|  |
| --- |
|  |

Оглавление

[**Что такое анализ данных и для чего он нужен**](#_heading=h.jp4n9ikngthd) **4**

[**Этапы работы с данными**](#_heading=h.jyl9ydreyr78) **5**

[**Import vs DirectQuery**](#_heading=h.2s8eyo1) **5**

[**Принципы работы в Power Query**](#_heading=h.86ct1hgpndak) **6**

[Принцип универсальности](#_heading=h.tvs59qg6k4uk) 6

[Принцип таблчности](#_heading=h.qog3xzzazmqf) 6

[**Что умеет и не умеет Power Query**](#_heading=h.8gnc2y4fckr3) **7**

[**Откуда можно загружать данные в PQ**](#_heading=h.efhagn6b7a4) **8**

[**Разбор примера из практики**](#_heading=h.cnvr9vd6mmnh) **8**

[**Решаемые задачи**](#_heading=h.vdn9odz8azdj) **10**

[**Типы слияния в Power Query**](#_heading=h.adj4ghsxbwvw) **10**

[Добавление](#_heading=h.rshvx27zb04l) 11

[Объединение](#_heading=h.iu46ycuofoaw) 11

[Типы объединения](#_heading=h.6493g34uk6no) 12

[Результаты типов объединений](#_heading=h.97m1frtzu0p4) 14

[**Ссылки и дублирование запросов**](#_heading=h.nyyncz39gl6x) **14**

[Дублирование](#_heading=h.nxa1e4yppmle) 15

[Ссылка](#_heading=h.q32dhc9qbmf2) 15

[Копирование](#_heading=h.h2xt3q4hhnp1) 15

[**Работа с датами в PQ**](#_heading=h.j20t4njfl119) **16**

[Распознавание дат](#_heading=h.qytonxjp7um) 16

[Преобразование дат](#_heading=h.1tnq1lrxd41u) 16

[Вычисления с датами](#_heading=h.2ujy20m7hjrj) 17

[**Глоссарий**](#_heading=h.tyjcwt) **18**

[**Дополнительные материалы**](#_heading=h.llkhqlmoefdi) **18**

Введение

На этой лекции вы найдете ответы на такие вопросы как / узнаете:

* 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

Термины, используемые в лекции

**Power Query** (PQ) — это технология подключения к данным. С её помощью можно обнаруживать, подключать, объединять и уточнять данные из разных источников для последующего анализа. Функции Power Query доступны в классических версиях Excel и Power BI.

**Анализ данных** — средство проверки гипотез и решения задач для исследователя данных.

**Аналитик** — специалист в области анализа данных, который на достаточной уровне владеет инструментальными и программными средствами анализа данных.

**Создание набора данных** — этап исследования, в котором данные нужно загрузить и подготовить к анализу.

**Моделирование данных** — этап исследования, в котором данные нужно проанализировать (найти ответы на стоящие вопросы).

**Подключение** — создание подключений к данным, которые хранятся в облаке, в какой-либо службе или на локальном компьютере.

**Преобразование** — трансформация или формирование данных в соответствии с поставленной задачей.

**Шаги запроса** — итерации при работе в PQ, формирующие загрузку, трансформацию и выгрузку данных.

**Добавление запросов** — стыковка двух и более таблиц с одинаковым наименованием столбцов. В Power Query операция добавления создаёт новый запрос, содержащий все строки из первого запроса, за которым следуют все строки из второго запроса.

**Объединение запросов** — это поиск и подстановка данных одной таблицы в другую по заданным параметрам.

# Что такое анализ данных и для чего он нужен

Анализ данных — достаточно широкое понятие, оно насчитывает множество определений. Как пользователи Power Query, выделим для себя одно:

|  |
| --- |
| **Анализ данных** — это средство проверки гипотез и вопросов, которые ставит перед нами наша деятельность, а также их решения. |

**Важно!** Всегда следует помнить: сами по себе гипотезы и вопросы в данных не содержатся. Если нам, как аналитикам, не хватает экспертности в данном бизнесе, нужно работать вместе с экспертом области, которую мы исследуем.

