**Заключительный проект в Power Query: объединяем все вместе**

Это фрагмент книги [Гил Равив. Power Query в Excel и Power BI: сбор, объединение и преобразование данных](https://baguzin.ru/wp/?p=24451).

[Предыдущий раздел](https://baguzin.ru/wp/?p=24709) [К содержанию](https://baguzin.ru/wp/?p=24451#content)

Примите поздравления. Вы приступаете к последней главе. Пришло время подтвердить ваши навыки по обработке данных и проверить возможности в заключительном проекте.

#### Отчет о доходах компании Wide World Importers

Вы – специалист по данным в компании Wide World Importers. После атаки хакеров вас попросили восстановить отчет о доходах компании на основе старых данных. Ваш предшественник вел отчеты в трех файлах Excel за 2015-2017 годы. В каждом из них содержалось 12 листов – по одному на каждый месяц. А выручка 2018 года была экспортирована в формате csv. Нужно импортировать данные с 36 листов, объединить их с данными 2018 года и создать единый отчет.

Загрузите папку C14E01 и целевую таблицу в файле C14E01 - Goal.xlsx. Изучите файлы 2015.xlsx, 2016.xlsx и 2017.xlsx. Каждый лист содержит сводную таблицу формата 2\*3:



Рис. 1. Исходные данные за 2015–2017 гг. в файлах Excel

В строках представлены категории и города клиентов, а в столбцах – цвет продукта, категория и наименование поставщика.

Данные о доходах за 2018 году находятся в файле со значениями, разделенными запятыми (CSV):



Рис. 2. Исходные данные за 2018 г. в файле \*.csv

Попробуйте решить проблему самостоятельно. Потратьте пару часов, пытаясь это сделать. Если вам потребуется помощь, можно прочитать текст ниже.

Не знаете, как поступить? Вы можете разбить задачу на этапы:

1. Импортируйте все файлы Excel из папки.
2. Создайте новый столбец с 36 строками объектов *Table* (по таблице на лист в трех файлах).
3. Для каждого листа удалите первую строку, столбец и строку итогов.
4. Примените функцию *FnUnpivotSummarizedTable*.
5. Импортируйте файл 2018.csv и выполните сведение столбца.
6. Соедините доходы за четыре года. Обратите внимание, что даты представлены по-разному.

#### Часть 1. Подготовка таблиц

Начнем с преобразования сводных помесячных данных за 2015-2017 гг. в таблицу фактов. Откройте новую книгу Excel. Пройдите *Данные* –> *Получить данные* –> *Из файла* –> *Из папки*. Выберите путь для папки C:\Data\C14\C14E01 и щелкните *Открыть*. В следующем окне щелкните *Преобразовать данные*. Откроется окно редактора Power Query.

Выберите элемент управления фильтром столбца *Name* и снимите галочку с файла 2018.csv. Щелкните Ok. Переименуйте запрос из *C14E01* в *Revenues*. В заголовке столбца *Content* щелкните на значке *Объединить файлы*. В окне *Объединить файлы* выделите папку *Параметр 1 [12]* и кликните Ok. Power Query автоматически создаст ряд запросов:



Рис. 3. Импорт всех листов из всех Excel-файлов из папки

Убедитесь, что активен запрос *Revenues* и удалите последние три столбца. Удалите расширение *xlsx* из столбца *Source.Name* для извлечения года и переименуйте столбец в *Year*. (Имеется несколько способов удалить расширение файла. Если вы не уверены, как это сделать, вернитесь к [главе 3](https://baguzin.ru/wp/?p=24512).) Переименуйте столбец с *Name* на *Month*. В столбце *Data* находятся объекты *Table*. Если кликнуть левой кнопкой мыши на пустое пространство справа от слова Table, в нижней части окна увидим предварительный просмотр таблицы:



Рис. 4. Первая строка каждой таблицы не несет информации и помешает отмене сведения столбца

Следует удалить первую строку и общие итоги в последнем столбце и последней строке. Пройдите *Добавление столбца* –> *Настраиваемый столбец*. Настройте окно *Настраиваемый* *столбец*:



Рис. 5. Параметры окна *Настраиваемый столбец*

Нажмите Ok. Новый столбец возвращает объект *Table* такой же, как и в столбце *Data* но без первой строки. Можете проверить, кликнув левой кнопкой мыши на пустое пространство справа от слова *Table* в столбце *Skip First Row*. Удалите столбец *Data*. Пройдите *Добавление столбца* –> *Настраиваемый столбец*. В окне *Настраиваемый* *столбец* в поле *Имя нового столбца* введите *No Grand Total Column*. В поле *Формула пользовательского столбца* введите:

= Table.RemoveColumns(

[Skip First Row], {

List.Last(Table.ColumnNames([Skip First Row]))

}

)

Эта формула удаляет последний столбец. С помощью функции Table.ColumnNames вы получаете имена столбцов таблицы в столбце *Skip First Row*, а List.Last возвращает имя последнего столбца. Опять же, можете проверить, что из всех таблиц удалены столбцы *Column19*, содержавшие общие итоги.

