Глава 7. Импорт больших текстовых файлов в Power Query

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

Предыдущая глава Содержание 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:

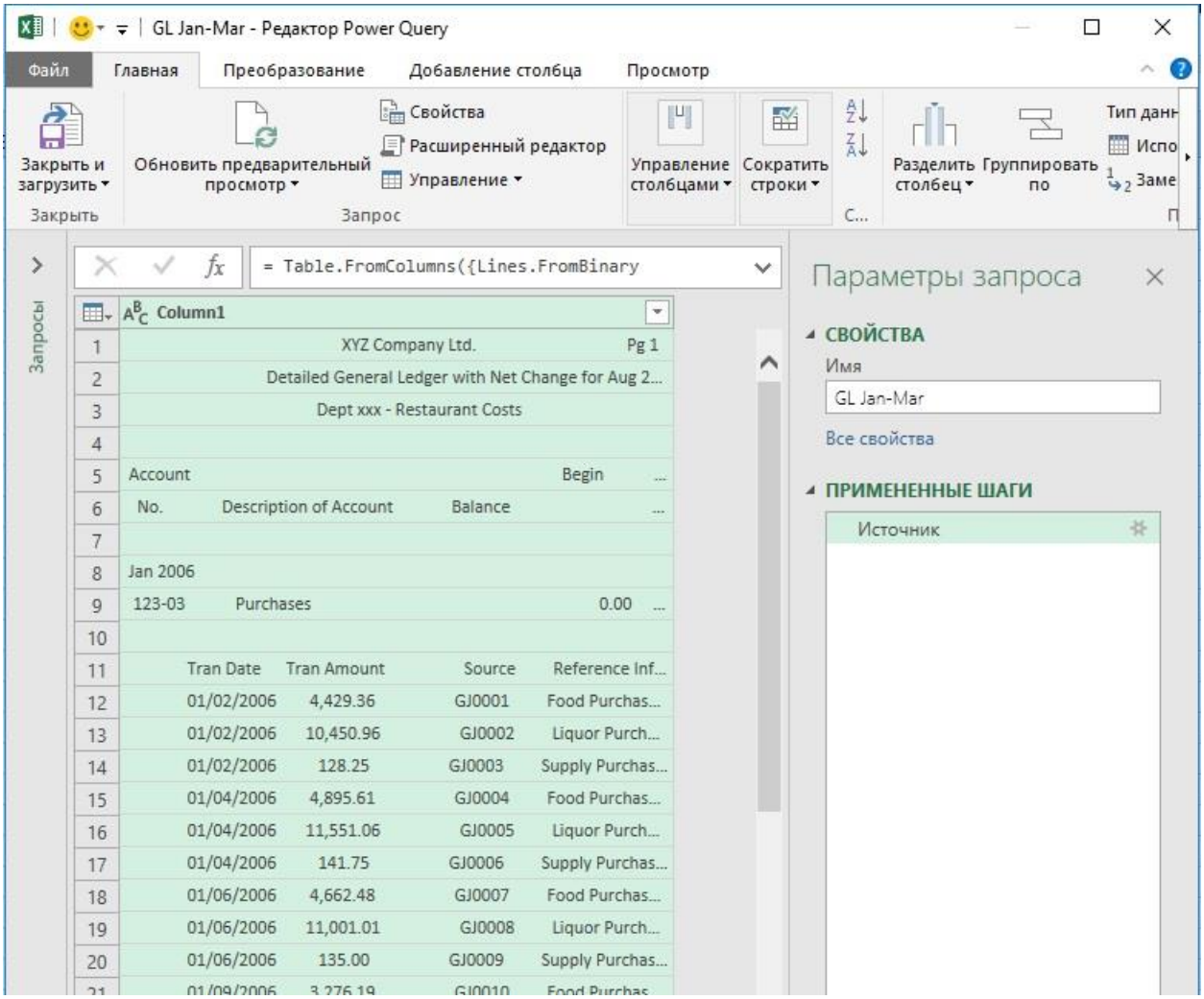


Рис. 7.1. Импорт неструктурированного текстового файла

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*:

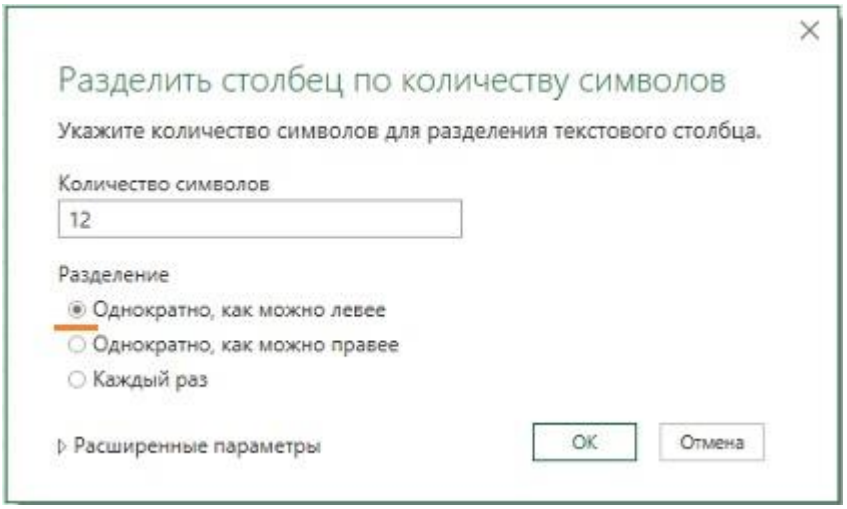


Рис. 7.2. Выделение первого столбца

Повторно усеките два новых столбца:

Column1.1	Column1.2			
Tran Date	Tran Amount	Source	Reference Information	Vendor...
01/02/2006	4,429.36	GJ0001	Food Purchase-North Douglas Distrib...	
01/02/2006	10,450.96	GJ0002	Liquor Purchase-Liquor Distribution ...	
01/02/2006	128.25	GJ0003	Supply Purchase-Avis & Davis Distributo...	
01/04/2006	4,895.61	GJ0004	Food Purchase-North Douglas Distrib...	
01/04/2006	11,551.06	GJ0005	Liquor Purchase-Liquor Distribution ...	
01/04/2006	141.75	GJ0006	Supply Purchase-Avis & Davis Distributo...	
01/06/2006	4,662.48	GJ0007	Food Purchase-North Douglas Distrib...	
01/06/2006	11,001.01	GJ0008	Liquor Purchase-Liquor Distribution ...	
01/06/2006	135.00	GJ0009	Supply Purchase-Avis & Davis Distributo	

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:

	Column1.1	Column1.2.1	Column1.2.2.1	Column1.2.2.2
1	Tran Date	Tran Amount	Source	Reference Information Vendor Name
2	01/02/2006	4,429.36	GJ0001	Food Purchase-North Douglas Distributors
3	01/02/2006	10,450.96	GJ0002	Liquor Purchase-Liquor Distribution Branch
4	01/02/2006	128.25	GJ0003	Supply Purchase-Avis & Davis Distributors Inc
5	01/04/2006	4,895.61	GJ0004	Food Purchase-North Douglas Distributors
6	01/04/2006	11,551.06	GJ0005	Liquor Purchase-Liquor Distribution Branch
7	01/04/2006	141.75	GJ0006	Supply Purchase-Avis & Davis Distributors Inc

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:

	Tran Date	Tran Amount	Source	Reference Information	Vendor Name
1	01/02/2006	4,429.36	GJ0001	Food Purchase-North	Douglas Distributors
2	01/02/2006	10,450.96	GJ0002	Liquor Purchase-Liquor	Distribution Branch
3	01/02/2006	128.25	GJ0003	Supply Purchase-Avis & Davis	Distributors Inc
4	01/04/2006	4,895.61	GJ0004	Food Purchase-North	Douglas Distributors
5	01/04/2006	11,551.06	GJ0005	Liquor Purchase-Liquor	Distribution Branch
6	01/04/2006	141.75	GJ0006	Supply Purchase-Avis & Davis	Distributors Inc
7	01/06/2006	4,662.48	GJ0007	Food Purchase-North	Douglas Distributors

