**Глава 7. Импорт больших текстовых файлов в Power Query**

Это продолжение перевода книги Кен Пульс и Мигель Эскобар. Язык М для Power Query. Главы не являются независимыми, поэтому рекомендую читать последовательно.

[Предыдущая глава](http://baguzin.ru/wp/?p=21013) [Содержание](http://baguzin.ru/wp/?p=20852#content) Следующая глава

Одной из самых больших проблем для профессионалов Excel является импорт и очистка неструктурированных текстовых файлов. В них зачастую:

* отсутствуют символы-разделители,
* в разных строках поля разделены различным количеством пробелов,
* присутствуют непечатаемые символы,
* повторяются строки заголовка.

Power Query автоматически решает большинство из этих проблем. Создайте новую книгу Excel, и новый запрос: *Данные* –> *Из текстового/CSV-файла*. Загрузите файл с примерами *GL Jan-Mar.txt*. В окне предварительного просмотра нажмите *Изменить*. Power Query помещает данные в один столбец:

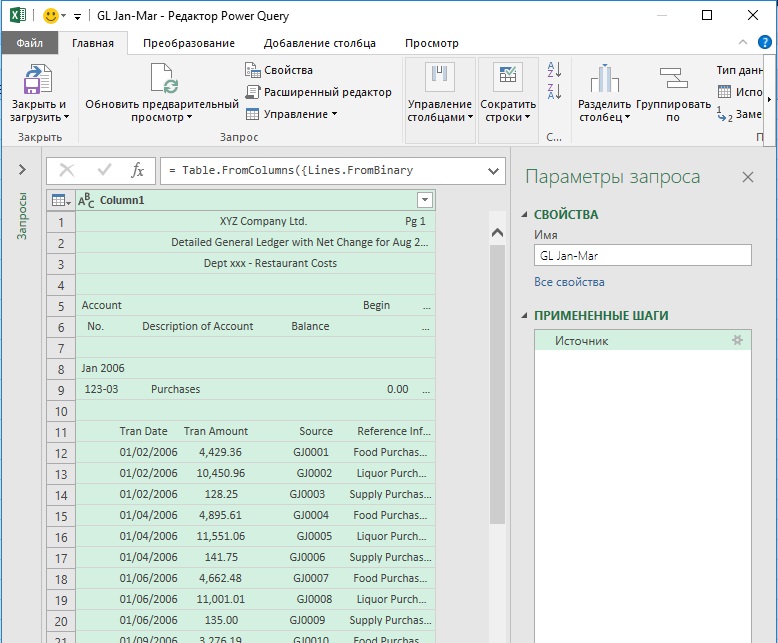


Рис. 7.1. Импорт неструктурированного текстового файла

Поскольку файл не содержит разделителей, Power Query не сделал никаких предположений о данных. Он предоставил вам возможность обработать данные вручную. На первом этапе основная цель – как можно быстрее представить данные в виде подобия столбцов. Верхние 10 строк, похоже, лишние, а вот 11-я строка напоминает заголовки. *Главная* –> *Удалить строки* –> *Удаление верхних строк* –> 10.

Далее следует избавиться от пробелов. В Excel это стандартная практика при обработке текста. Например, функция СЖПРОБЕЛЫ() удаляет все начальные, конечные и повторяющиеся пробелы, а ПЕЧСИМВ() – непечатные символы. Power Query также умеет это делать. Щелкните правой кнопкой мыши *Column1* –> *Преобразование* –> *Очистить*. А затем *Column1* –> *Преобразование* –> *Усечь*. Усечение Power Query работает немного иначе функции СЖПРОБЕЛЫ() в Excel. Усечение удаляет только начальные и конечные пробелы, оставляя внутренние пробелы без изменений.

Следующий шаг – разделение столбцов. Поскольку дата содержит 10 символов, можно попробовать оставить чуть больше, например, 12 для первого столбца. *Главная* –> *Разделить столбец* –> *По количеству символов* –> 12. Обратите внимание на настройку переключателя *Разделение*:

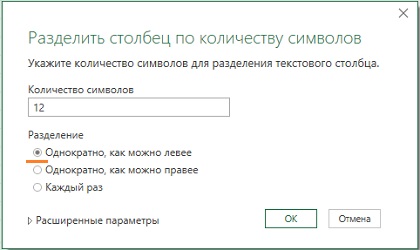


Рис. 7.2. Выделение первого столбца

Повторно усеките два новых столбца:

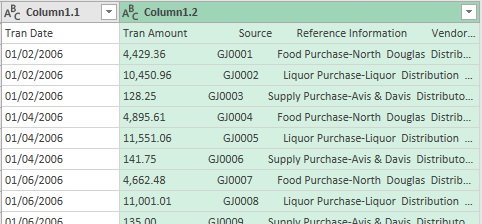


Рис. 7.3. Промежуточный вид запроса

Подберите количество символов для выделения столбца *Tran Amount*, усеките столбцы после разделения, и т.д. У вас получится приблизительно так:

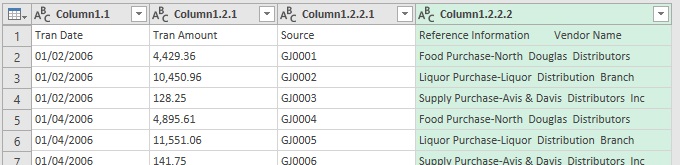


Рис. 7.4. Выделено четыре столбца

Для разделения последнего столбца используйте иной трюк: *Главная* –> *Разделить столбец* –> *По разделителю* –> ˽˽ (два пробела). Усеките два образовавшихся столбца. Обратите внимание, каждый раз, когда вы разделяли столбцы, Power Query автоматически добавлял еще один шаг – изменял тип нового столбца на текст. Это лишнее действие, поэтому удалите все шаги *Измененный тип*. Далее *Главная* –> *Использовать первую строку в качестве заголовков*. С этапом разбиения месива данных на столбцы вы справились:

Изображение выглядит как снимок экрана

Автоматически созданное описание

Рис. 7.5. Неструктурированные данные разбиты на столбцы

Если вы прокрутите вниз, то обнаружите, что в данных много строк мусора:

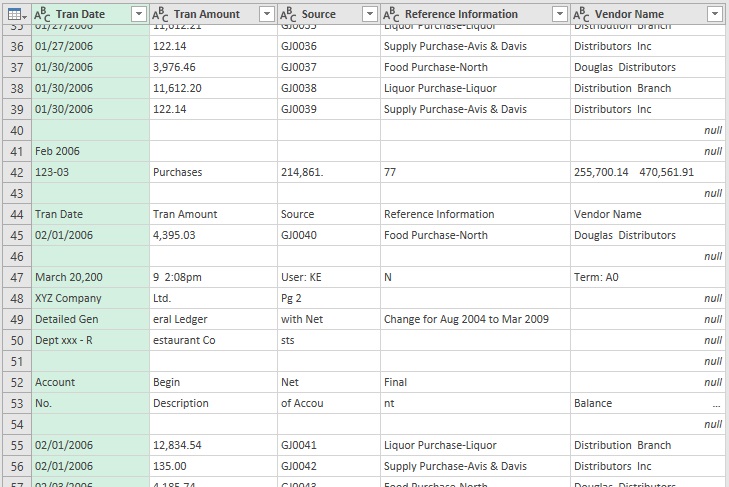


Рис. 7.6. Строки заголовка второй страницы, смешанные с данными

Щелкните правой кнопкой мыши столбец *Tran Date* –> *Тип изменения* –> *Используя локаль* –> *Дата* –> *Английский (США)*. Подробнее см. [Глава 2. Изменение настроек Power Query, действующих по умолчанию](http://baguzin.ru/wp/?p=20889). Появится куча ошибок:

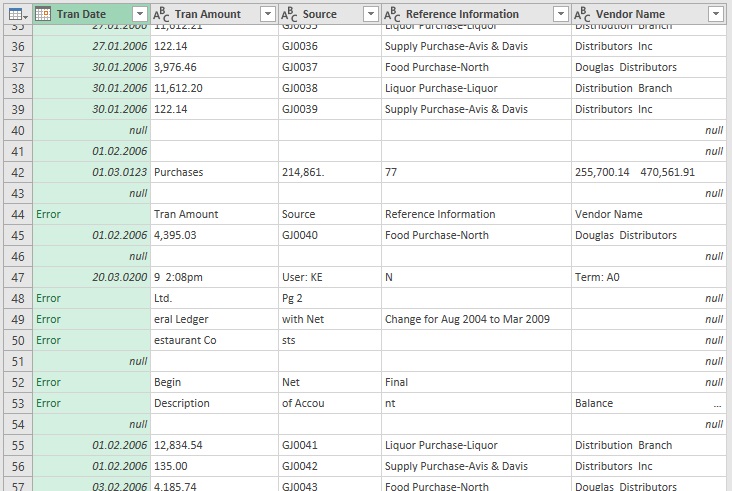


Рис. 7.7. Часть ячеек в столбце *Tran Date* содержат ошибки или значение *null*

В отличие от многих других программ, ошибки в Power Query очень функциональны. Их можно контролировать, на них можно реагировать. Изучив ошибки на рис. 7.7, вы увидите, что они появились только в строках, которые являются мусором. Значение null также сигнализируют, что эти строки не нужны. Выделите столбец *Tran Date* –> *Главная* –> *Удалить ошибки.* Отфильтруйте столбец *Tran Date* –> снимите флажок null.