Эксперт — ключевая фигура в процессе анализа. Аналитик — связующее звено между экспертами и данными. Он собирает у экспертов гипотезы, а затем выдвигает требования к данным. Эффективные аналитические решения можно получить при взаимодействии ЭКСПЕРТ-ГИПОТЕЗА-АНАЛИТИК.

Исключая научную часть исследования данных, задача аналитика — помочь руководителю в принятии управленческого решения. Например, рассчитать эффективность (рентабельность) работы сети магазинов, просчитать окупаемость закупки, сегментировать клиентов, рассчитать, была ли акция успешной, проверить, влияет ли доход клиента на кредитоспособность и так далее.

# Этапы работы с данными

При верхнеуровневом подходе к исследованию данных, его можно разделить на два этапа — создание набора данных и моделирование.

Создание набора данных — этап исследования, в котором данные нужно загрузить и подготовить к анализу.

Данные, которые агрегируются в компаниях в различных источниках (например, в базах данных), часто имеют свои особенности, так как редко накапливаются для решения задач анализа, а скорее для коммерческих целей: ведения учёта, составления отчётности. Такие данные, как правило, содержат ошибки, аномалии, противоречия. Следовательно, одна из прикладных задач аналитика — это подготовка качественных данных для анализа. Под качественными будем понимать полные и достоверные данные. Ещё одна задача аналитика — это работа с большими объёмами данных, которые, например, Excel уже не способен обрабатывать.

**Важно!** Технология Power Query на данном этапе помогает справится с поставленными задачами.

**Задачи решаемые на этапе создания набор данных:**

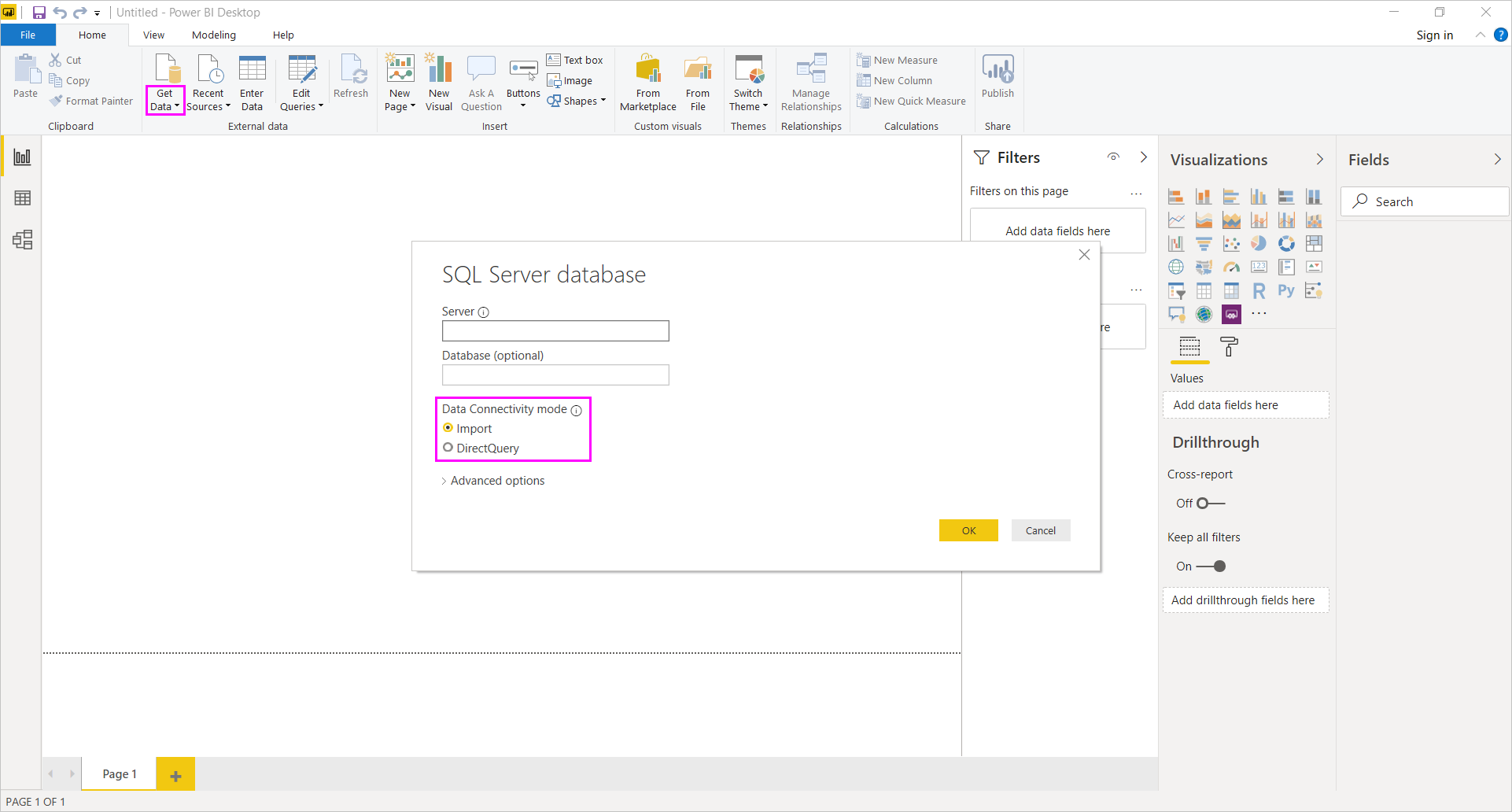
* выбор данных — подключение к источнику данных;
* очистка данных — преобразование и приведение данных к нужному виду;
* загрузка данных — определение того, какие именно таблицы необходимо загружать;
* после подготовки данных необходимо приступать к моделированию — этапу исследования, в котором данные анализируются (ищутся ответы на стоящие вопросы).