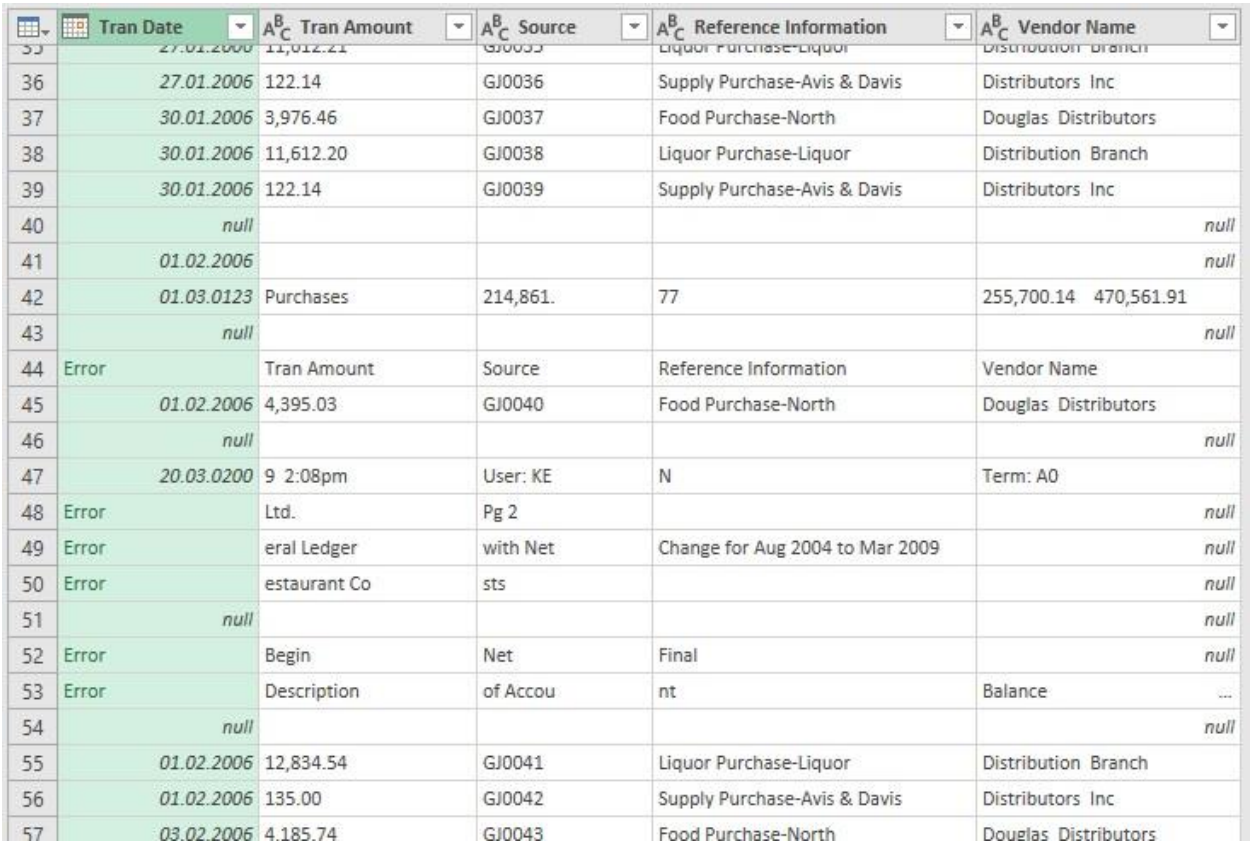
Рис. 7.5. Неструктурированные данные разбиты на столбцы

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

	Tran Date	Tran Amount	Source	Reference Information	Vendor Name
36	01/27/2006	122.14	GJ0036	Supply Purchase-Avis & Davis	Distributors Inc
37	01/30/2006	3,976.46	GJ0037	Food Purchase-North	Douglas Distributors
38	01/30/2006	11,612.20	GJ0038	Liquor Purchase-Liquor	Distribution Branch
39	01/30/2006	122.14	GJ0039	Supply Purchase-Avis & Davis	Distributors Inc
40					null
41	Feb 2006				null
42	123-03	Purchases	214,861.	77	255,700.14 470,561.91
43					null
44	Tran Date	Tran Amount	Source	Reference Information	Vendor Name
45	02/01/2006	4,395.03	GJ0040	Food Purchase-North	Douglas Distributors
46					null
47	March 20,200	9 2:08pm	User: KE	N	Term: A0
48	XYZ Company	Ltd.	Pg 2		null
49	Detailed Gen	eral Ledger	with Net	Change for Aug 2004 to Mar 2009	null
50	Dept xxx - R	estaurant Co	sts		null
51					null
52	Account	Begin	Net	Final	null
53	No.	Description	of Accou	nt	Balance ...
54					null
55	02/01/2006	12,834.54	GJ0041	Liquor Purchase-Liquor	Distribution Branch
56	02/01/2006	135.00	GJ0042	Supply Purchase-Avis & Davis	Distributors Inc

