

Tutorial: Loading Data into a Star Schema

Product attributes					Geography Data				Date	Measure
A	B	C	D	E	F	G	H	I	J	K
ProductID	Product	Category	Segment	Manufacturer	Zip	City	State	Region	Date	Amount
791	Natura RP-	Rural	Productivity	Natura	18336	Matamoras, PA	PA	East	9/30/2008	566.57433
759	Natura RP-	Rural	Productivity	Natura	18337	Milford, PA	PA	East	5/24/2004	1217.5678
2345	Aliqui UE-1	Urban	Extreme	Aliqui	18337	Milford, PA	PA	East	5/21/2009	3569.5443
609	Maximus U	Urban	Convenience	Maximus	18403	Archbald, PA	PA	East	1/23/2008	7179.8282
2045	Currus UE-	Urban	Extreme	Currus	18407	Carbondale, PA	PA	East	12/14/2004	3359.5464
1109	Pirum RP-5	Rural	Productivity	Pirum	18411	Clarks Summit, PA	PA	East	3/25/2002	1679.5632
792	Natura RP-	Rural	Productivity	Natura	18413	Clifford, PA	PA	East	3/24/2004	1510.7249
992	Natura UC-	Urban	Convenience	Natura	18419	Factoryville, PA	PA	East	7/24/2008	2435.5556

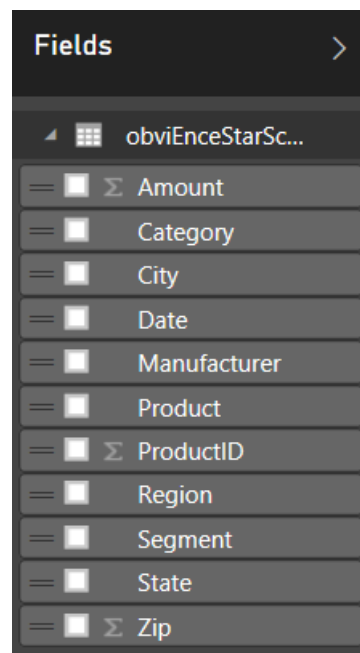
Very often our analysis starts with a flat data set that contains all of the pertinent columns in a single table that looks like the one above.

As we can see, we can analyze this data using three different lenses or dimensions:

1. Product
2. Geography
3. Date

Unfortunately, if we load the data as is, the end result does not look particularly user friendly as it essentially presents us with a flat list of all fields available for Analysis.

In this tutorial, we will learn how to massage the data using Power Query so that we can build a more user-friendly model.



- 1 The first thing to do is to load our data set into a Power Query environment so we can start massaging it. In our case, the dataset is contained in a CSV file. Click on Get Data -> Text/CSV and open obviEnceStarSchemaLab.csv and then click *Transform Data*

h1>obviEnceStarSchemaLab.csv

File Origin
1252: Western European (Windows)
Delimiter
Comma
Data Type Detection
Based on first 200 rows