Удалите столбец *Skip First Row*. Пройдите *Добавление столбца* –> *Настраиваемый столбец*. В окне *Настраиваемый* *столбец* в поле *Имя нового столбца* введите *Summarized Table*. В поле *Формула пользовательского столбца* введите:

= Table.RemoveLastN([No Grand Total Column], 1)

Проверить удаление последний строки не удастся, так как предварительный просмотр таблицы показывает лишь несколько первых строк. Удалите столбец *No Grand Total Column*.

Альтернативный способ удаления первой строки, столбца и строки итогов состоит в написании пользовательской функции. Находясь в редакторе Power Query пройдите *Главная* –> *Создать* *источник* –> *Другие источники* –> *Пустой запрос*. В строку формул введите код:

(inputTable) =>

let

SkipFirstRow = Table.Skip(inputTable, 1),

NoGrandTotalColumn = Table.RemoveColumns(

SkipFirstRow, {

List.Last(Table.ColumnNames(SkipFirstRow))

}

),

Result = Table.RemoveLastN(NoGrandTotalColumn, 1)

in

Result

Назовите запрос *FnCleanSummarizedTable.* Создайте дубль запроса *Revenues*. В запросе *Revenues (2)* удалите шаги от *Добавлен пользовательский объект* до последнего. Пройдите *Добавление столбца* –> *Вызвать настраиваемую функцию*:



Рис. 6. Удаление первой строки и итогов с помощью пользовательской функции

#### Часть 2. Вызов функции отмены сведения столбца

Для отмены сведения столбцов применим пользовательскую функцию FnUnpivotSummarizedTable, рассмотренную [ранее](https://baguzin.ru/wp/?p=24620). Пройдите *Главная* –> *Закрыть и загрузить в…* –> *Только создать подключение*. Сохраните файл под именем C14E01 - Result.xlsx. Загрузите файл C07E03 - Solution.xlsx с сайта [Microsoft Press](https://www.microsoftpressstore.com/store/collect-combine-and-transform-data-using-power-query-9781509307951#downloads) или файл C07E02 - Result.xlsx из [архива](https://baguzin.ru/wp/wp-content/uploads/2022/02/Sozdanie-funktsii-otmeny-svertyvaniya-stolbtsov.-Primery.7z). Откройте файл Excel. На панели *Запросы* *и подключения* щелкните правой кнопкой мыши на FnUnpivotSummarizedTable и выберите *Копировать*. Вернитесь в файл C14E01 - Result.xlsx. Откройте редактор PQ. Правой кнопкой мыши щелкните где-либо в фоновой области панели *Запросы* и выберите *Вставить*.

В качестве альтернативы, находясь в редакторе PQ файла C14E01 - Result.xlsx, пройдите *Главная* –> *Создать источник* –> *Другие источники* –> *Пустой запрос*. А затем *Главная* –> *Расширенный редактор*. Выделите и удалите всё содержимое в окне *Расширенный редактор*. Вставьте код:

let

 Источник = (Source, RowFields, ColumnFields, ValueField)=>

let

 #"Filled Down" = Table.FillDown(Source, List.FirstN(Table.ColumnNames(Source), List.Count(RowFields) - 1)),

 #"Merged Columns" = Table.CombineColumns(#"Filled Down", List.FirstN(Table.ColumnNames(#"Filled Down"), List.Count(RowFields)), Combiner.CombineTextByDelimiter(":", QuoteStyle.None),"Merged"),

 #"Transposed Table" = Table.Transpose(#"Merged Columns"),

 #"Filled Down1" = Table.FillDown(#"Transposed Table",List.FirstN(Table.ColumnNames(#"Transposed Table"), List.Count(ColumnFields)-1)),

 #"Promoted Headers" = Table.PromoteHeaders(#"Filled Down1", [PromoteAllScalars=true]),

