

Mastering Data Analytics

Copyright Notice & Disclaimer

- *The content contained in this module is provided only for educational purposes of Mastering Data Analytics' training courses. You may not copy, reproduce, distribute, publish, display, perform, modify, create derivative works, transmit, or in any way exploit any such content, nor may you distribute any part of this content over any network, including a local area network, sell or offer it for sale, or use such content to construct any kind of database.*
- *For permission to use the content, please contact via email: training@mastering-da.com*



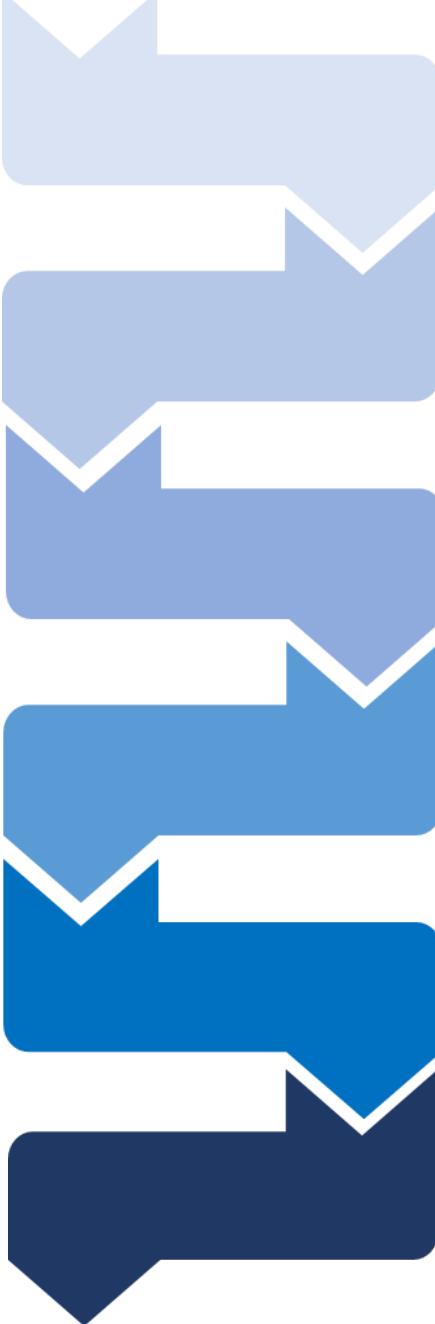


MASTERING DATA ANALYTICS



BUSINESS INTELLIGENCE

Module 2: End-to-End BI Workflow
in Power BI (Part 1)

- 
- 1 Fundamental BI & Analytical Thinking
 - 2 End-to-End BI Workflow in Power BI
 - 3 Descriptive Statistics & Analytics
 - 4 Diagnostics Analytics
 - 5 Analytical Idea Presentation
(Dashboard – Insight - Story)
 - 6 Business Intelligence Capstone



POWER BI DESKTOP

PART 1: Data Preparation

- 3.1 [Power Query Overview](#)
- 3.2 [Get Data](#)
- 3.3 [PQ – Basic Transform Data](#)
- 3.4 [Data Issues](#)

PART 2: Data Modeling

- 4.1 [Data Model Overview](#)
- 4.2 [Fact & Dimension](#)
- 4.3 [Schema](#)
- 4.5 [Cardinality](#)
- 4.6 [Cross Filter Direction](#)
- 4.7 [Hierarchies](#)

2. End-to-End Business Intelligence Workflow in Power BI

01

Data Preparation

1. Power Query Overview
2. Get Data
3. PQ – Basic Transform Data
4. Profiling Data
5. Data Issues
 - 4a. Bad Shape + Dirty Data
 - 4b. Missing Data + Outliers
5. Combine Data from Folder
6. Blending Data
7. Checklist

02

Data Modelling

1. Data Model Overview
2. Fact & Dimension
3. Schema
4. Cardinality
5. Cross Filter Direction
6. Hierarchies

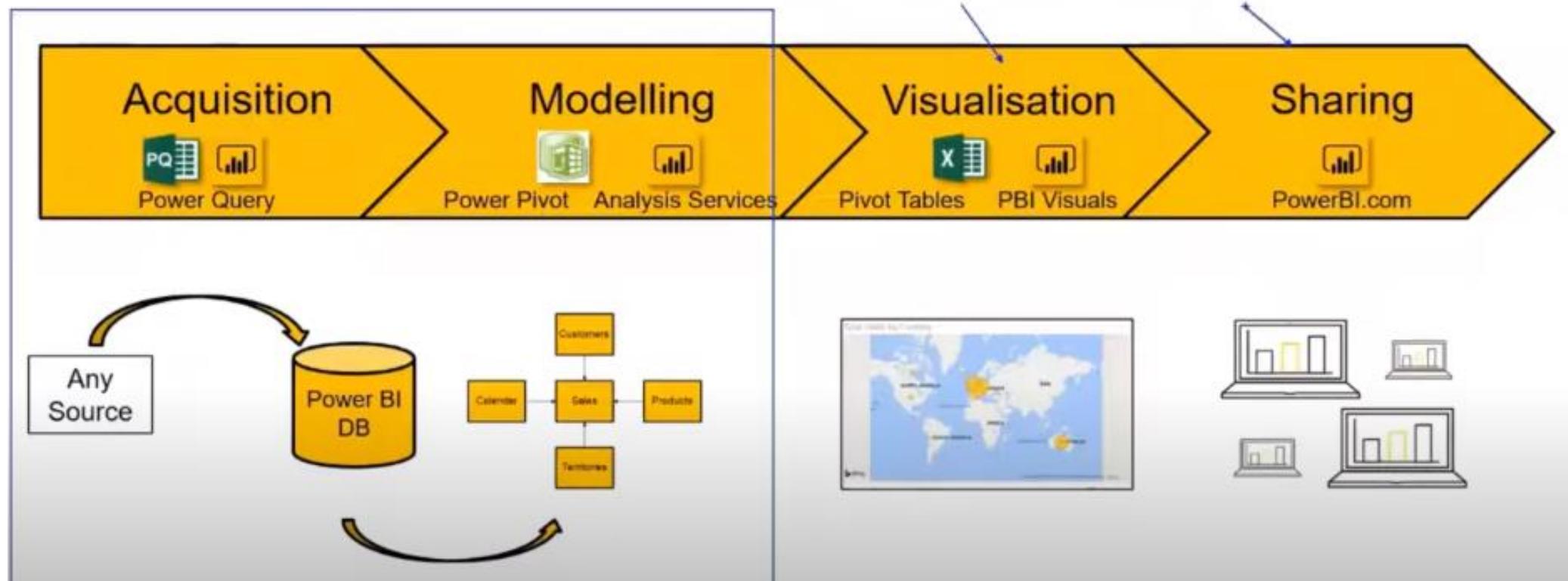
03

End-to-End in Power BI Cloud

1. Introduce PBI Ecosystem (PBI Service)
2. Fact & Dimension
3. Prep Data (On Pro & Premium)
4. Data Modeling
5. Report and Dashboard
6. Refresh Scorecard & Metrics
7. Sharing, Collaboration, (PBI Mobile)
8. Deployment Pipelines