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. Появится куча ошибок:

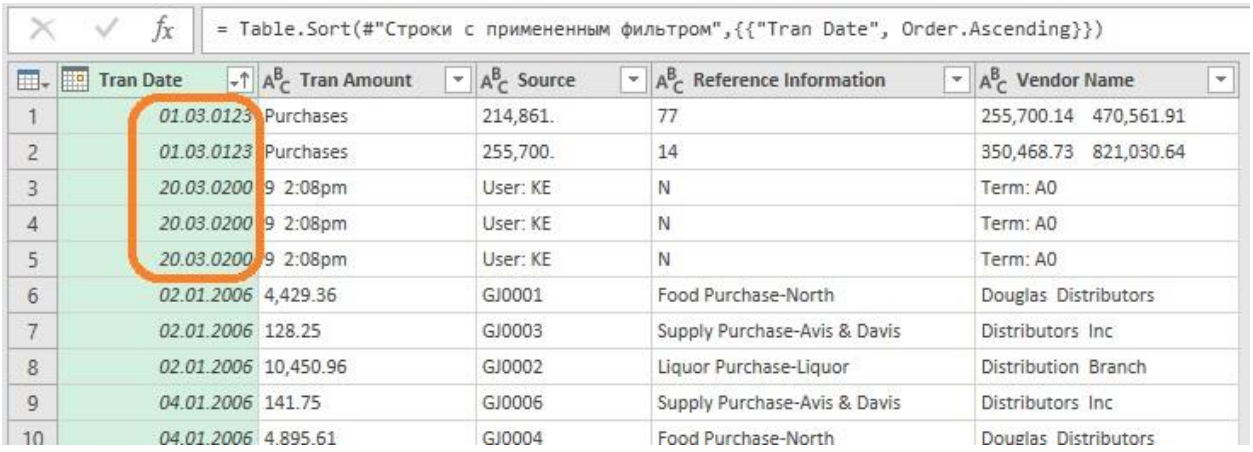


	Tran Date	Tran Amount	Source	Reference Information	Vendor Name
36	27.01.2006	122.14	GJ0036	Supply Purchase-Avis & Davis	Distributors Inc
37	30.01.2006	3,976.46	GJ0037	Food Purchase-North	Douglas Distributors
38	30.01.2006	11,612.20	GJ0038	Liquor Purchase-Liquor	Distribution Branch
39	30.01.2006	122.14	GJ0039	Supply Purchase-Avis & Davis	Distributors Inc
40	null				null
41	01.02.2006				null
42	01.03.0123	Purchases	214,861.	77	255,700.14 470,561.91
43	null				null
44	Error	Tran Amount	Source	Reference Information	Vendor Name
45	01.02.2006	4,395.03	GJ0040	Food Purchase-North	Douglas Distributors
46	null				null
47	20.03.0200	9 2:08pm	User: KE	N	Term: A0
48	Error	Ltd.	Pg 2		null
49	Error	eral Ledger	with Net	Change for Aug 2004 to Mar 2009	null
50	Error	estaurant Co	sts		null
51	null				null
52	Error	Begin	Net	Final	null
53	Error	Description	of Accou	nt	Balance ...
54	null				null
55	01.02.2006	12,834.54	GJ0041	Liquor Purchase-Liquor	Distribution Branch
56	01.02.2006	135.00	GJ0042	Supply Purchase-Avis & Davis	Distributors Inc
57	03.02.2006	4.185.74	GJ0043	Food Purchase-North	Douglas Distributors

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.



	Tran Date	Tran Amount	Source	Reference Information	Vendor Name
1	01.03.0123	Purchases	214,861.	77	255,700.14 470,561.91
2	01.03.0123	Purchases	255,700.	14	350,468.73 821,030.64
3	20.03.0200	9 2:08pm	User: KE	N	Term: A0
4	20.03.0200	9 2:08pm	User: KE	N	Term: A0
5	20.03.0200	9 2:08pm	User: KE	N	Term: A0
6	02.01.2006	4,429.36	GJ0001	Food Purchase-North	Douglas Distributors
7	02.01.2006	128.25	GJ0003	Supply Purchase-Avis & Davis	Distributors Inc
8	02.01.2006	10,450.96	GJ0002	Liquor Purchase-Liquor	Distribution Branch
9	04.01.2006	141.75	GJ0006	Supply Purchase-Avis & Davis	Distributors Inc
10	04.01.2006	4.895.61	GJ0004	Food Purchase-North	Douglas Distributors