# Import vs DirectQuery

Import и DirectQuery – это способы забора данных в Power BI:

|  |  |
| --- | --- |
| **Import:**   * загружает все данные из базы внутрь Power BI; * больше весит файл pbix; * подходит для работы с данными внутри Power BI. | **DirectQuery:**   * запрашивает данные напрямую из источника; * файл pbix весит мало; * вся работа с данными должна быть проделана до Power BI. |

Выбор метода забора данных необходимо сделать при загрузке данных из крупного источника, например, SQL Server:



# 

# Принципы работы в Power Query

Общую логику работы в Power Query можно представить следующим образом: сначала мы загружаем данные в Power Query, потом приводим в нужный вид, а затем выгружаем. В будущем данные итерации можно повторить, просто обновив запрос. Работая с этим алгоритмом действий, следует придерживаться двух принципов:

* универсальности;
* табличности.

## Принцип универсальности

Работа с шагами запроса в Power Query должна быть максимально универсальной. То есть мы должны стремится к таким решениям, чтобы при изменении исходных данных не менять запрос.

## Принцип таблчности

Все данные после проработки в Power Query должны быть выгружены в виде таблиц. Таблица — это способ структурирования данных по правилам:

* одна строка = одно событие или объект;
* столбцы содержат сравнимые показатели;
* для данных в столбце осмысленная арифметическая операция;
* нет объединений;
* таблица состоит из списка заголовков и списка списков значений.

# Что умеет и не умеет Power Query

Чтобы лучше понимать, какие задачи предстоит решать в Power Query, разберём, какие манипуляции с данными здесь можно проводить:

1. **Загружать данные почти из любых источников**: файлов, баз данных, интернета и так далее.
2. **Собирать данные из нескольких листов, папок, файлов.** Консолидировать данные из разных источников (например, из папки).
3. **Трансформировать загруженные данные** — производить итерации типа сортировки, фильтрации, группировки, сворачивания или разворачивания столбцов по заданному признаку и многое другое.
4. **Связывать и объединять таблицы между собой**. Формировать данные на основании нескольких уже загруженных таблиц.
5. **Работать с датами и текстом.** Производить итерации типа зачистки, исправления регистра, удаления лишних пробелов, вычисления длины строки и так далее.
6. **Выполнять простые вычисления с данными.** Производить арифметические операции со столбцами, вычислять среднее, а также другие простые операции, включая логику ЕСЛИ.