Four Stages of Self Service BI

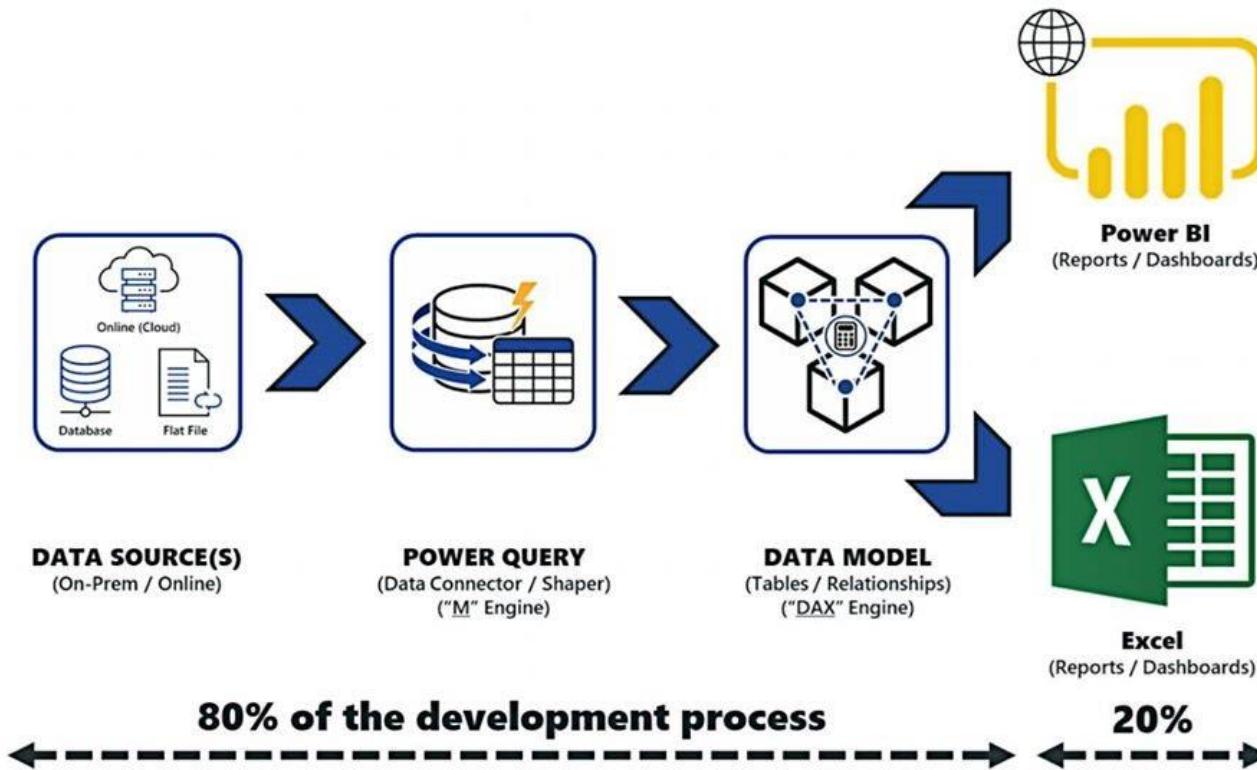


DATA PREPARATION

Power Query Overview



REPORT DEVELOPMENT PROCESS



POWER QUERY BENEFITS

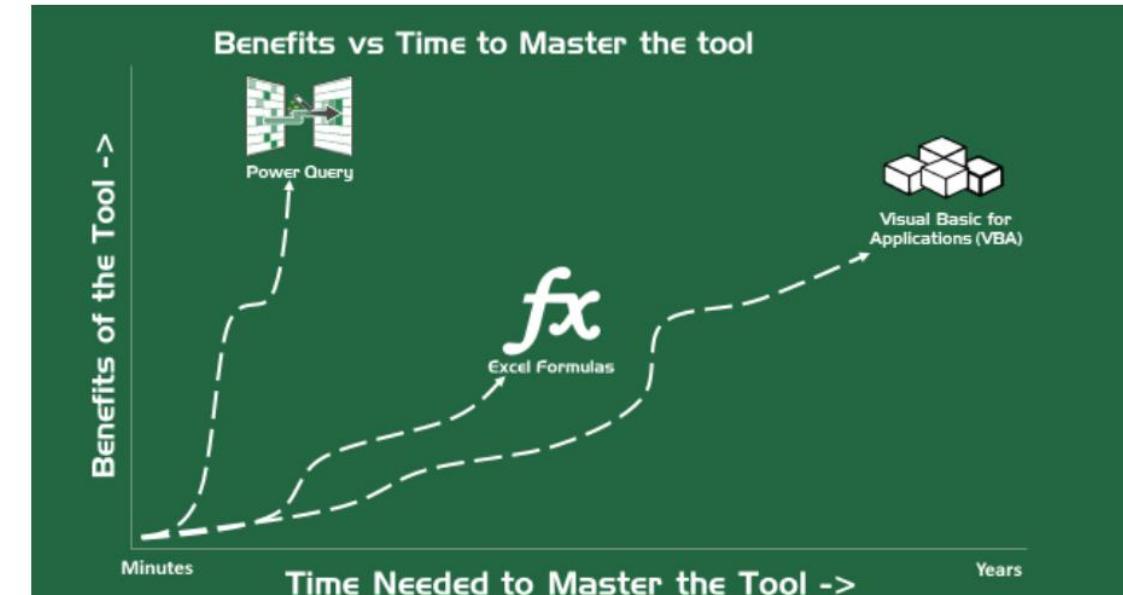


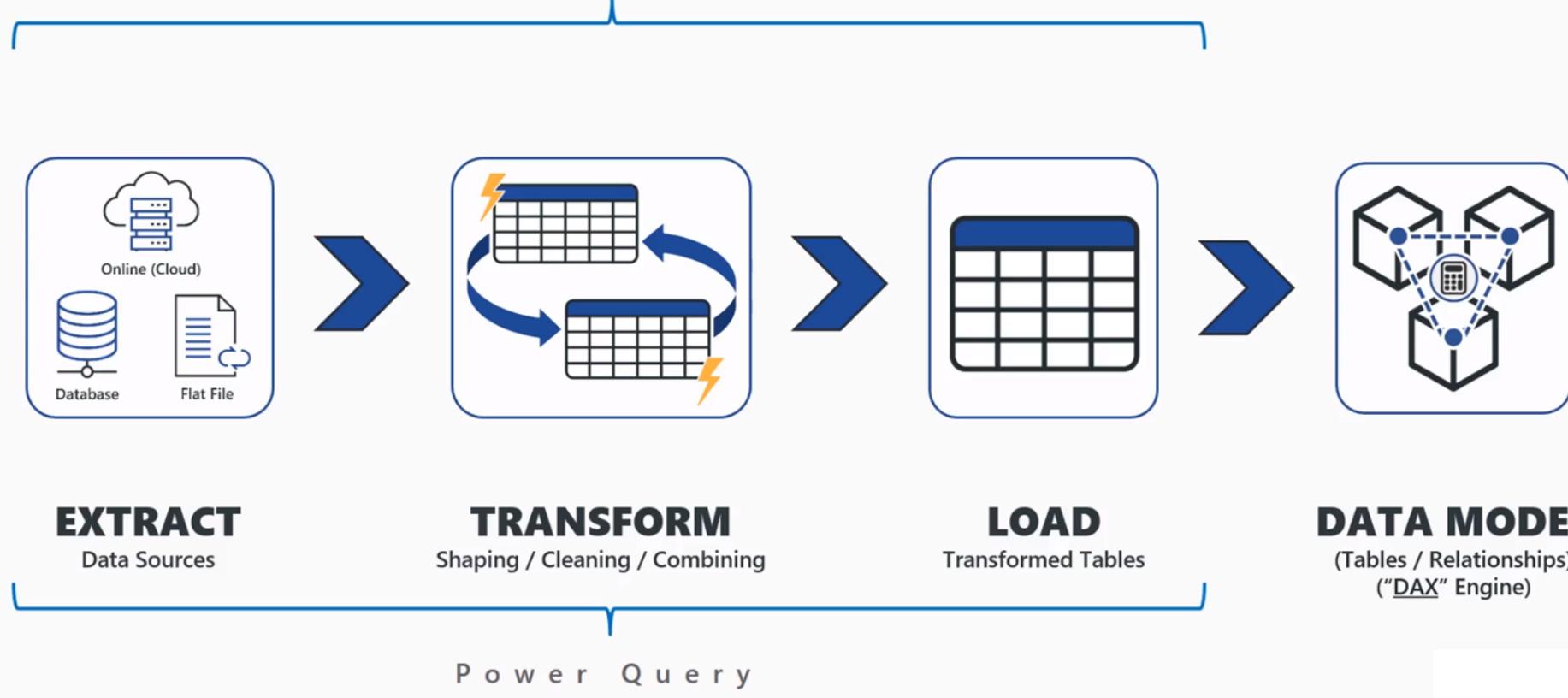
Figure 3 Power Query was designed to be an easy-to-use data transformation and manipulation tool.

DATA PREPARATION

Power Query Overview



A Transformative Tool For Data



DATA PREPARATION

Power Query Overview



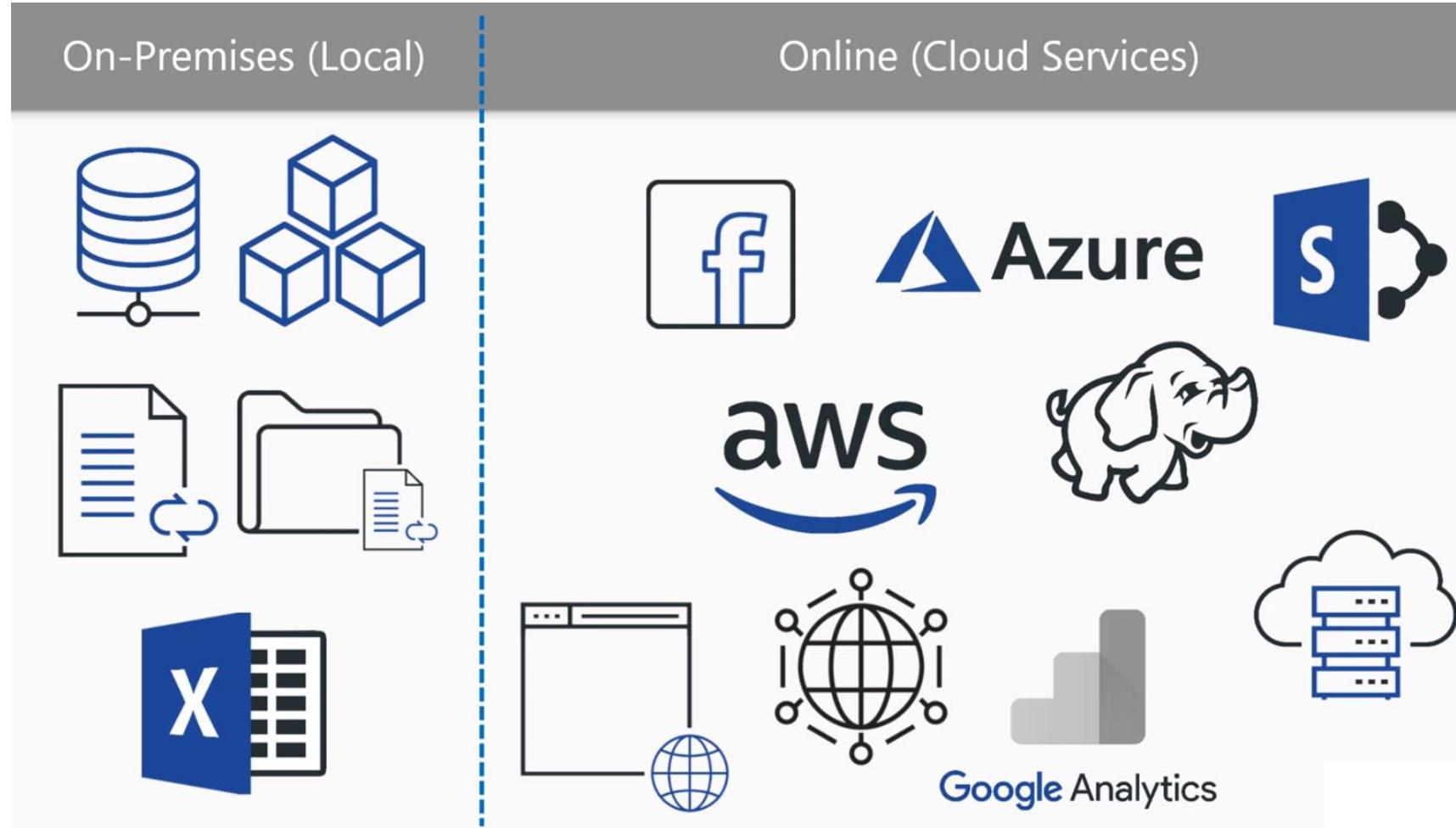
Power Query is an ETL tool. ETL stands for Extract, Transform and Load.