 #"Unpivoted Other Columns" = Table.UnpivotOtherColumns(#"Promoted Headers",List.FirstN(Table.ColumnNames(#"Promoted Headers"), List.Count(ColumnFields)) , "Attribute", ValueField),

 #"Split Column by Delimiter" = Table.SplitColumn(#"Unpivoted Other Columns", "Attribute", Splitter.SplitTextByDelimiter(":", QuoteStyle.Csv), RowFields),

 #"Renamed Columns" = Table.RenameColumns(#"Split Column by Delimiter", List.Zip({List.FirstN(Table.ColumnNames(#"Split Column by Delimiter"), List.Count(ColumnFields)), ColumnFields})),

 #"Changed Type" = Table.TransformColumnTypes(#"Renamed Columns",{{ValueField, type number}})

in

 #"Changed Type"

in

 Источник

Нажмите *Готово*. Переименуйте *Запрос1* в *FnUnpivotSummarizedTable*.

Итак, вы тем или иным способом создали функцию *FnUnpivotSummarizedTable.* На панели *Запросы* выберите *Revenues*. Пройдите *Добавление столбца* –> *Настраиваемый столбец*. В окне *Настраиваемый столбец* введите имя столбца *Unpivoted* и код:

= FnUnpivotSummarizedTable([Summarized Table], {"Customer Category", "City"}, {"Color", "Supplier Category", "Supplier Name"}, "Revenue")

Удалите столбец *Summarized Table*. Разверните столбец *Unpivoted*. На панели *Развернуть* отмените флажок *Использовать имя исходного столбца в качестве префикса* и щелкните Ok. В результате в панели предварительного просмотра можно увидеть, что для всех рабочих листов операция отмены свертывания столбцов выполнена корректно.



Рис. 7. Таблица доходов за 2015-2017 гг.

Теперь можно обработать доходы за 2018.

#### Часть 3. Последовательность свертывания столбцов для доходов 2018 года

Займемся файлом 2018.csv. Этот файл содержит разделенные запятыми значения пар атрибут/значение, и можно предположить, что каждая многострочная запись о доходах начинается с атрибута *Date* (см. рис. 2).

Находясь в редакторе PQ пройдите *Главная* –> *Создать источник* –> *Файл* –> *Текстовый или CSV-файл*. Выберите файл 2018.csv и нажмите *Импорт*. В окне 2018.CSV нажмите Ok. Переименуйте запрос в *2018 Revenues*. Пройдите *Преобразование* –> *Использовать первую строку в качестве заголовка*. Пройдите *Добавление столбца* –> *Столбец индекса*. Пройдите *Добавление столбца* –> *Условный столбец*. Настройте окно:



Рис. 8 Окно *Добавление условного столбца*

Щелкните Ok. Выберите столбец *Row ID* и пройдите *Преобразование* –> *Заполнить* –> *Вниз*. Удалите столбец *Индекс*. Выберите столбец *Attribute* и пройдите *Преобразование* –> *Столбец сведения*. В окне Столбец сведения в поле *Столбец значений* выберите *Value*. Раскройте раздел *Расширенные параметры*. В поле *Функция агрегированного значения* выберите *Не агрегировать*. Щелкните Ok. Удалите столбец *Row ID*.

Теперь можно объединить результаты за 2015-2017 гг. и за 2018 г.

#### Часть 4. Объединение доходов

Теперь имеются два запроса, причем *Revenues* включает столбцы *Year* и *Month*, а *2018 Revenues* – столбец *Date*. Выберите запрос *Revenues*. Выделите первые два столбца, *Year* и *Month*, и пройдите *Преобразование* –> *Объединить столбцы*. В окне *Объединить столбцы* выберите разделитель пробел, введите *Date* в поле *Имя нового столбца*, щелкните Ok. Пройдите *Главная* –> *Добавить запросы* –> *Добавить запросы*. В окне *Добавление* в качестве второй выберите таблицу *2018 Revenues.* Щелкните Ok.

Измените тип столбца *Revenue* на *Десятичное число*, а столбца *Date* – на *Дата*. Загрузите запросы: *Revenues* – *Только создать подключение* и *Добавить эти данные в модель данных*, а запрос *2018 Revenues* – *Только создать подключение*. Файл решения C14E01 - Solution.xlsx.