**При этом мы не можем сделать следующее:**

1. **Редактировать данные, загруженные напрямую**. То есть, менять данные нужно в источнике, из которого они загружаются.
2. **Выполнять сложные статистические и математические расчёты.** Сумму, среднее, медиану и другие простые вычисления можно произвести, однако сложные финансовые вычисления — нет. После выгрузки данных из PQ в PB их можно добавить. Но это уже дальнейшая работа с формулами Dax.
3. **Визуализировать** — Power Query не про визуализацию, а про данные. Наглядно представить данные диаграммами можно на следующих этапах работы в Power BI (после выгрузки данных из PQ в PB).

# Откуда можно загружать данные в PQ

**Базы данных** — наиболее лёгкий и удобный путь получения информации, так как уже содержит упорядоченные данные. PQ умеет работать с большинством баз данных: Oracle, MySQL, Microsoft SQL Server, PostgreSQL,MongoDB, Microsoft Access и так далее.

**Интернет** —в Power Query есть возможность подгружать нужные данные из веб-страниц.

**ERP,CRM** —по умолчанию PQ умеет запрашивать и соединять данные из распространённых программ (SAP, Microsoft Dynamics) и CRM-системы Salesforce. 1С поддерживают обмен данными по протоколу OData (Open Data Protocol) — специально разработанному открытому набору правил для запросов и обновления данных между компьютерами по сети.

**Файлы:**

* данные из файлов Microsoft Excel любых типов: XLS, XLSX, XLSB, XLSM;
* общепринятые веб-форматы: HTML, XML, JSON;
* обычный TXT или разделённый запятыми или точкой с запятой CSV;
* возможен импорт из PDF (с промежуточным этапом в виде распознавания PDF в Microsoft Word).

# Разбор примера из практики

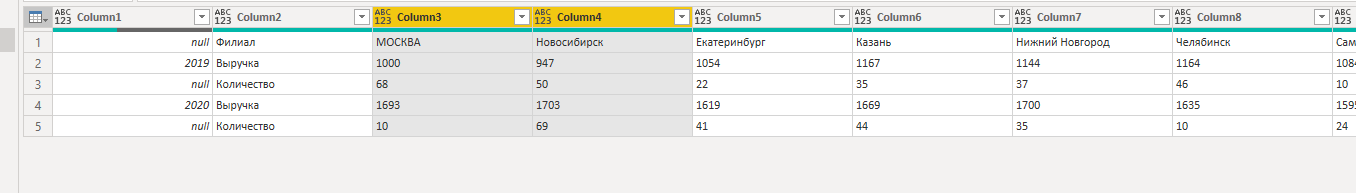
****

Данные находятся в не табличном виде, так как есть объединения, количество и выручка в текстовом формате и так далее.

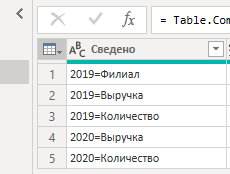
**Задача** — привести данные в табличный вид.

**Алгоритм действий:**

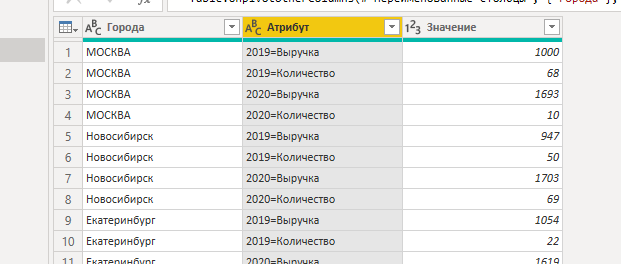
1. Транспонируем данные.