- **Extract** – Data can be extracted from a **variety of sources**: Databases, CSV files, Text files, Excel, Website and even PDF.
- **Transform** – After the data has been extracted, it can be **cleaned up** (i.e., remove spaces, split columns, change date formats, fill blanks, find and replace etc) and reshaped (i.e., unpivot, remove columns etc). When data is extracted from different sources it is unlikely to be consistent, the transform process is used to make it ready for use.
- **Load** – Once the data has been extracted and transformed, it needs to be put somewhere so that you can use it. From an Excel perspective, it can be **pushed into a worksheet, a data model, or another query**.

To summarize, Power Query takes data from different sources and turns it into something which can be used. As a tool, this is pretty useful already. But here is the best part. **Once the ETL process has been created, it can be run over and over again with a single click. Which can save hours of work every week.**

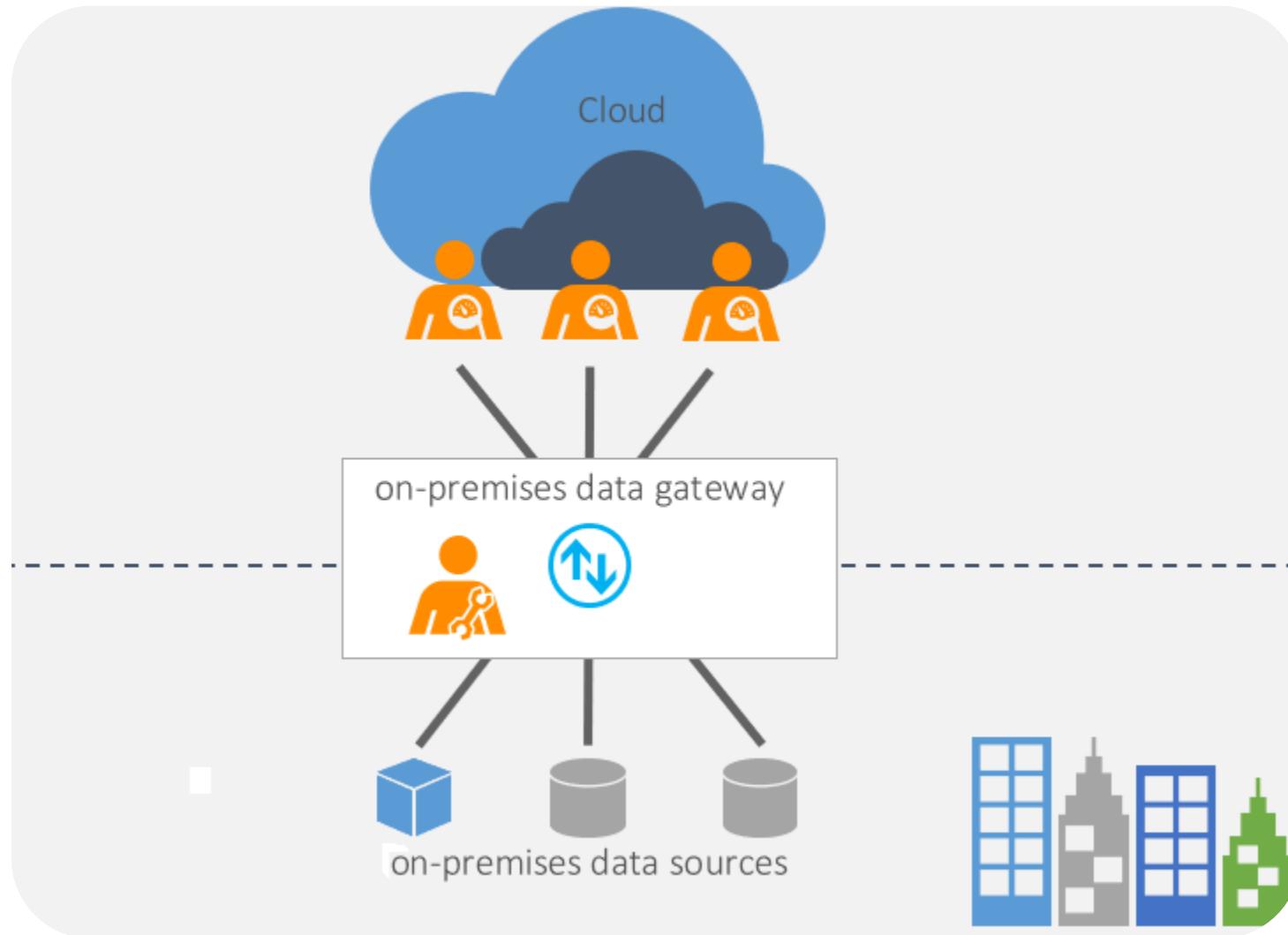
DATA PREPARATION

Get Data - Data Sources



DATA PREPARATION

Get Data - Data Gateway

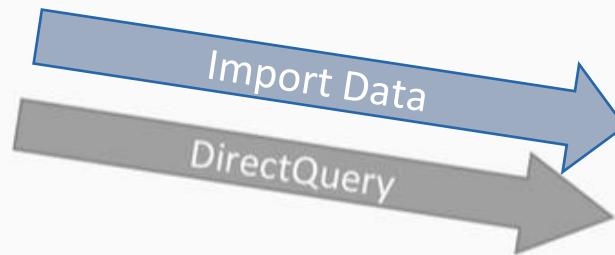


DATA PREPARATION

Composite Models



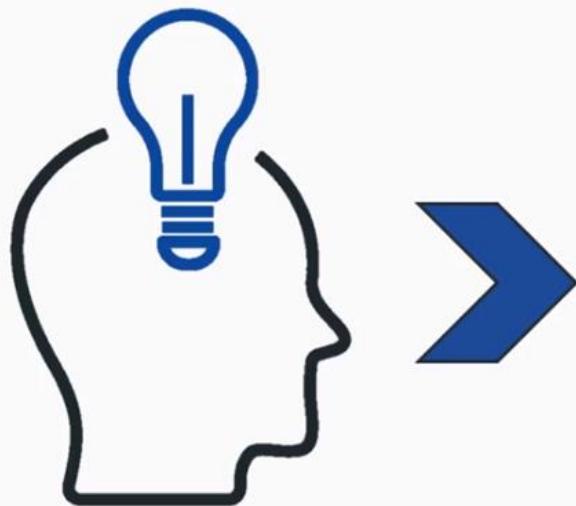
Composite Models



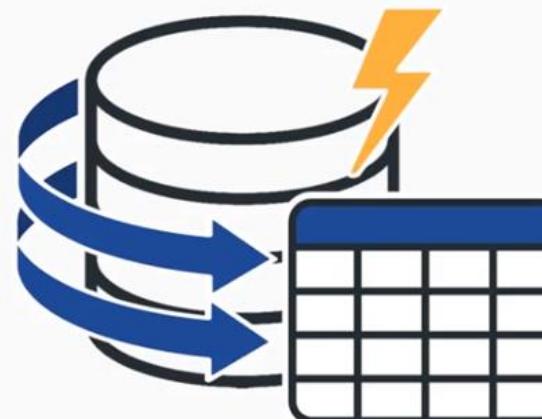


Power Query Fundamentals

DATA PREPARATION - Basic
Transform Data



BUSINESS LOGIC
(Shape, Transform, Clean)



POWER QUERY
(Data Connector / Shaper)
("M" Engine)

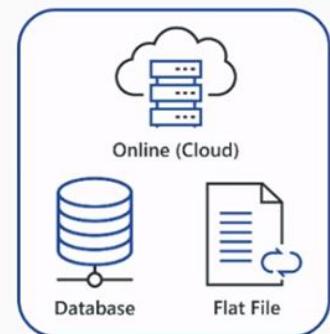


APPLIED STEPS
(Action Recorder)
(Sequential Steps)

DATA PREPARATION - Basic Transform Data

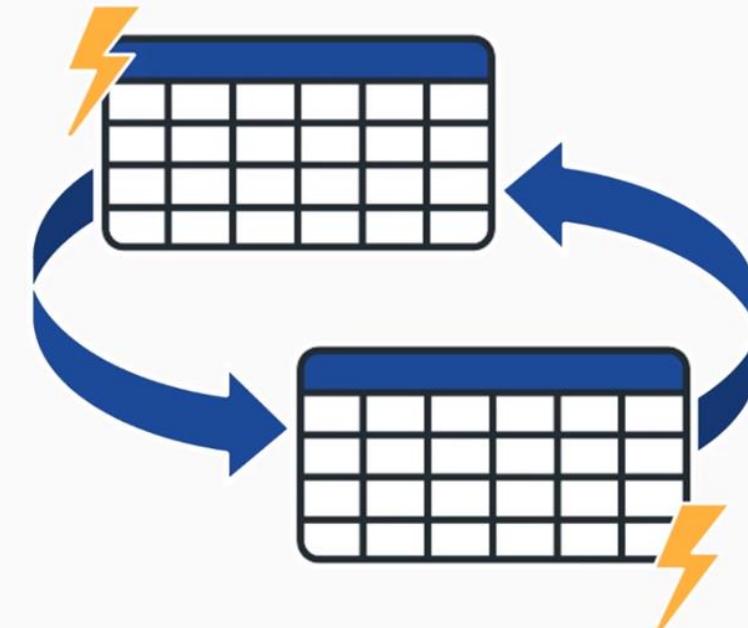


Report Development Process



DATA SOURCE(S)
(On-Prem / Online)

POWER QUERY
(Data Connector / Shaper)
("M" Engine)



TRANSFORM
Shaping / Cleaning / Combining



Essential Transformations For (Most) Tables



1) Choose Columns



2) Rename Columns

Prod_Name → Product Name



3) Change Data Type



APPLIED STEPS

(Action Recorder)

DATA PREPARATION - Basic Transform Data

Add/ Modify Columns



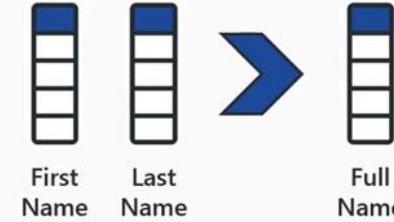
Common Ways To **Add** or **Modify** Columns



1) Add Columns



2) Merge Columns



Apple | iPhone XXIII → iPhone XXIII



3) Modify Columns

Sony | Playstation ∞ → Playstation ∞

APPLIED STEPS
(Action Recorder)

DATA PREPARATION - Basic Transform Data

Insert in Applied Steps



Screenshot of Power Query Editor showing the 'Transform' tab selected. The 'Address Line 2' column is highlighted.