На данный момент всё еще осталось несколько нерелевантных строк, чтобы найти их, сначала отсортируйте *Tran Date* по возрастанию (рис. 7.8), а затем по убыванию, и снимите флажки с ненужных строк.

Изображение выглядит как снимок экрана

Автоматически созданное описание

Рис. 7.8. Несколько нерелевантных строк

Щелкните правой кнопкой мыши столбец *Tran Date* –> *Переименовать* –> *Date.* Щелкните правой кнопкой мыши столбец сумма *Tran Amount* –> *Переименовать* –> *Amount.* Щелкните правой кнопкой мыши столбец *Amount* –> *Тип изменения* –> *Используя локаль* –> *Десятичное число* –> *Английский (США)*. Отфильтруйте столбец *Amount* –> снимите флажок null. Все строки мусора удалены.

Присмотревшись к данным в двух последних столбцах, вы понимаете, что разбиение на столбцы было выполнено неверно:

Изображение выглядит как снимок экрана

Автоматически созданное описание

Рис. 7.9. Два последних столбца разделены неверно

Объедините столбцы: выделите столбец *Reference Information*, удерживая нажатой клавишу Ctrl, выделите столбец *Vendor Name*. Важно помнить, что порядок выделения столбцов определяет, какой из них будет первым при объединении. Пройдите по меню *Преобразование* –> *Объединить столбцы* –> *Разделитель* –> *Пробел*. Снова разделите столбцы, используя разделитель дефис. Дайте разделенным столбцам разумные названия: *Category* и *Vendor*.

Возможно в названии некоторых поставщиков (Vendor), используется дефис, поэтому в окне *Разделить столбец по разделителю* установите переключатель в положение *Самый левый разделитель*:

Изображение выглядит как снимок экрана

Автоматически созданное описание

Рис. 7.10. Настройка разделения столбца

Помните, что при настройке этой операции, вы не ограничены разделителями, состоящими из одного символа. Если требуется, вы можете ввести в это поле хоть целое слово.

Возможно, в столбце *Vendor* остались сдвоенные пробелы. Вы не можете использовать усечение, так как оно не работает между словами в столбце. Щелкните правой кнопкой мыши *Vendor* –> *Замена значений* –> *Значение для поиска* – два пробела подряд –> *Заменить на* – один пробел –> *Ok*. Если вы подозреваете, что два пробела подряд могут встречаться несколько раз, повторите шаг с заменой. Дайте запросу говорящее имя, например, *Transactions*.

Цель достигнута – у вас чистый набор данных (рис. 7.11), который можно загрузить в таблицу: *Главная* –> *Закрыть и загрузить*.

Изображение выглядит как снимок экрана

Автоматически созданное описание

Рис. 7.11. Очищенные данные

Данные загрузятся на лист Excel и будут отформатированы как Таблица:

Изображение выглядит как снимок экрана

Автоматически созданное описание

Рис. 7.12. Данные загружены в Таблицу

Чтобы проверить качество данных, щелкните в любой ячейке Таблицы, *Вставить* –> Сводная таблица –> На существующий лист –> Диапазон G2. Настройте сводную таблицу:

Изображение выглядит как снимок экрана

Автоматически созданное описание

Рис. 7.13. Сводная таблица для проверки качества данных

О качестве данных говорят следующие признаки:

* Даты в строках, сгруппированы по месяцам (если бы хотя бы одно значение в столбце не было датой, группировка не выполнилась)
* Три поставщика в строках (а не их разнобой)
* Три категории по столбцам
* Суммы в ячейках (если бы хотя бы одно значение не было числом, здесь отражалось бы количество ячеек)

Всё, что было сделано до сих пор, было возможно выполнить и без Power Query. Преимущество описанной методики вы ощутите, когда через квартал к вам поступит новый текстовый файл. В мире Excel это означает еще один утомительный час импорта, очистки и форматирования. В мире Power Query вам нужно выполнить всего лишь несколько простых операций:

* Щелкните правой кнопкой мыши запрос *Transactions* –> *Изменить*
* Перейдите к первому шагу и щелкните значок шестеренки
* Измените имя файла GL Apr-Jun.txt
* *Главная* –> *Закрыть и загрузить*

Изображение выглядит как снимок экрана

Автоматически созданное описание

Рис. 7.14. Сводная таблица на данных второго квартала

Появились новые поставщики, новые транзакции и новые даты. Если вам нужны данные, как за первый, так и за второй кварталы, воспользуйтесь импортом всех файлов из папки, как описано в [главе 4](http://baguzin.ru/wp/?p=20923).