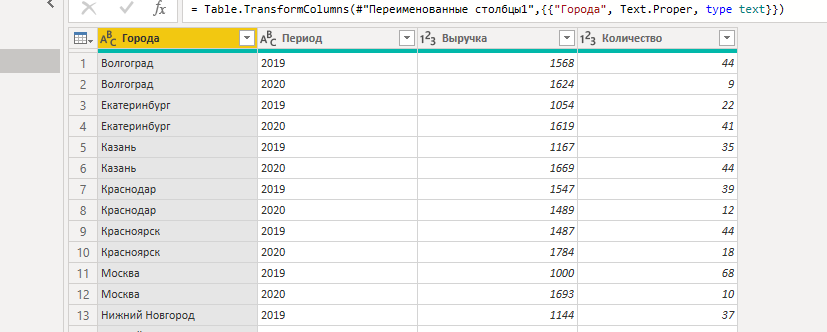
1. Заполняем в колонке с периодом (первой) значения, кнопками ВНИЗ, затем ВВЕРХ.
2. Объединяем первый и второй столбец (колонку с периодом и наименованием значений) через любой пользовательский символ.



1. Транспонируем данные.
2. Повышаем заголовок (кнопкой «‎Использовать первую строку в качестве заголовка»‎).
3. Отменяем свёртывание других столбцов.



1. Разъединяем столбец «Атрибут»‎.
2. Меняет формат данных на текстовый у столбца с периодом.
3. Производим итерацию «Столбец сведения».
4. Меняем заголовки таблицы.



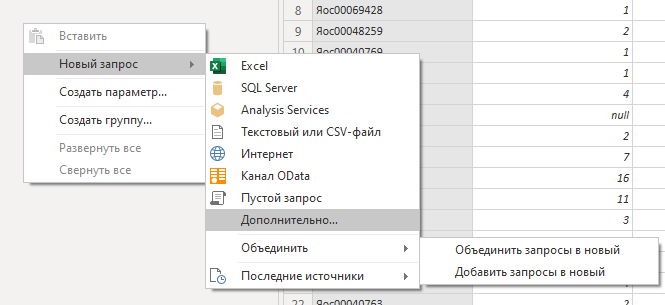
# Решаемые задачи

1. **Данные находятся в разных таблицах**. Например, мы работаем в организации с сетью магазинов по всей стране. Каждый магазин формирует свой файл по продажам, и нам необходимо консолидировать данные.
2. **Результат одного запроса может быть исходным для другого**. Например, если у нас есть консолидированные данные, но необходимо, не меняя источник, исключить, например, Москву.

# Типы слияния в Power Query

В PQ существует два типа слияния запросов (таблиц) — объединение и добавление.

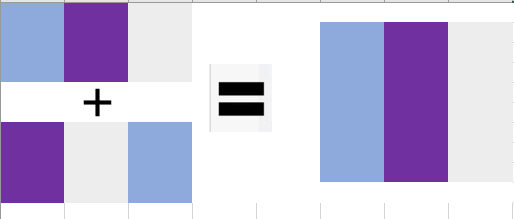
Обе команды доступны на вкладке «‎‎Главная» либо при создании нового вопроса в соответствующем пункте меню. То есть операцию и объединения, и добавления можно произвести, как создавая новый запрос на основании уже имеющихся (не производить итераций над уже имеющимися запросами), так и путём итераций над запросами (добавляя и объединяя с существующим запросом другие).



## Добавление

Главная задача, которую решает добавление запросов (таблиц) — это автоматическая консолидация нескольких таблиц (например, нескольких листов одной книги Excel, нескольких таблиц из разных книг, консолидация запросов из разных источников и так далее).

*Важно! Добавлять можно 2 и более таблиц. Названия добавляемых столбцов не должны совпадать. Порядок и количество столбцов в каждой таблице могут быть различными.*



Если у нас есть несколько таблиц с разными наименованиями столбцов, то добавление консолидирует данные тех столбцов, наименования которых совпадают. А тех, которые не совпадают, добавит справа в общем запросе.

## Объединение

Основная задача, которую решает объединение — это поиски и подстановка данных одной таблицы в другую по заданным параметрам. Например, объединение запросов по левому внешнему соединению — это аналог функции ВПР в Excel.