The 'QUERY SETTINGS' pane is open, showing the 'PROPERTIES' section with 'Name' set to 'Store'. A blue arrow points to the 'STEPS' section, which lists 'Renamed Columns' and 'Changed Type'.

An 'Insert Step' dialog box is displayed, asking if the user wants to insert a step. It contains the message: "Are you sure you want to insert a step? Inserting an intermediate step may affect subsequent steps, which could cause your query to break." The 'Insert' button is highlighted with a yellow box and a cursor icon.

Address Line 2	Employee Count	Selling Area Size	Manager First Name	Manager Last Name
Saint Petersburg Citycenter, Shopping mall	95	480	Miguel	Saenz
West Ln, Baildon, Shipley Citycenter, United	32	680	Josh	Edwards
Nizhny Novgorod Downtown, Shopping mall	95	480	Kim	Akers
Leidui Shopping mall, Taipei, Taiwan	95	480	Ming-Yang	Xie
Green Bay Downtown	19	455	Nuno	Bento
East Queen Anne Big shopping mall	17	462	Roy	Antebi
Lackawanna Park St	26	680	Pilar	Pinilla
Mino shopping mall	25	700	Roy	Antebi
Kingsgate St Shopping mall	22	500	Jill	Shrader
Kennewick, Downtown	26	680	Prithvi	Raj
Citycenter Shopping mall Ramstein-Miesenbach, Germany	15	500	Greg	Winston
West Lake Hills shopping mall	19	455	Luca	Argentiero
4800 Spenard Road, Shopping mall	47	1125	Chris	Bryant
Yakima, shopping mall	47	1125	Zainal	Arifin
Pasadena, TX St Shopping mall	47	680	Barak	Regev
DT St Sydney, Australia	25	460	Ties	Arts
	17	1125	Don	Roessler
	47	700	Daniel	Taylor
	95	480	Mojca	Gostinacar
	19	455	Chris	Ashton
	47	1125	Mike	Ray
	33	560	Robert	Brown
	25	700	Denis	Dehenne
	1125	1125	Alan	Brewer
	47	400	Daniel	Goldschmidt

DATA PREPARATION - Basic Transform Data

Move in Applied Steps



Screenshot of the Microsoft Power BI Data Flow interface showing a data transformation step.

The interface includes:

- A main table view showing columns: Close Date, Zip Code, Store Phone, Address Line 1, and Address Line 2.
- An "Applied Steps" pane on the right containing "QUERY SETTINGS" for a step named "Store".
- A context menu open over the "Removed Contoso From Step" step, listing options like Move Up, Move Down, Extract Previous, View Native Query, and Properties... A red arrow points to the "Move Up" option.

Table Data Preview:

Close Date	Zip Code	Store Phone	Address Line 1	Address Line 2
null	2014522	510-555-0118	Saint Petersburg Citycenter, Shopping mall	Saint Petersburg Citycenter, Shopping mall
null	null	138-555-0118	West Ln, Baildon, Shipley Citycenter, United	West Ln, Baildon, Shipley Citycenter, United
null	147820	511-555-0119	Nizhny Novgorod Downtown, Shopping mall	Nizhny Novgorod Downtown, Shopping mall
null	109807	830-555-0128	Leidui Shopping mall, Taipei, Taiwan	Leidui Shopping mall, Taipei, Taiwan
null	54001	286-555-0189	Green Bay Downtown	Green Bay Downtown
null	97001	320-555-0195	East Queen Anne Big shopping mall	East Queen Anne Big shopping mall
null	32254	454-555-0119	Lackawanna Park St	Lackawanna Park St
null	97001	150-555-0189	Mino shopping mall	Mino shopping mall
7/12/2009	22302	822-555-0145	Kingsgate St Shopping mall	Kingsgate St Shopping mall
null	97001	212-555-0187	Kennewick, Downtown	Kennewick, Downtown
null	66877	555-555-0113	Citycenter Shopping mall Ramstein-Miesenbach, Germany	Citycenter Shopping mall Ramstein-Miesenbach, Germany
null	97001	612-555-0100	West Lake Hills shopping mall	West Lake Hills shopping mall
null	null	110-555-0115	4800 Spenard Road, Shopping mall	4800 Spenard Road, Shopping mall
null	97001	168-555-0183	Yakima, shopping mall	Yakima, shopping mall
null	14901	139-555-0120	Columbia St	Columbia St
null	97001	913-555-0172	Sunnyside, WA Citycenter	Sunnyside, WA Citycenter
null	7960	646-555-0185	Washington St	Washington St
null	97001	903-555-0145	Toppenish, Citycenter	Toppenish, Citycenter
null	201800	716-555-0127	China Shanghai, Jiading District Yongguang Rd	China Shanghai, Jiading District Yongguang Rd
4/12/2009	97001	206-555-0180	North Bend, Citycenter	North Bend, Citycenter
null	11238	158-555-0191	Flatbush Ave St	Flatbush Ave St
null	70001	373-555-0142	Egger Acres St	Egger Acres St
null	6759	164-555-0114	CTUDED Rd, Litchfield County	CTUDED Rd, Litchfield County
null	70001	314-555-0113	Pasadena, TX St Shopping mall	Pasadena, TX St Shopping mall
null	59000	1 (11) 500 555-	PT St Sydney, Australia	PT St Sydney, Australia
null	12140	712-555-0113	Rensselaer County St	Rensselaer County St
null	null	318-555-0137	22 Market Sq Shipley, West Yorkshire, BD18 3QJ, United Kingdom	22 Market Sq Shipley, West Yorkshire, BD18 3QJ, United Kingdom

DATA PREPARATION - Basic Transform Data

Data Types



Data Types



Numeric Data Types

- Whole Number
- Decimal Number
- Fixed Decimal Number (Floating point stored as integer)
- Boolean

Date/Time Data Types

- Date – Internally stored as an integer
- Time – Internally stored as a fraction between 0 and 1
- Date Time

Other Data Types

- Text
- **Any – You should never see this in a data model. Bad things can happen!!**

*Set your
Data Types
in the
Query Editor*

*Set your
Data Formats
(\$ %, etc)
in the Data Model*

Pro Tip: Data type is different from data format

DATA PREPARATION - Basic Transform Data

Data Type vs Data Format



Contoso Final Report - Power BI Desktop

Thao Phuong

File Home Help Table tools Column tools

Name DateKey Date

Summarization Don't summarize

Format dd,mmm d,yyyy

Date formats

- *3/14/2001 (m/d/yyyy)
- Wednesday, March 14, 2001 (ddd, mmmm d, yyyy)
- March 14, 2001 (mmm d, yyyy)
- Wednesday, 14 March, 2001 (ddd, d mmmm, yyyy)
- 14 March, 2001 (d mmmm, yyyy)
- 3/14/01 (m/d/yy)
- 03/14/01 (mm/dd/yy)
- 03/14/01 (mm/dd/yy)
- 01/03/14 (yy/mm/dd)
- 2001-03-14 (yyyy-mm-dd)
- 14-Mar-01 (dd-mmmm-yy)
- 14/03/2001 (dd/mm/yyyy)
- Saturday, April 7, 2007 (mmm yyyy)
- March 2001 (mmmmm yyyy)
- 2001-03 (yyyy-mm)
- March 14 (mmmmm d)
- 01 (yy)
- 2001 (yyyy)

Sort by column Groups Manage relationships New column

Fields

Search

DAX Measures DIM Calendar DIM Channel DIM Geography DIM Product MAS... DIM Store FACT Sales

File Home Transform Add Column View

Transpose Data type: Date Replace Values Unpivot Columns

Group Use First Row By as Headers Count Rows Rename Pivot Column Convert to List

Move ABC ABC Extract Move Split Column Format Parse

Any Column Text Column Number Column

Statistics Standard Scientific Information

10² 0.0 Rounding Trigonometry

Date Time Duration

Expand Aggregate Extract Values

Calculated Day of Week

Properties Name: 2 Databl... All Properties

Applied Steps Source: Filtered Hidden Files1 Invoke Custom Function1 Renamed Columns1 Removed Other Columns1 Expanded Table Column1 Changed Type Duplicated Column Calculated Day of Week

	1.2 SalesAmount	123 GeographyKey	Date	123 Date - Copy
1	658.08	1431	787	1/1/2007
2	1527.96	2847.15	558	1/1/2007
3	65.24	116.8	800	1/1/2007
4	1903.86	4128.5	800	1/1/2007
5	136.8	268	916	1/1/2007
6	136.44	408.5865	800	1/1/2007
7	28.56	50.4	586	1/1/2007
8	819.45	1782	835	1/1/2007
9	1783.84	4845.6	529	1/1/2007
10	91.76	170.962	916	1/1/2007
11	608.76	1164.15	734	1/1/2007
12	265.08	467.964	586	1/1/2007
13	347.64	680.4	754	1/1/2007
14	69.32	135.96	790	1/1/2007
15	811.17	1724.8	932	1/1/2007
16	1117.35	2317.56	710	1/1/2007
17	265.08	506.961	922	1/1/2007
18	1102.92	2343.15	710	1/1/2007
19	6116.64	11697.66	756	1/1/2007
20	265.08	513.4605	903	1/1/2007
21	303.84	581.1	909	1/1/2007

DATA PREPARATION - Basic Transform Data

Detect Data Type



Screenshot of the Power Query Editor interface showing a table of product data. The table has 34 rows and 11 columns, with the first column being the primary key. The columns are: Product Subcategory Key, Manufacturer, Brand Name, Class Name, Color Name, Size, Weight, Weight Unit Measure ID, Unit Cost, and Unit Price. A 'Ctrl + A' keyboard shortcut is overlaid on the table area, indicating a selection action. The 'QUERY SETTINGS' pane on the right shows the query is named 'Product' and includes steps for Source, Navigation, and Removed Other Columns.