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:

Reference Information	Vendor Name
Supply Purchase-Avis & Davis	Distributors Inc
Liquor Purchase-Liquor	Distribution Branch
Food Purchase-North	Douglas Distributors
Supply Purchase-Avis & Davis	Distributors Inc
Liquor Purchase-Liquor	Distribution Branch
Food Purchase-North	Douglas Distributors
Supply Purchase-Avis & Davis	Distributors Inc

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*:

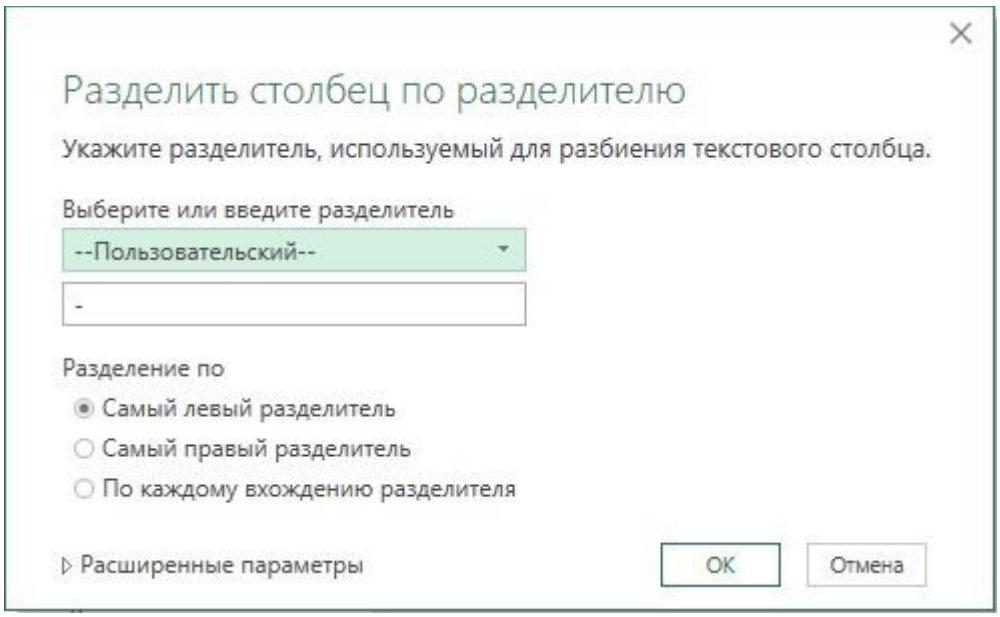


Рис. 7.10. Настройка разделения столбца

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.

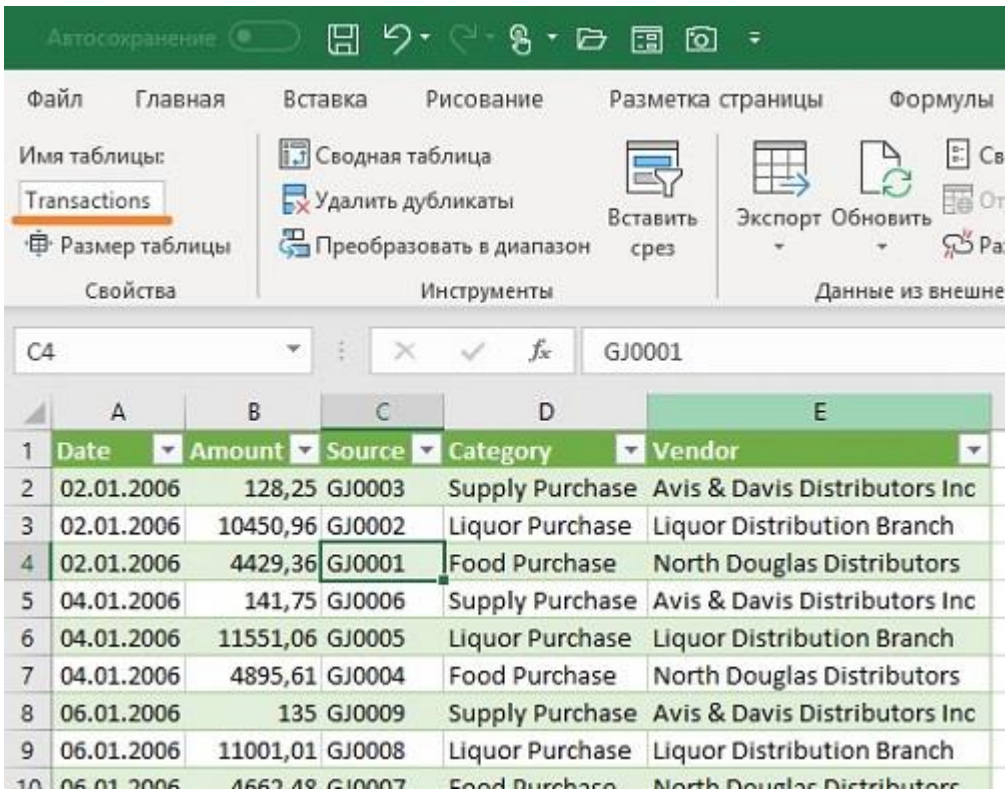
Возможно, в столбце *Vendor* остались сдвоенные пробелы. Вы не можете использовать усечение, так как оно не работает между словами в столбце. Щелкните правой кнопкой мыши *Vendor –> Замена значений –> Значение для поиска* – два пробела подряд *–> Заменить на* – один пробел *–> Ok*. Если вы подозреваете, 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*.