*Важно! Названия столбцов не имеют значения. Объединение можно производить как по одному столбцу, так и по связке нескольких столбцов.*

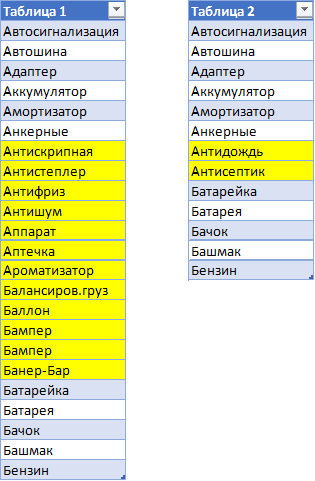
В Power Query есть несколько способов объединения: внешнее левое, внешнее правое и так далее.

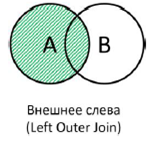
Объединение за один раз можно произвести только с двумя таблицами, то есть если мы хотим объединить информацию из трёх таблиц (запросов), то нам придётся сделать два объединения.

Объединение таблиц (запросов) происходит по ключу — тому столбцу, который есть и в первой таблице, и во второй. Или по связке таких столбцов.

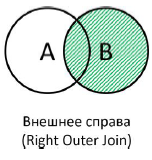
### Типы объединения

Разберём типы всех объединений в Power Query на примере двух простых таблиц.



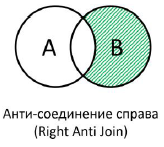


**Внешнее слева** — аналог функции ВПР в Excel. В результате объединения двух таблиц получим все элементы из «Таблицы1»‎ и в дополнительном столбце найденные совпадения из «Таблицы2»‎. Данный тип соединения часто используется в работе с данными.

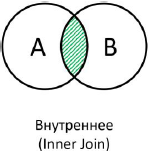


**Внешнее справа** — в результате объединения с данными из «Таблицы2» найдутся совпадающие данные «Таблицы1». Это обратный вариант соединения «внешнее слева»‎.

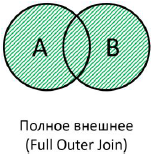
**Анти-соединение слева** — результат соединения, те позиции, которые есть в «Таблице1» и при этом отсутствуют в «Таблице2».