	Product Subcategory Key	Manufacturer	Brand Name	Class Name	Color Name	Size	Weight	Weight Unit Measure ID	Unit Cost	Unit Price
1	Contoso, Ltd	Contoso	Economy	Silver	2.2 x 2.2 x 4	4.5 ounces	11	21.57		
2	Contoso, Ltd	Contoso	Economy	Silver	0.8 x 3.6 x 1.1	5.6 ounces	30.58	59.99		
3	Contoso, Ltd	Contoso	Economy	Blue	0.8 x 3.6 x 1.1	7.4 ounces	55.72	77.68		
4	Contoso, Ltd	Contoso	Regular	White	3.8 x 0.6 x 2.2	11 ounces	50.56	109.95		
5	Contoso, Ltd	Contoso	Regular	Black	5.5 x 1.7 x 5.3	1 pounds	61.62	134		
6	Contoso, Ltd	Contoso	Regular	Blue	3.8 x 0.6 x 2.2	11 ounces	91.93	199.9		
7	Contoso, Ltd	Contoso	Regular	Red	1.2 x 0.8 x 2.8	2 ounces	91.93	199.9		
8	Contoso, Ltd	Contoso	Regular	Pink	0.7 x 3.6 x 2.8	8.8 ounces	84.49	255		
9	Contoso, Ltd	Contoso	Regular	Yellow	3.5 x 2 x 0.5	2.9 pounds	48.92	55.95		
10	Contoso, Ltd	Contoso	Deluxe	Blue	4.1 x 2.4 x 0.4	5 ounces	99.14	299.23		
11	Contoso, Ltd	Contoso	Regular	blue	2.2 x 1.8 x 4	1 pounds	106.69	232		
12	Wide World Importers	Wide World Importers	Economy	Silver	150X1x15	30 grams	76.45	149.95		
13	Wide World Importers	Wide World Importers	Regular	Silver	6 x 0.5 x 0.5	2.2 ounces	91.95	199.95		
14	Wide World Importers	Wide World Importers	Deluxe	Yellow	2.5 x 0.5 x 1	1.3 ounces	98.07	296		
15	Wide World Importers	Wide World Importers	Economy	White	150X1x15	30 grams	79.53	156		
16	Wide World Importers	Wide World Importers	Regular	Yellow	0.5 x 0.5 x 0.1	2.2 ounces	83.24	181		
17	Northwind Traders	Northwind Traders	Economy	White	1.1 x 1 x 1	1.6 ounces	13.1	25.69		
18	Northwind Traders	Northwind Traders	Economy	Red	1.9 x 9 x 7	2.1 pounds	22.05	47.95		
19	Northwind Traders	Northwind Traders	Economy	Red	5.6 x 3.5 x 3.1	1 pounds	17.45	37.95		
20	Northwind Traders	Northwind Traders	Economy	Yellow	5 x 5 x 2	9.6 ounces	18.65	40.55		
21	Northwind Traders	Northwind Traders	Regular	Red	11 x 7.7 x 2.5	1.2 pounds	45.98	99.99		
22	Northwind Traders	Northwind Traders	Regular	Black	10 x 8 x 3	10.4 ounces	49.69	149.99		
23	Northwind Traders	Northwind Traders	Regular	Red	6 x 4 x 10	2 pounds	49.69	149.99		
24	Wide World Importers	Wide World Importers	Economy	White	5 x 5.5 x 2.1	0.17 pounds	34.36	67.4		
25	Wide World Importers	Wide World Importers	Regular	White	3.3 x 3 x 4.7	5.9 ounces	55.18	120		
26	Wide World Importers	Wide World Importers	Regular	White	7 x 5 x 1.5	1 pounds	52.88	115		
27	Wide World Importers	Wide World Importers	Regular	Yellow	10.4 x 8.1 x 3.1	11.4 ounces	61.16	132.99		
28	Wide World Importers	Wide World Importers	Deluxe	Blue	11 x 7.8 x 3	2 pounds	82.83	249.99		
29	Adventure Works	Adventure Works	Economy	Silver	null	27.9 pounds	86.67	169.99		
30	Adventure Works	Adventure Works	Black	null	20.1 pounds	61.17	119.99			
31	Adventure Works	Adventure Works	Regular	White	null	9.6 pounds	128.76	279.99		
32	Adventure Works	Adventure Works	Economy	Brown	null	2 pounds	73.11	143.4		
33	Adventure Works	Adventure Works	Economy	Brown	null	20.9 pounds	101.97	200		
34	Adventure Works	Adventure Works	Regular	Rewnn	null	25.7 minivnnts	160.98	349.95		

DATA PREPARATION - Basic Transform Data

Detect Data Type



Learn Power BI Report - Power Query Editor

File Home Transform Add Column View Help

Transpose Reverse Rows Data Type: Any Replace Values Unpivot Columns Group Use First Row By as Headers Count Rows Detect Data Type Rename Pivot Column Move Convert to List Split Column Format ABC Extract abc Parse Statistics Standard Scientific Trigonometry Rounding Date Time Duration Information Date & Time Column String

Queries [11]

- Transform File from FACTSales [3]
 - Sample Query [2]
 - Sample File Parameter1 (Sample)
 - Sample File
 - Transform Sample File from FACT...
 - Transform File from FACTSales
- Other Queries [7]
 - Channel
 - Geography
 - Product
 - Product Category
 - Product Sub Category
 - Store
 - Sales

	ABC 123 Store Key	ABC 123 Store Manager	ABC 123 Store Name	ABC 123 Open Date	ABC 123 Close Date	ABC 123 Zip Code	ABC 123 Sto
1	304		14 Contoso Saint Petersburg Store	6/7/2004		null	2014522
2	209		227 Contoso Bailldon Store	9/22/2004		null	138-555
3	305		15 Contoso Nizhny Novgorod Store	3/15/2004		null	147820
4	302		24 Contoso Taipei Store	7/19/2004		null	109807
5	42		71 Contoso Green Bay Store	8/3/2003		null	54001
6	1		35 Contoso Seattle No.1 Store	4/12/2004		null	97001
7	103		120 Contoso Jacksonville Store	10/8/2003		null	32254
8	2		35 Contoso Seattle No.2 Store	2/14/2004		null	97001
9	184		198 Contoso Alexandria Store	11/6/2004	7/12/2009	22302	822-555
10	3		36 Contoso Kennewick Store	2/12/2004		null	97001
11	235		252 Contoso Ramstein Store	8/13/2004		null	66877
12	4		37 Contoso Bellevue Store	3/1/2004		null	97001
13	183		111 Contoso Anchorage Store	12/2/2004		null	110-555
14	6		39 Contoso Yakima Store	5/15/2005		null	97001
15	126		144 Contoso Elmira Store	2/24/2004		null	14901
16	8		41 Contoso Sunnyside Store	7/2/2004		null	97001
17	144		160 Contoso Morristown Store	11/24/2004		null	7960
18	9		42 Contoso Toppenish Store	8/18/2004		null	97001
19	281		290 Contoso Shanghai No.2 Store	3/2/2004		null	201800
20	12		45 Contoso North Bend Store	11/21/2004	4/12/2009	97001	206-555
21	121		140 Contoso Brooklyn Store	9/19/2004		null	11238

DATA PREPARATION - Basic Transform Data

Merge Column



Screenshot of a data preparation tool interface showing a table of data and a context menu open over the last column.

The table has columns: Address Line 2, Employee Count, Selling Area Size, Manager First Name, and Manager Last Name.

The context menu for the Manager First Name column includes:

- Copy
- Remove Columns
- Remove Other Columns
- Add Column From Examples...
- Remove Duplicates
- Remove Errors
- Replace Values...
- Fill
- Change Type
- Transform
- Merge Columns** (highlighted with a red circle)
- Group By...
- Unpivot Columns
- Unpivot Other Columns
- Unpivot Only Selected Columns
- Move

The "QUERY SETTINGS" pane shows:

- PROPERTIES**: Name = Store, All Properties
- APPLIED STEPS**: Source, Navigation, Removed Other Columns, Renamed Columns, Changed Type

Table Data:

Address Line 2	Employee Count	Selling Area Size	Manager First Name	Manager Last Name
ersburg Citycenter, Shopping mall	95	480	Miguel	
Baldon, Shipley Citycenter, United	32	680	Josh	
ovgorod Downtown, Shopping mall	95	480	Kim	
opping mall, Taipei, Taiwan	95	480	Ming-Yang	
ry Downtown	19	455	Nuno	
en Anne Big shopping mall	17	462	Roy	
nna Park St	26	680	Pilar	
opping mall	25	700	Roy	
e St Shopping mall	22	500	Jill	
ck, Downtown	26	680	Prithvi	
er Shopping mall Ramstein-Miesenbach, Germany	15	500	Greg	
te Hills shopping mall	19	455	Luca	
enard Road, Shopping mall	47	1125	Chris	
shopping mall	47	1125	Zainal	
a St	26	680	Barak	
e, WA Citycenter	17	460	Ties	
ton St	47	1125	Don	
ih, Citycenter	25	700	Daniel	Taylor
anghai, Jiading District Yongguang Rd	95	480	Mojca	Gostincar
nd, Citycenter	19	455	Chris	Ashton
Ave St	47	1125	Mike	Ray
res St	33	560	Robert	Brown
Rd, Litchfield County	25	700	Denis	Dehenne
a, TX St Shopping mall	47	1125	Alan	Brewer
dney, Australia	95	480	Daniel	Goldschmidt
er County St	19	455	Sean	Chai
et Sq Shipley, West Yorkshire, BD18 3QJ, United Kingdom	90	680	Bernard	Tham