ProductID	Product	Category	Segment	Manufacturer	Zip	City	State	Region	Date	Amount
791	Natura RP-79	Rural	Productivity	Natura	18336	Matamoras, PA	PA	East	9/30/2008	566.5743342
759	Natura RP-47	Rural	Productivity	Natura	18337	Milford, PA	PA	East	5/24/2004	1217.567824
2345	Aliqui UE-19	Urban	Extreme	Aliqui	18337	Milford, PA	PA	East	5/21/2009	3569.544304
609	Maximus UC-74	Urban	Convenience	Maximus	18403	Archbald, PA	PA	East	1/23/2008	7179.828201
2045	Currus UE-05	Urban	Extreme	Currus	18407	Carbondale, PA	PA	East	12/14/2004	3359.546404
1109	Pirum RP-55	Rural	Productivity	Pirum	18411	Clarks Summit, PA	PA	East	3/25/2002	1679.563204
792	Natura RP-80	Rural	Productivity	Natura	18413	Clifford, PA	PA	East	3/24/2004	1510.724893
992	Natura UC-55	Urban	Convenience	Natura	18419	Factoryville, PA	PA	East	7/24/2008	2435.555644
609	Maximus UC-74	Urban	Convenience	Maximus	18426	Greentown, PA	PA	East	8/31/2007	7179.828201
674	Maximus UC-39	Urban	Convenience	Maximus	18426	Greentown, PA	PA	East	11/28/2007	5709.842901
952	Natura UC-15	Urban	Convenience	Natura	18337	Milford, PA	PA	East	5/17/2001	2099.559004
533	Maximus UE-21	Urban	Extreme	Maximus	18407	Carbondale, PA	PA	East	5/17/2003	4197.858021
650	Maximus UC-15	Urban	Convenience	Maximus	18407	Carbondale, PA	PA	East	3/15/2006	4122.258777
2388	Aliqui UC-36	Urban	Convenience	Aliqui	18407	Carbondale, PA	PA	East	8/10/2000	2519.554804
1203	Pirum UC-05	Urban	Convenience	Pirum	18411	Clarks Summit, PA	PA	East	4/15/2002	2561.554384
599	Maximus UC-64	Urban	Convenience	Maximus	18414	Dalton, PA	PA	East	7/25/2007	7095.829041
2388	Aliqui UC-36	Urban	Convenience	Aliqui	18424	Gouldsboro, PA	PA	East	10/8/2007	2603.553964
633	Maximus UC-98	Urban	Convenience	Maximus	18426	Greentown, PA	PA	East	6/30/2009	4535.534644
2388	Aliqui UC-36	Urban	Convenience	Aliqui	18426	Greentown, PA	PA	East	4/10/2006	2519.554804
573	Maximus UC-38	Urban	Convenience	Maximus	18428	Hawley, PA	PA	East	10/9/2000	2349.876501

The data in the preview has been truncated due to size limits.

Load
Transform Data
Cancel

2 Now, let us make sure that all of our columns have a correct data type associated with them:

- Set the Date field to be Date
- Set the Amount field to be a Decimal Number
- Set Product ID to be a Whole Number

(most likely the field's type will be already correctly identified by tool)

Name the Query **SalesData** and also **disable** Load to Report by right clicking the query->Properties and unchecking *Enable load to report*

