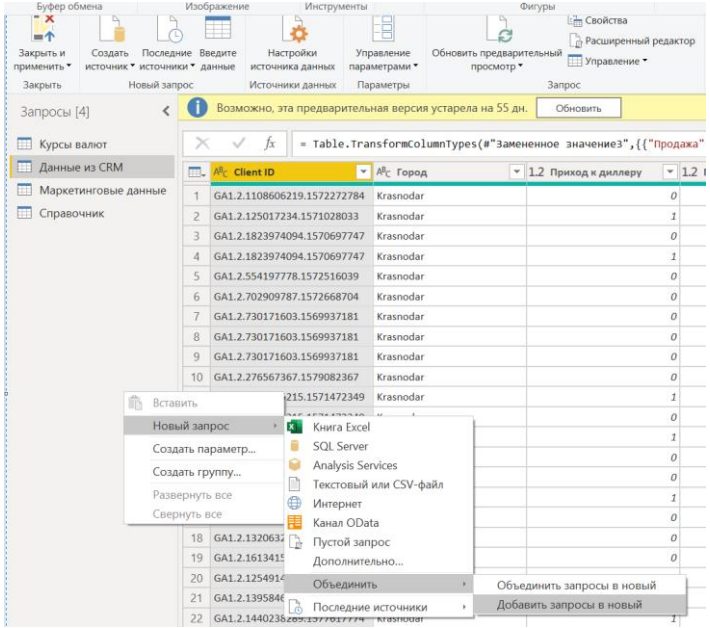
# Добавление данных в запрос

1. Загрузите в качестве источников данных необходимые файлы.
2. Перейдите в редактор Power Query.




1. Нажмите правой кнопкой мыши области Запросы.
2. Выберите Новый запрос → Объединить → Добавить запросы в новый.
3. Выберите таблицы для объединения строк из двух таблиц в одну.
4. Нажмите ОК.
5. Дайте название новой таблице.

Перед выходом из редактора нажмите Закрыть и применить, чтобы сохранить изменения.



# Как это выглядит

1. У нас есть таблица с данными из Google Analytics по посещению сайта
2. У нас есть таблица с данными о продажах
3. Из двух таблиц мы делаем одну

1

| | Client ID | Device Category | Date | Конверсия |
|---|---|---|---|---|
| 1 | 1.1108606219.1572272784 | mobile | 01.01.2020 | 1 |
| 2 | 2.125017234.1571028033 | mobile | 26.01.2020 | 1 |
| 3 | GA1.2.1823974094.1570697747 | mobile | 08.02.2020 | 0 |
| 4 | GA1.2.1823974094.1570697747 | mobile | 16.02.2020 | 1 |
| 5 | GA1.2.554197778.1572516039 | mobile | 08.02.2020 | 1 |
| 6 | GA1.2.702909787.1572668704 | mobile | 25.01.2020 | 0 |
| 7 | GA1.2.730171603.1569937181 | mobile | 05.01.2020 | 0 |
| 8 | GA1.2.730171603.1569937181 | mobile | 22.01.2020 | 0 |
| 9 | GA1.2.730171603.1569937181 | mobile | 02.02.2020 | 0 |
| 10 | GA1.2.276567367.1579082367 | mobile | 15.01.2020 | 0 |
| 11 | GA1.2.1081975215.1571472349 | mobile | 16.01.2020 | 1 |
| 12 | GA1.2.1081975215.1571472349 | mobile | 18.01.2020 | 0 |

+

2

| | Client ID | 1.2 Продажа |
|---|---|---|
| 1 | GA1.2.1108606219.1572272784 | null |
| 2 | GA1.2.125017234.1571028033 | null |
| 3 | GA1.2.1823974094.1570697747 | 0 |
| 4 | GA1.2.1823974094.1570697747 | null |
| 5 | GA1.2.554197778.1572516039 | null |
| 6 | GA1.2.702909787.1572668704 | 0 |
| 7 | GA1.2.730171603.1569937181 | 0 |

=

3

| | Client ID | Device Category | Date | Конверсия | 1.2 Продажа |
|---|---|---|---|---|---|
| 1 | GA1.2.1823974094.1570697747 | null | null | null | 0 |
| 2 | GA1.2.702909787.1572668704 | null | null | null | 0 |
| 3 | GA1.2.730171603.1569937181 | null | null | null | 0 |
| 4 | GA1.2.730171603.1569937181 | null | null | null | 0 |
| 5 | GA1.2.730171603.1569937181 | null | null | null | 0 |
| 6 | GA1.2.276567367.1579082367 | null | null | null | 0 |
| 7 | GA1.2.1081975215.1571472349 | null | null | null | 0 |
| 8 | GA1.2.1081975215.1571472349 | null | null | null | 0 |
| 9 | GA1.2.1537731947.1571378524 | null | null | null | 1 |
| 10 | GA1.2.1067260226.1573748856 | null | null | null | 0 |
| 11 | GA1.2.1611513157.1574872756 | null | null | null | 1 |

В DAX есть схожий функционал связей, изучим на следующем занятии.

# Демонстрация интерфейса

И скачаем данные из Google документа и преобразуем в таблицы

# Практика: перейдем в интерфейс Power BI и Power Query

# Спасибо за внимание