DATA PREPARATION - Basic Transform Data

Split Column



In Power Query, you can split a column through different methods. In this case, the column(s) selected can be split by a delimiter

The screenshot shows the Power Query Editor interface with the following details:

- File Bar:** Home, Transform, Add Column, View, Tools, Help.
- Toolbars:** Close & Apply, New Source, Recent Sources, Enter Data, Data source settings, Manage Parameters, Refresh Preview, Properties, Advanced Editor, Choose Columns, Remove Columns, Keep Rows, Remove Rows, Sort.
- Queries List:** Accounts (selected).
- Table View:** Accounts table with rows 1-4.
- ContextMenu (Main):** Split Column (highlighted), Group By, Use First Row as Headers, Data Type: Text, Replace Values.
- ContextMenu (Sub):** By Delimiter (highlighted), By Number of Characters, By Positions, By Lowercase to Uppercase, By Uppercase to Lowercase, By Digit to Non-Digit, By Non-Digit to Digit.
- ContextMenu (Accounts):** Copy, Remove, Remove Other Columns, Duplicate Column, Add Column From Examples..., Remove Duplicates, Remove Errors, Change Type, Transform, Replace Values..., Replace Errors..., Split Column (highlighted), Group By..., Fill, Unpivot Columns, Unpivot Only Selected Columns, Rename..., Move, Drill Down, Add as New Query.
- ContextMenu (Sub):** By Delimiter... (highlighted), By Number of Characters..., By Positions..., By Lowercase to Uppercase, By Uppercase to Lowercase, By Digit to Non-Digit, By Non-Digit to Digit.

DATA PREPARATION - Basic Transform Data

Replace



Replace Values

Replace one value with another in the selected columns.

Value To Find: Contoso

Replace With:

> Advanced options

OK Cancel

Store Key	Store Manager	Store Name	Open Date	Address	City, Country
304	14	Contoso Saint Petersburg Store	3/13/2005	null	97001 106-555-0185
209	227	Contoso Baildon Store	2/24/2004	null	14901 139-555-0120
305	15	Contoso Nizhny Novgorod Store	7/2/2004	null	97001 913-555-0172
302	24	Contoso Taipei Store	11/24/2004	null	7960 646-555-0185
42	71	Contoso Green Bay Store	8/18/2004	null	97001 903-555-0145
1	35	Contoso Seattle No.1 Store	3/2/2004	null	201800 716-555-0127
103	120	Contoso Jacksonville Store	11/21/2004	4/12/2009	97001 206-555-0180
2	35	Contoso Seattle No.2 Store	9/19/2004	null	11238 158-555-0191
184	198	Contoso Alexandria Store	11/27/2004	null	70001 373-555-0142
3	36	Contoso Kennewick Store	7/23/2004	null	6759 164-555-0114
235	252	Contoso Ramstein Store	5/21/2004	null	CTUDED Rd, Litchfield County
4	37	Contoso Bellevue Store	3/18/2004	null	70001 314-555-0113
183	111	Contoso Anchorage Store	1/25/2004	null	Pasadena, TX St Shopping mall
6	39	Contoso Yakima Store	3/21/2004	null	12140 712-555-0113
126	144	Contoso Elmira Store			Rensselaer County St
8	41	Contoso Sunnyside Store			22 Market Sq Shipley, West Yorkshire, BD
144	160	Contoso Morristown Store			
9	42	Contoso Toppenish Store			
281	290	Contoso Shanghai No.2 Store			
12	45	Contoso North Bend Store			
121	140	Contoso Brooklyn Store			
89	107	Contoso Round Rock Store			
178	193	Contoso Litchfield County Store			
83	101	Contoso Pasadena Store			
300	294	Contoso Sydney No.2 Store			
127	145	Contoso Poestenkill Store			
208	226	Contoso West Yorkshire Store			

DATA PREPARATION - Basic Transform Data

Trim & Clean



Trim - Remove leading and trailing whitespaces from each cell in the selected columns!
Clean - Remove non-printable characters in the selected columns!

Query Editor

File Home Transform Add Column View

Table Any Column

Queries [1] Table1

Category Subcategory Value

1 Administration Water

2 Administration Water

3 Administration Cellphone

4 Administration Cellphone

Untrimmed Data	Trimmed Data
Marry Angela	Marry Angela
Joey Tribiani	Joey Tribiani
Mathew Pery	Mathew Pery
Jenier Aniston	Jenier Aniston

Uncleaned Data	Cleaned Data
•I live in India	I live in India
‣I live in India	I live in India
⇒I live in India	I live in India
↳I live in India	I live in India

DATA PREPARATION - Basic Transform Data

Filter Table



Why Should We Filter Data?



APPLY FILTERS
(Filtering = Reducing Rows)



DATA PREPARATION - Basic Transform Data

Filter



Screenshot of a data preparation tool interface showing a table of sales data and a query settings panel.

The table contains columns: ReturnQuantity, ReturnAmount, DiscountQuantity, DiscountAmount, TotalCost, SalesAmount, GeographyKey, Date, and a row header column.

The Query Settings panel shows:

- Properties:** Name is FACTSales.
- Applied Steps:** A list of transformations applied to the source data, including:
 - Filtered Hidden Files1
 - Invoke Custom Function1
 - Renamed Columns1
 - Removed Other Columns1
 - Expanded Table Column1
 - Changed Type
 - Filtered Date >= 2007 (highlighted)

A context menu is open over the SalesAmount column, showing options like Sort Ascending, Sort Descending, Clear Sort, Clear Filter, Remove Empty, and Number Filters. The Number Filters dropdown is expanded, showing a list of numerical values with checkboxes, and includes a search bar and a note: "List may be incomplete".

	ReturnQuantity	ReturnAmount	DiscountQuantity	DiscountAmount	TotalCost	SalesAmount	GeographyKey	Date
9	0	0	0	0			787	1/1/20
3	1	999	0	1			558	1/1/20
4	0	0	0	7			800	1/1/20
18	0	0	0	1			800	1/1/20
80	1	3.35	0	0			916	1/1/20
18	0	0	0	3			800	1/1/20
2	1	28	0	1				
9	0	0	0	0				
8	0	0	0	4				
4	0	0	0	4				
6	0	0	0	1				
4	0	0	0	2				
4	1	189	0	2				
4	0	0	0	0				
9	0	0	0	4				
13	0	0	0	4				
4	0	0	0	2				
13	0	0	0	2				
24	0	0	0	3				
4	0	0	0	1				
4	0	0	0	2				
4	0	0	0	2				
16	1	75.99	0	4				
8	1	22.99	0	4				
9	0	0	0	4				
8	1	300	0	4				
13	1	627	0	2				
2	1	999.9	0	4				
13	0	0	0	4				

DATA PREPARATION - Basic Transform Data

Advanced Filter



673	8	0	0	4	538.4	1783.84	4845.6	529
44.99	4	0	0	4	8.998	91.76	170.962	916
199	6						X 15	734
129.99	4						64	586
189	4						0.4	754
33.99	4						96	790
196	9						4.8	932
186.9	13						56	710
129.99	4						61	922
184.5	13						15	710
499.9	24						66	756
129.99	4						05	903
149	4						1.1	909
24.99	4						61	921
129.99	4						62	921
75.99	16						42	894
22.99	8						28	529
184.5	9						3.6	852
300	8	1	300	4	240	965.72	2160	772
627	13	1	627	2	188.1	2492.88	7962.9	710
999.9	2	1	999.9	4	199.98	1019.56	1799.82	905
116	13	0	0	4	69.6	693.42	1438.4	712
219	4	0	0	0	0	446.6	876	933
33.99	4	1	33.99	0	0	69.32	135.96	758
22.89	36	1	22.89	3	3.4335	272.88	820.6065	894
427	8	0	0	0	0	1131.76	3416	754
19.95	8	0	0	2	7.98	73.36	151.62	752
139.99	9	0	0	2	13.999	642.33	1245.911	845
21	9	0	0	2	2.1	86.94	186.9	947

DATA PREPARATION - Basic Transform Data

Advanced Filter



Learn Power BI Report - Power Query Editor

File Home Transform Add Column View Help

Transpose Replace Values Unpivot Columns Merge Columns

Group By Use First Row as Headers Reverse Rows Detect Data Type Fill Move

Count Rows Rename Pivot Column Convert to List

Table Any Column Text Column Number Column Date & Time Column Structured Column

Statistics Standard Scientific Trigonometry Rounding Information

Date Time Duration

Expand Aggregate Extract Values Run Script

Queries [11]

Source.Name

1 SalesKey 2 DateKey 3 channelKey 4 StoreKey 5 ProductKey 6 PromotionKey 7 UnitCost 8 UnitPrice

4687 39083 1 93 700 2 73.12

6576 39083 1 253 1845 5 509.32

10301 39083 2 199 1822 2 16.31

12973 39083 2 199 1444 2 105.77

13143 39083 1 125 2517 2 1.71

13661 39083 2 199 1590 2 7.58

39083 2 306 1734 10 14.28

39083 1 198 1012 2 91.05

39083 4 309 1088 10 222.98

39083 1 125 2438 2 22.94

39083 1 254 288 5 101.46

39083 2 306 1914 10 66.27

39083 1 230 1440 10 86.91

39083 1 40 928 2 17.33

39083 1 171 728 2 90.13

39083 4 310 968 5 85.95

46007 39083 1 116 1962 2 66.27

46635 39083 4 310 1036 5 84.84

50055 39083 1 233 420 10 254.86

50751 39083 1 282 1902 2 66.27

53274 39083 1 162 2162 2 75.96

56110 39083 1 185 1284 2 12.74

18 2007 Contoso Data.xlsx

19 2007 Contoso Data.xlsx

20 2007 Contoso Data.xlsx

21 2007 Contoso Data.xlsx

22 2007 Contoso Data.xlsx

OK Cancel

Equals... Does Not Equal... Begins With... Does Not Begin With... Ends With... Does Not End With... Contains... Does Not Contain...