#### Сравнение таблиц и отслеживание действий хакеров

В предыдущем упражнении вы восстановили отчет о доходах после кибератаки. Согласно исследованию группы по кибербезопасности, злоумышленники подделали значения доходов в хранилище данных. Ваша задача состоит в сравнении доходов в вашем отчете с измененными данными в файле C14E02 - Compromied.xlsx для определения, какие записи подделаны хакерами. Хотя найти различия можно с помощью сводных таблиц или мер DAX, попытайтесь решить задачу с помощью Power Query.

Вот несколько подсказок, которые помогут вам:

1. Хакеры внесли изменения в столбец *Revenue*. Можно объединить две таблицы для нахождения несоответствий. Используйте все столбцы, кроме столбца *Revenue*, для слияния, а затем разверните столбец *Revenue* из второй таблицы. Примените фильтр к двум столбцам *Revenue* для нахождения различий.
2. Чтобы найти новые строки, добавленные хакерами, можно объединить таблицу из вашего отчета с скомпрометированной таблицей. Тип соединения – анти-соединение справа. Таким образом, будут обнаружены строки, которые есть в скомпрометированной таблице и нет в вашем отчете.

Импортируйте таблицу *Revenues* из файла C14E01 - Solution.xlsx. Переименуйте запрос в *Correct*. Импортируйте таблицу *Revenues* из файла C14E02 - Compromised.xlsx. Переименуйте запрос в *Compromised*. На панели *Запросы* выберите *Correct*. Пройдите *Главная* –> *Объединить запросы* –> *Объединить запросы в новый*. В окне *Слияние* в таблице *Correct*, удерживая нажатой клавишу <Ctrl>, выбирайте столбцы Date, Color, Supplier Category, Supplier Name, Custom Category и City. Обязательно сохраняйте порядок следования столбцов. В качестве второй выберите таблицу *Compromised*. В таблице *Compromised*, удерживая нажатой клавишу <Ctrl>, выбирайте поля Date, Color, Supplier Category, Supplier Name, Custom Category и City.

В поле *Тип соединения* выберите *Внутреннее (только совпадающие строки).* Щелкните Ok. Переименуйте новый запрос в *Compromised Rows* и разверните столбец *Compromised*. На панели *Развернуть* отмените выбор всех столбцов, кроме *Revenue*. Снимите галочку *Использовать* *исходное имя столбца как префикс*. Переименуйте новый столбец в *Compromised.Revenue*.

Чтобы отфильтровать строки, в которых доходы в столбцах *Revenue* и *Compromised.Revenue* не равны необходимо применить фильтр, который недоступен в пользовательском интерфейсе. Можно сгенерить код фильтра через пользовательский интерфейс, а затем изменить его в строке формул. Щелкните мышью на элементе фильтра в заголовке столбца *Revenue* и выберите *Числовые фильтры* –> *Не равно* –> 12345. В строке формул увидим:

= Table.SelectRows(#"Переименованные столбцы", each [Revenue] <> 12345)

Измените формулу на

= Table.SelectRows(#"Переименованные столбцы", each [Revenue] <> [Compromised.Revenue])

Остались только строки с различными значениями в *Revenue* и *Compromised.Revenue*.



Рис. 9. Строки с различными значениями в *Revenue* и *Compromised.Revenue*

#### Обнаружение следов деятельности хакеров во взломанной таблице

Теперь найдем строки в скомпрометированной таблице, которые были добавлены хакерами и не найдены в исходной таблице. Продолжите предыдущее упражнение. На панели *Запросы* выберите *Correct*. Пройдите *Главная* –> *Объединить запросы* –> *Объединить запросы в новый*. В окне *Слияние* в таблице *Correct*, удерживая нажатой клавишу <Ctrl>, выбирайте столбцы Date, Color, Supplier Category, Supplier Name, Custom Category и City. Обязательно сохраняйте порядок следования столбцов. В качестве второй выберите таблицу *Compromised*. В ней удерживая нажатой клавишу <Ctrl>, выбирайте столбцы Date, Color, Supplier Category, Supplier Name, Custom Category и City.

В поле *Тип соединения* выберите *Анти-соединение справа (только строки во второй таблице)*. Щелкните Ok. Переименуйте запрос в *Hacker's Instructions*. Удалите все столбцы кроме *Compromised*. Нажмите значок развернуть рядом с заголовком *Compromised.* Извлеките все поля, но снимите галочку *Использовать* *исходное имя столбца как префикс*. Найдите сообщение хакера.



Рис. 10. Сообщение хакера

Загрузите запрос *Hacker's Instructions* в таблицу на лист Excel. Файл решения C14E02 - Solution.xlsx.