Query Properties

Name

SalesData

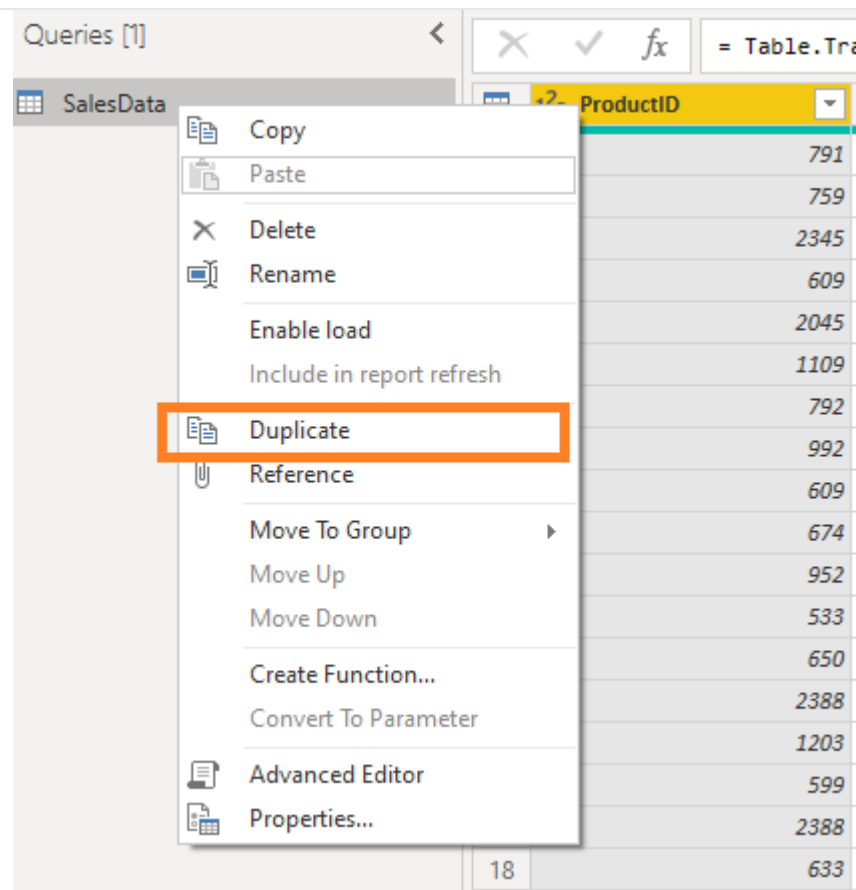
Description

☐ Enable load to report

☐ Include in report refresh ⓘ

OK Cancel

- 3 Now let us create a **Product** dimension by right clicking on the **SalesData** query and selecting *Duplicate*



- 4 Rename the new query **Product** from **SalesDta(2)**

Query Settings
×

PROPERTIES

Name

Product

All Properties

APPLIED STEPS

Source
⚙

Promoted Headers
⚙

✕ Changed Type

- 5 Now let us click on *Home->Choose Columns* and uncheck those entries that do not have anything to do with **Product** and click OK

Choose Columns

Choose the columns to keep

☐ (Select All Columns)