⚠ List may be incomplete. Load more

DATA PREPARATION - Basic Transform Data

Add Column from Example



Screenshot of the Power Query Editor showing the "Add Column From Examples" dialog.

The dialog shows a preview of the first 10 rows of a table with columns: Name & postal abbreviation[12], Name & postal abbreviation[12]2, Cities Capital, Cities Largest[16], Established[C], Population [D][14], Total area[15] mi², and Text Before Delimiter.

The "Text Before Delimiter" column contains the state names: Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, and Delaware.

The formula bar at the top shows: Transform: Text.BeforeDelimiter(["#Name & postal abbreviation[12]"], ",")

	Name & postal abbreviation[12]	Name & postal abbreviation[12]2	Cities Capital	Cities Largest[16]	Established[C]	Population [D][14]	Total area[15] mi ²	Text Before Delimiter
3	Alabama	AL	Montgomery	Birmingham	Dec 14, 1819	4,874,747	52,420	Alabama
4	Alaska	AK	Juneau	Anchorage	Jan 3, 1959	739,795	665,384	Alaska
5	Arizona	AZ	Phoenix	Phoenix	Feb 14, 1912	7,016,270	113,990	Arizona
6	Arkansas	AR	Little Rock	Little Rock	Jun 15, 1836	3,004,279	53,179	Arkansas
7	California	CA	Sacramento	Los Angeles	Sep 9, 1850	39,536,653	163,695	California
8	Colorado	CO	Denver	Denver	Aug 1, 1876	5,607,154	104,094	Colorado
9	Connecticut	CT	Hartford	Bridgeport	Jan 9, 1788	3,588,184	5,543	Connecticut
10	Delaware	DE	Dover	Wilmington	Dec 7, 1787	961,939	2,489	Delaware

DATA PREPARATION - Basic Transform Data

Add Column Column



Analyzing_Sales_Data - Power Query Editor

File Home Transform Add Column View Tools Help

Column From Examples Custom Invoke Custom Function General

Conditional Column Index Column Duplicate Column Format From Text Statistics Standard Scientific Trigonometry From Number Rounding From Date & Time Date Time Duration

From Text Parse From Date & Time Text Analytics Vision Azure Machine Learning AI Insights

Queries [2]

Products Orders

	ABC 123 ProductID	A ^B C ProductName	A ^B C QuantityPerUnit	t ² 3 UnitsInStock
1	1 Chai	10 boxes x 20 bags	35	
2	2 Chang	24 - 12 oz bottles	17	
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24		Guaraná Fantástica	12 - 355 ml cans	
25				

Custom Column

Add a column that is computed from the other columns.

New column name: Custom

Custom column formula: = [ProductID][ProductName]

Available columns:

- ProductID
- ProductName
- QuantityPerUnit
- UnitsInStock

<< Insert OK Cancel

No syntax errors have been detected.

4 COLUMNS, 77 ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED ON THURSDAY, DECEMBER 12, 2019



Power Query M Formula Language

- It is a functional, case-sensitive language
- Used for performing data mashup
- Performs major operations such as importing data, filtering data, and combining data from varied sources
- A Power Query M formula language query consists of formula expression steps that creates a mashup query

DATA PREPARATION - Basic Transform Data

M-code



Untitled - Query Editor

Transform Add Column View

Monospaced Always allow Advanced Editor Query Dependencies

Show whitespace

Data Preview Parameters Advanced Dependencies

```
= Table.RemoveColumns(Source, {"CustomerKey"})
```

	ProductKey	OrderDate	SalesTerritoryKey	OrderQuantity	1.2
1	592	2004-05-06	9	1	
2	592	2004-05-06	9	1	
3	405	2004-05-06	9	1	
4	479	2004-04-06	9	1	
5	482	2004-04-06	9	1	
6	595	2004-06-06	9	1	
7	489	2004-07-06	9	1	
8	491	2004-08-06	9	1	

The Power Query Formula Language 'M' is a "Functional" Language

Untitled - Query Editor

Transform Add Column View

Monospaced Always allow Advanced Editor Query Dependencies

Show whitespace

Data Preview Parameters Advanced Dependencies

```
= Table.RemoveColumns(Source, {"CustomerKey"})
```

	ProductKey	OrderDate	SalesTerritoryKey	OrderQuantity	1.2
1	592	2004-05-06	9	1	
2	592	2004-05-06	9	1	
3	405	2004-05-06	9	1	
4	479	2004-04-06	9	1	
5	482	2004-04-06	9	1	
6	595	2004-06-06	9	1	
7	489	2004-07-06	9	1	
8	491	2004-08-06	9	1	

'M' Language is case sensitive

DATA PREPARATION - Basic Transform Data

Writing Power Query Functions



Power Query M Functions (Continued)

Advanced Editor

Orders

Display Options ?

```
let
    Source = Excel.Workbook(File.Contents("D:\Packt Data\Sample - Superstore.xlsx"), null, true),
    Orders_Sheet = Source{[Item="Orders",Kind="Sheet"]}[Data],
    #"Promoted Headers" = Table.PromoteHeaders(Orders_Sheet, [PromoteAllScalars=true]),
    #"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"Order Date", type date}, {"Ship Date", type date},
    {"Sales", type number}, {"City", type text}, {"Quantity", type number}, {"Discount", type number}})
in
    #"Changed Type"
```

Accessing data function

Table function

✓ No syntax errors have been detected.

Done Cancel

DATA PREPARATION - Basic Transform Data

Data profiling



Screenshot of the Power Query Editor showing data profiling features:

- View tab:** The "Column distribution" checkbox is checked (1).
- Layout tab:** The "Column quality" checkbox is checked (2).
- Data Preview pane:** The "CustomerKey" column is selected. A red box highlights the distribution statistics: 1000 distinct, 1000 unique. Other columns show similar distributions.
- Bottom pane:** A context menu is open over the "Date" column, showing options like Copy, Keep Duplicates, Remove Duplicates, etc. (3).

Screenshot of the Options dialog box:

- Preview features:** The "Enable column profiling" checkbox is checked (4).
- GLOBAL section:** Includes Data Load, Power Query Editor, DirectQuery, R scripting, Python scripting, Security, Privacy, Updates, Usage Data, Diagnostics, and Auto recovery.
- CURRENT FILE section:** Includes Data Load, Regional Settings, Privacy, Auto recovery, Query reduction, and Report settings.
- Buttons:** OK (6) and Cancel.

DATA PREPARATION - Basic Transform Data

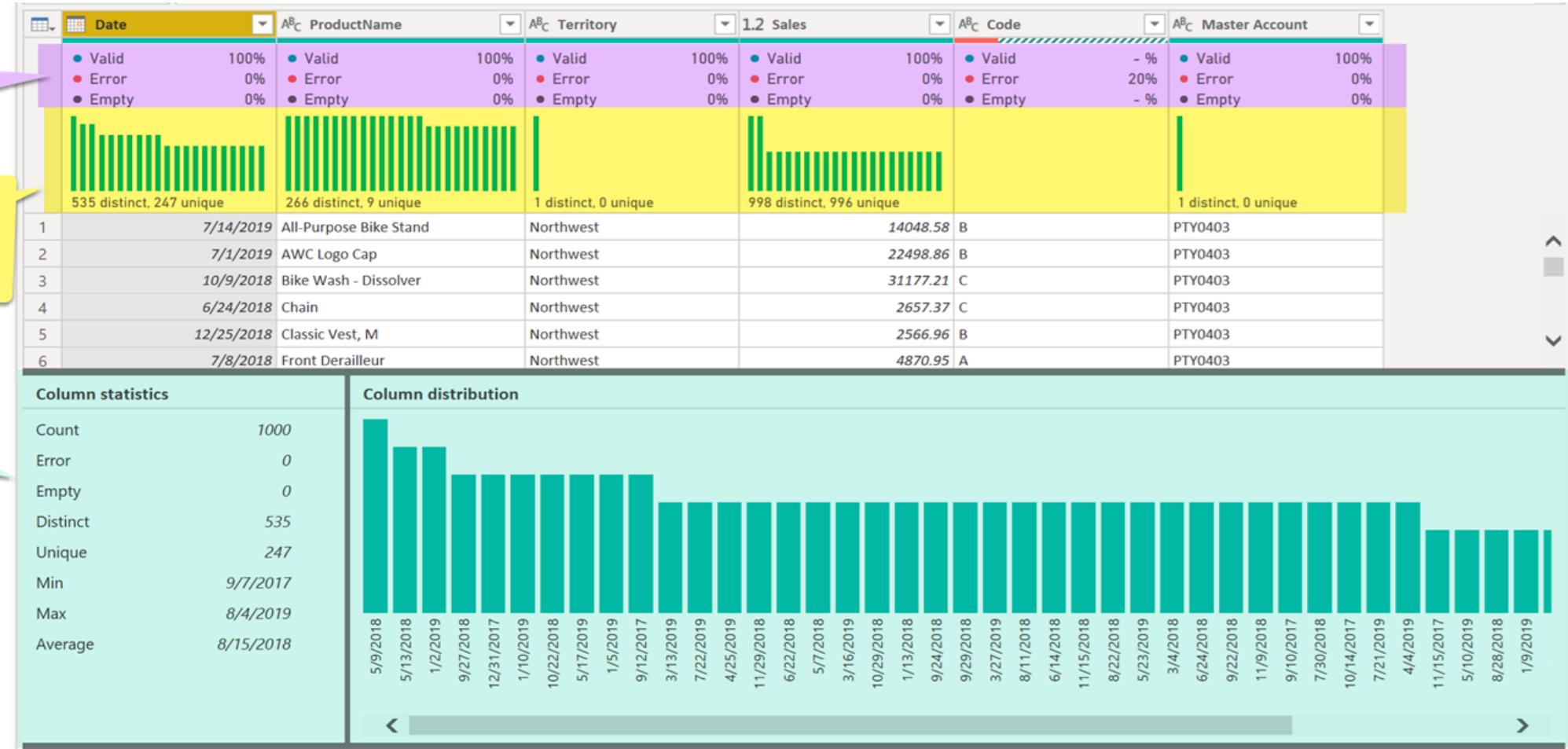
Data profiling



Column quality

Column distribution

Column profile



DATA PREPARATION - Basic Transform Data

Query Dependencies



Contoso Final Report - Power Query Editor

File Home Transform Add Column View Tools Help

Query Settings Layout Data Preview

Queries [18]

- Lookup Tables [6]
 - DIM Product M...
 - DIM Geography
 - DIM Calendar
 - DAX Measures
 - DIM Store
 - DIM Channel
- Data Tables [1]
 - FACT Sales
- Staging Tables [3]
 - DIM Product
 - DIM Product Sub...
 - DIM Product Cat...
- Data Sources [3]
- Transform File fr...
- Sample Query [2]
 - Sample File P...
 - Sample File
 - Transform Sam...
 - Transform File f...
- Calendar Dates [3]
 - Calendar Start...
 - Calendar End D...
 - Fiscal Year End...
- Contoso Data So...