**Анти-соединение справа** — выводит данные «Таблицы2» за вычетом данных из «Таблицы1».

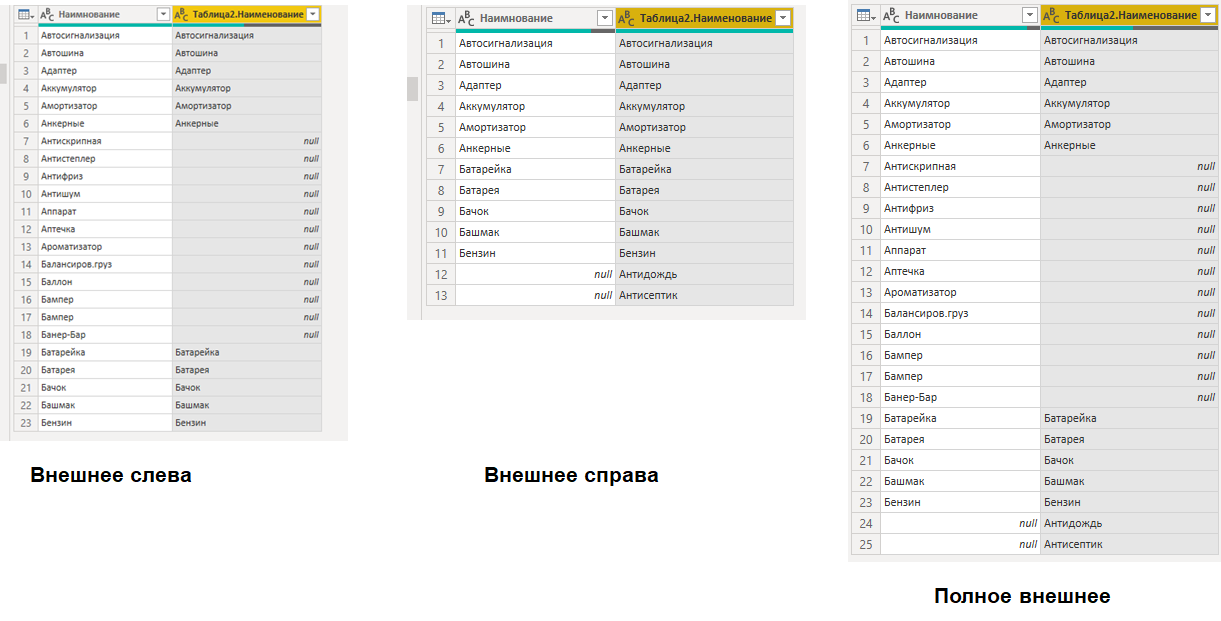


**Внутреннее** — тип слияния выводит только совпадающие данные в двух таблицах.



**Полное внешнее** — выводит данные из обеих таблиц. То есть те, что совпадают, встанут друг напротив друга в одной строке, а остальные будут иметь в паре NULL.

### Результаты типов объединений

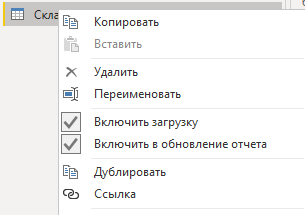


# 

# 

# Ссылки и дублирование запросов

Есть ещё ряд полезных итераций с запросами: дублирование, ссылка на запрос или копирование запроса. Все эти итерации доступны, если щёлкнуть правой кнопкой мыши на запрос в PQ.



Данные манипуляции решают ряд задач. Разберём их ниже.

## Дублирование

Например, у есть запрос, состоящий из множества шагов, но нам необходимо те же шаги сделать в другой таблице (изменить источник данных). Для этого необходимо дублировать исходный запрос и с помощью настроек изменить источник данных.

*Важно! При дублировании запросы совершенно независимы друг от друга, то есть правки, вносимые в один из запросов, не повлияют на другой.*

## Ссылка

Например, у нас есть консолидированные данные в одном общем запросе.

Один руководитель просит сделать одни расчёты с этими данными, второй — совершенно другие. В этой ситуации можно сделать две ссылки из главного источника данных и на их основании построить нужные отчёты.

*Важно! Ссылается на другой запрос, делая его источником, то есть при изменении шагов запроса-источника, поменяются данные и в запросах-ссылках.*

## Копирование

Команда «‎копировать»‎ работает аналогично дублированию, однако есть ключевое различие.

Если применить копирование к обычному запросу, который никак не связан с другими запросами, мы получим просто его копию. Однако если запрос имеет связи, то копирование создаст копию запроса и всех влияющих на него связей.

# Работа с датами в PQ

PQ легко справляется с задачами при работе с данными типа «‎Дата»‎ и «‎Дата и Время»‎. За эти преобразования отвечают кнопки «‎Дата» и «‎Время», находящиеся на вкладках ‎«Преобразование» и «Добавление столбца».

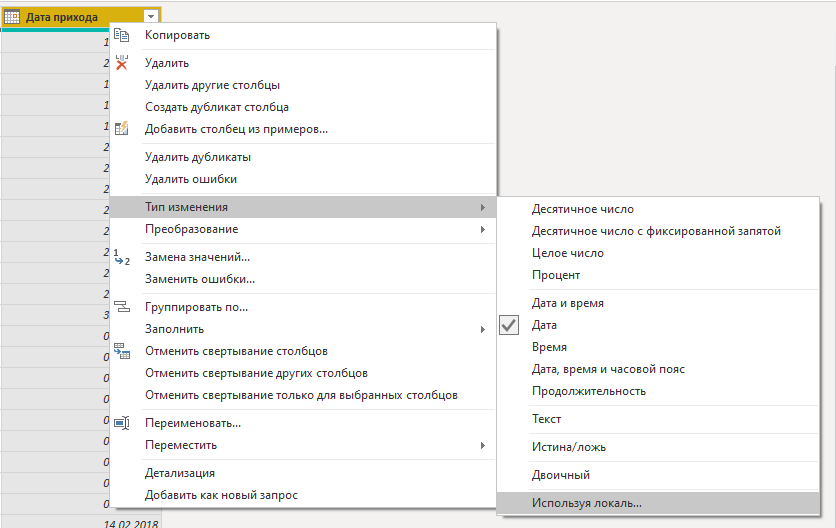
**Задачи которые можно решить:**

1. Извлечь из даты отдельные составляющие (день, месяц, год, квартал и так далее);
2. Определить номер недели по дате;
3. Произвести вычисления.