☒ ProductID

☒ Product

☒ Category

☒ Segment

☒ Manufacturer

☐ Zip

☐ City

☐ State

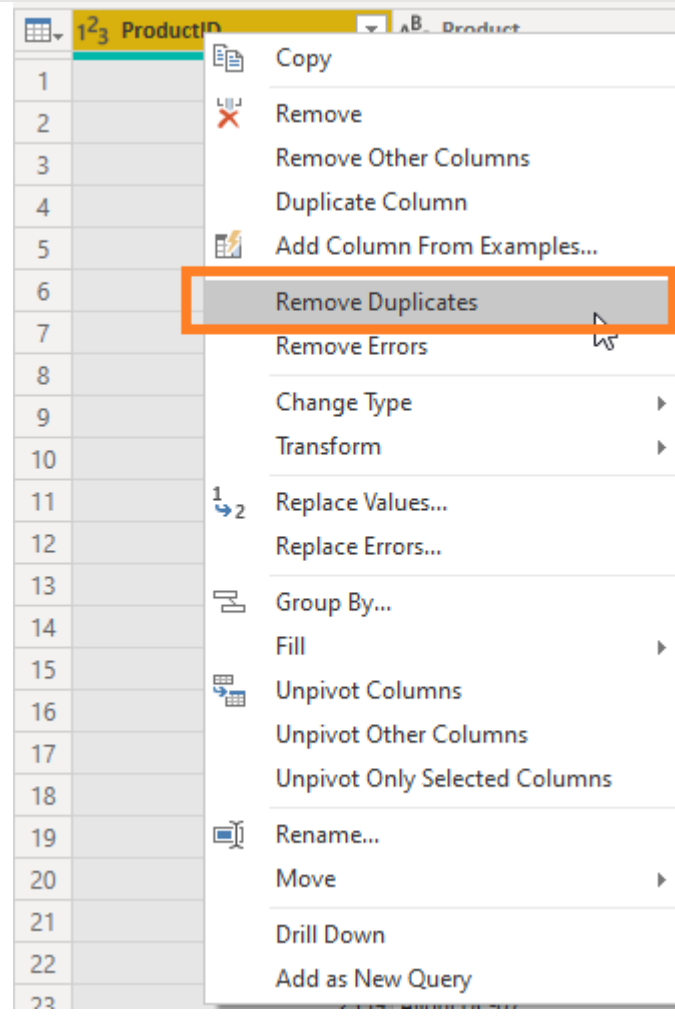
☐ Region

☐ Date

☐ Amount

OK

Cancel

6	<p>The next step is to remove duplicate records by clicking on <i>ProductID</i> and then <i>Remove Duplicate</i></p> <p>We now have a list of all unique Products in our data set.</p>	 <p>The screenshot shows a data table with a context menu open over the 'ProductID' column. The menu options are: Copy, Remove, Remove Other Columns, Duplicate Column, Add Column From Examples..., Remove Duplicates (highlighted with an orange box), Remove Errors, Change Type, Transform, Replace Values..., Replace Errors..., Group By..., Fill, Unpivot Columns, Unpivot Other Columns, Unpivot Only Selected Columns, Rename..., Move, Drill Down, and Add as New Query. The 'Remove Duplicates' option is highlighted with an orange box.</p>
7	<p>Now let us repeat this process to build out Geography (make sure to remove duplicate <i>Zip codes</i>) and Date (remove duplicate dates) dimensions</p> <p>(You have to do it on your own so you can practice)</p>	

8

Let us remove all unnecessary fields from our original data set now that we have moved them into separate dimensions:

1. Duplicate the **SalesData** Power Query again and rename the new one to say **SalesFact**
2. Remove all fields other than what we need to link back to our dimensions and let us also keep our measure (**Amount**)

Choose Columns

Choose the columns to keep



☐ (Select All Columns)

☒ ProductID

☐ Product

☐ Category

☐ Segment

☐ Manufacturer

☒ Zip

☐ City

☐ State

☐ Region

☒ Date

☒ Amount

OK

Cancel

- 9 Right click on **Product**, **Geography**, **Date** and **SalesFact** queries and click *Enable Load*

Click on *Home*->*Close & Apply*

SalesData	1 ² 3 ProductID
Product	1 791
Geography	2 759
Date	3 2345
SalesFact	4 609
	2045
	1109
	792
	992
	609
	674
	952
	533
	650
	2388
	1203
	599
	2388
	633
	2388
	573
	431
	2352

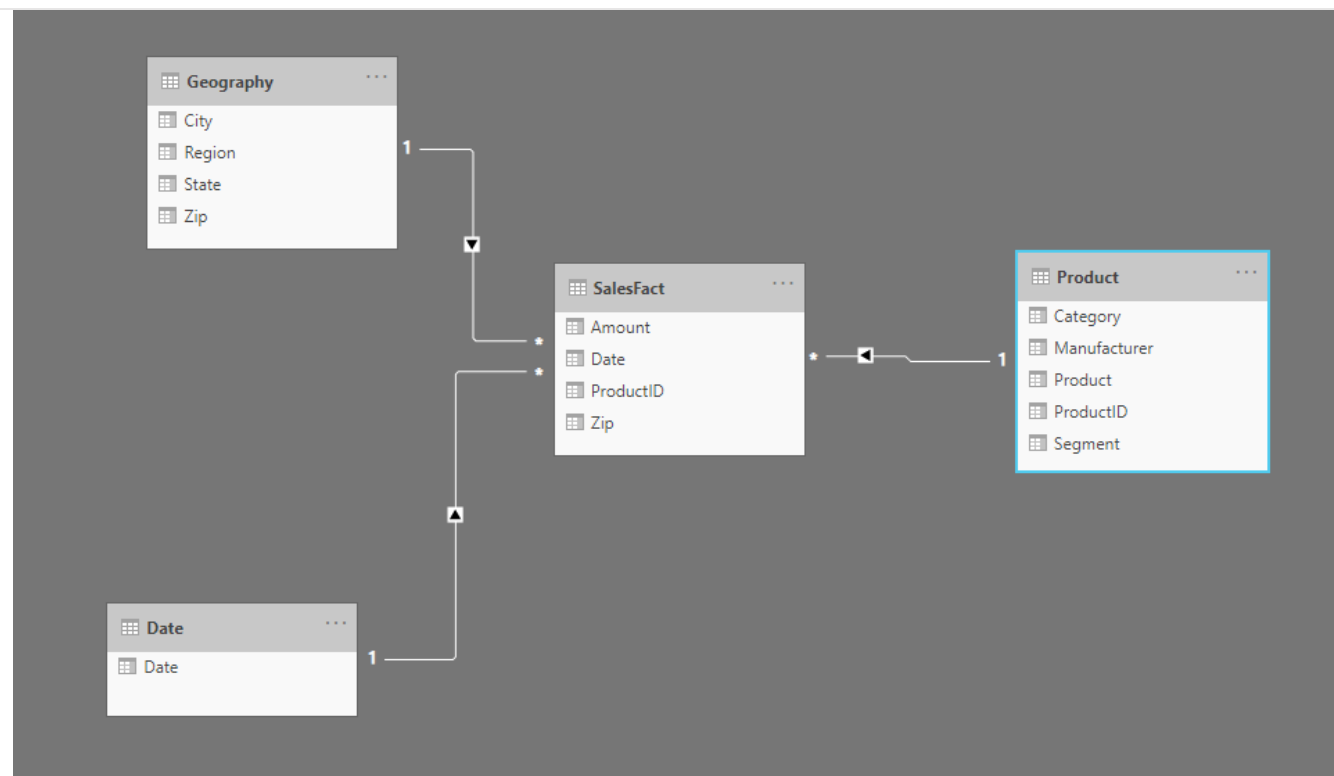
Copy
Paste
Delete
Rename
Enable load
Include in report refresh
Duplicate
Reference
Move To Group
Move Up
Move Down
Create Function...
Convert To Parameter
Advanced Editor
Properties...