Date	Amount	Source	Category	Vendor
02.01.2006	128,25	GJ0003	Supply Purchase	Avis & Davis Distributors Inc
02.01.2006	10450,96	GJ0002	Liquor Purchase	Liquor Distribution Branch
02.01.2006	4429,36	GJ0001	Food Purchase	North Douglas Distributors
04.01.2006	141,75	GJ0006	Supply Purchase	Avis & Davis Distributors Inc
04.01.2006	11551,06	GJ0005	Liquor Purchase	Liquor Distribution Branch
04.01.2006	4895,61	GJ0004	Food Purchase	North Douglas Distributors

Рис. 7.11. Очищенные данные

Данные загрузятся на лист Excel и будут отформатированы как Таблица:



	A	B	C	D	E
1	Date	Amount	Source	Category	Vendor
2	02.01.2006	128,25	GJ0003	Supply Purchase	Avis & Davis Distributors Inc
3	02.01.2006	10450,96	GJ0002	Liquor Purchase	Liquor Distribution Branch
4	02.01.2006	4429,36	GJ0001	Food Purchase	North Douglas Distributors
5	04.01.2006	141,75	GJ0006	Supply Purchase	Avis & Davis Distributors Inc
6	04.01.2006	11551,06	GJ0005	Liquor Purchase	Liquor Distribution Branch
7	04.01.2006	4895,61	GJ0004	Food Purchase	North Douglas Distributors
8	06.01.2006	135	GJ0009	Supply Purchase	Avis & Davis Distributors Inc
9	06.01.2006	11001,01	GJ0008	Liquor Purchase	Liquor Distribution Branch

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:

Сумма по полю Amount	Food Purchase	Liquor Purchase	Supply Purchase	Общий итог
⊟янв				
Avis & Davis Distributors Inc			1 665	1 665
Liquor Distribution Branch		158 293		158 293
North Douglas Distributors	54 904			54 904
⊟фев				
Avis & Davis Distributors Inc			1 849	1 849
Liquor Distribution Branch		186 132		186 132
North Douglas Distributors	67 719			67 719
⊟мар				
Avis & Davis Distributors Inc			3 384	3 384
Liquor Distribution Branch		242 316		242 316
North Douglas Distributors	104 769			104 769
Общий итог	**227 393**	**586 741**	**6 897**	**821 031**

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*

Сумма по полю Amount	Food Purchase	Liquor Purchase	Supply Purchase	Общий итог
⊟апр				
Avis & Davis Distributors Inc			1 661	1 661
Liquor Distribution Branch		191 993		191 993
North Douglas Distributors	55 196			55 196
⊟май				
Avis & Davis Distributors Inc			261	261
Liquor Distribution Branch		177 125		177 125
North Douglas Distributors	62 668			62 668
ACME&Co Supply Haus LLC			1 580	1 580
Sysco	5 848			5 848
⊟июн				
Liquor Distribution Branch		226 608		226 608
ACME&Co Supply Haus LLC			3 376	3 376
Sysco	102 759			102 759
Общий итог	**226 472**	**595 725**	**6 879**	**829 075**

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.