## Распознавание дат

Почти все текстовые форматы записи дат распознаются стандартными средствами PQ, даже если дата была записана в обратной последовательности.

*Важно! Не распознаются даты, если они идут в другом региональном формате, а ваш ПК имеет стандартные российские настройки (и наоборот). В этом случае необходимо воспользоваться командой «‎Используя локаль»‎:*



## Преобразование дат

Есть целый ряд возможностей преобразования дат в PQ.

**Раздел «День»‎**

* День — номер дня в месяце.
* День недели — порядковый номер дня недели, начиная с нуля, т. е. Пн=0, Вт=1 и т. д.
* День года — порядковый номер дня в году.
* Начало дня — дата-время начала дня (полночь).
* Конец дня — дата-время начала следующего дня (полночь).
* Название дня — день недели словом.

**Раздел «Неделя»‎**

* Неделя года — номер недели в году.

*Важно! Power Query использует нумерацию недель по американскому стандарту, который отличается от принятого в России и Европе стандарта ISO 8601.*

* Неделя месяца — порядковый номер (начиная с 1) недели в месяце.

**Раздел «Квартал»‎**

* Квартал года — номер квартала в году числом от 1 до 4.
* Начало квартала — дата первого дня квартала.
* Конец квартала — дата последнего дня квартала.

**Раздел «Месяц»‎**

* Месяц — номер месяца от 1 до 12.
* Начало месяца — дата первого дня текущего месяца и года.
* Дней в месяце — количество дней в месяце от 28 до 31.
* Название месяца — название месяца словом.
* Конец месяца — дата последнего дня месяца

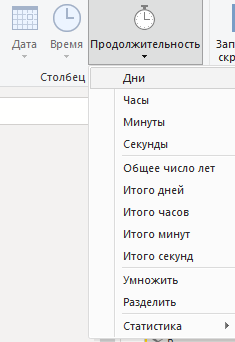
**Раздел «Год»‎**

* Год — номер года в числовом формате.
* Начало года — дата первого дня (1 января) текущего года.
* Конец года — дата конца года.

## Вычисления с датами

Часто встречающаяся задача — это вычисление разницы между двумя датами: возраст, стаж, срок хранения, длительность доставки и так далее. PQ легко работает с такими вычислениям путём простых арифметических операций.

*Важно! В PQ есть тип данных «‎Продолжительность»‎ для хранения длительности и продолжительности какого-либо процесса. Когда мы производим арифметические операции с датами, результатом будет тип данных «‎Продолжительность»‎. Для дальнейшего форматирования и преобразования этого типа данных‎, на вкладке преобразования используйте выпадающий список «‎Продолжительность».*



Частным случаем подсчёта длительности является вычисление возраста (разницы между сегодня и датой рождения). Для этого вычисления существует отдельный пункт меню «Возраст», который вычисляет, сколько дней, часов, минут, секунд прошло с даты рождения. Чтобы привести к необходимому формату, надо воспользоваться пунктом меню «‎Продолжительность»‎.

# Дополнительные материалы

* [Power Query — обзор и обучение](https://support.microsoft.com/ru-ru/office/power-query-%E2%80%94-%D0%BE%D0%B1%D0%B7%D0%BE%D1%80-%D0%B8-%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5-ed614c81-4b00-4291-bd3a-55d80767f81d)
* [Подключение к данным](https://docs.microsoft.com/ru-ru/power-bi/fundamentals/desktop-what-is-desktop)
* [Блог Power Bi](https://powerbi.microsoft.com/ru-ru/blog/)