10 Click on Relationship icon on the left



and review the auto detected relationships.

Note that in our case **Date** table is not linked to **SalesFact**. In your case one or more table may not have gotten linked to the SalesFact. That's ok, you will just need to create the relationship yourself by dragging and dropping the corresponding fields between the dimension and fact tables.



11 (Optional Step)

Alternatively, you can create a Relationship between **SalesFact** and **Date** tables (or any other tables where relationship did not get created automatically) by clicking on Home ->Manage Relationships -> New.



Select Tables and Columns used in the date relationship and click *OK* and then click *Close*.

Create Relationship

Select tables and columns that relate to one another.

SalesFact

ProductID	Zip	Date	Amount
791	18336	Tuesday, September 30, 2008	566.5743342
759	18337	Monday, May 24, 2004	1217.567824
2345	18337	Thursday, May 21, 2009	3569.544304
609	18403	Wednesday, January 23, 2008	7179.828201
2045	18407	Tuesday, December 14, 2004	3359.546404

Date

Date
Tuesday, September 30, 2008
Monday, May 24, 2004
Thursday, May 21, 2009
Wednesday, January 23, 2008
Tuesday, December 14, 2004

Advanced options

Cardinality

Many to One (*:1)

Cross filter direction

Single

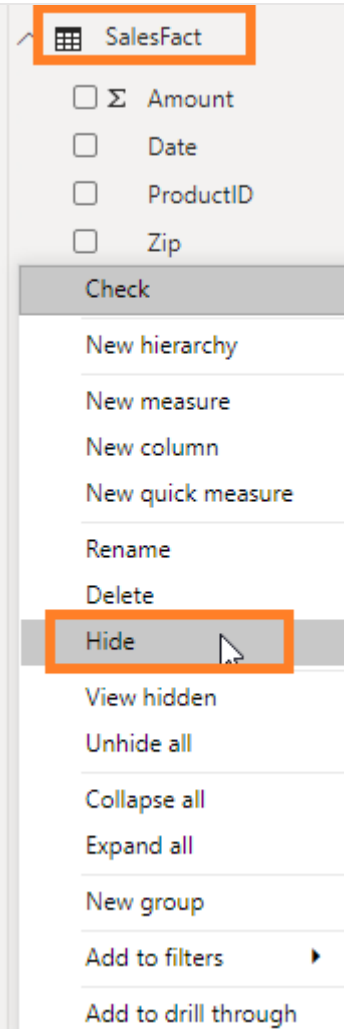
☒ Make this relationship active

StarSchemaLab - PBI Desktop.docx -

OK

Cancel

- 12 Hide irrelevant attributes in the **SalesFact** table (*Date, ProductID, Zip*) as demonstrated in the picture to the right



Now you can build some visualizations and explore the dataset. Our **Date** dimension is missing some key attributes such as Year, Month and Quarter. How can we address this and enhance the model with those attributes?