An error occurred in the 'C' Details: C:\Contoso Data Model\

Query Dependencies

The diagram illustrates the dependencies between different queries. It shows a complex network of connections between various data sources (e.g., DIM Product, DIM Geography, Fact Sales) and transformation steps (e.g., Union All, Sort, Calculate). The connections indicate the flow of data from the source to the final destination in the Contoso Data Model.

text},

Query Settings

PROPERTIES

- Name: DIM Product

APPLIED STEPS

- Source
- Navigation
- Changed Type

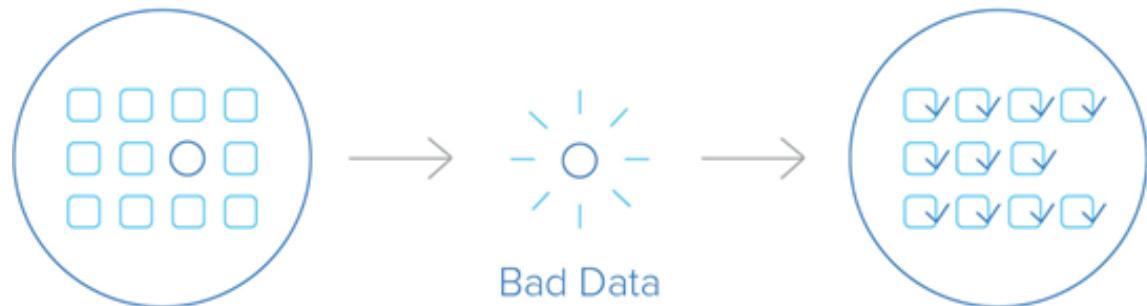
Layout Close

DATA PREPARATION - Basic Transform Data

Cleaning Data



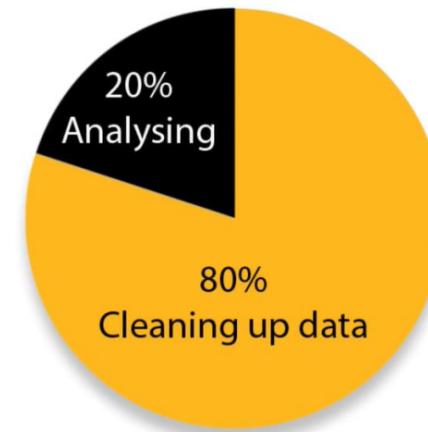
Data Cleansing



Data Cleaning made simple. Quickly and easily remove data that may distort your analysis.



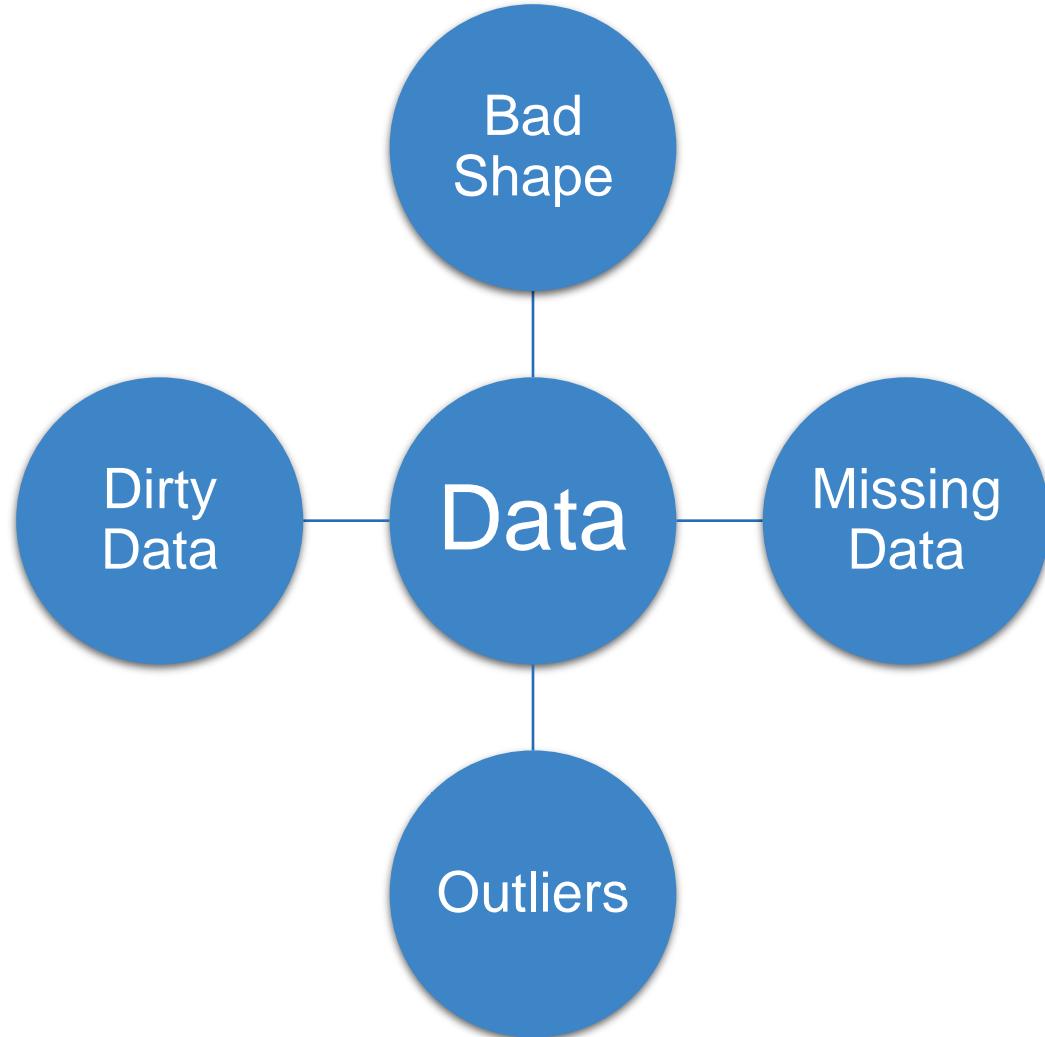
HOW DATA SCIENTISTS SPEND TIME



Data cleaning:
Making Data reliable and ready to efficiently be used in applications



DATA PREPARATION – Data Issue



DATA PREPARATION

Understand Structure of Data



Category	Type	2001	2002	2003	Grand total
Fruit	Apples	150	153	162	465
	Bananas	332	336	344	1012
	Pears	267	266	279	812
	Subtotal	749	755	785	2289
Vegetables	Cucumber	140	141	152	433
	Lettuce	246	245	258	749
	Tomatoes	156	161	168	485
	Subtotal	542	547	578	1667
Grand total		1291	1302	1363	3956



Year	Month	Category	Type	Sales
2001	January	Fruit	Apples	12
2001	January	Fruit	Pears	21
2001	January	Fruit	Bananas	29
2001	January	Vegetables	Cucumber	9
2001	January	Vegetables	Tomatoes	13
2001	January	Vegetables	Lettuce	22
2001	February	Fruit	Apples	11
2001	February	Fruit	Pears	21
2001	February	Fruit	Bananas	31
2001	February	Vegetables	Cucumber	8
2001	February	Vegetables	Tomatoes	12
2001	February	Vegetables	Lettuce	20
2001	March	Fruit	Apples	9
2001	March	Fruit	Pears	19
2001	March	Fruit	Bananas	32
2001	March	Vegetables	Cucumber	8
2001	March	Vegetables	Tomatoes	11
2001	March	Vegetables	Lettuce	21
2001	April	Fruit	Apples	9
2001	April	Fruit	Pears	18
2001	April	Fruit	Bananas	32
2001	April	Vegetables	Cucumber	10
2001	April	Vegetables	Tomatoes	12
2001	April	Vegetables	Lettuce	21
2001	May	Fruit	Apples	10
2001	May	Fruit	Pears	20

Cross Table /Contingency Table / Pivot Table

A **cross table** is a two-way table (matrix) consisting of columns and rows. Also known as a pivot table or a multi-dimensional table.

Regular Table

DATA PREPARATION

Understand Structure of Data



Figure 3.5

Stacked Data

	A	B
1	Gender	Salary
2	Male	81600
3	Female	61600
4	Female	64300
5	Female	71900
6	Male	76300
7	Female	68200
8	Male	60900
9	Female	78600
10	Female	81700
11	Male	60200
12	Female	69200
13	Male	59000
14	Male	68600
15	Male	51900
16	Female	64100
17	Male	67600
18	Female	81100
19	Female	77000
20	Female	58800
21	Female	87800
22	Male	78900

Figure 3.6

Unstacked Data

	A	B
1	Female Salary	Male Salary
2		81600
3	64300	76300
4	71900	60900
5	68200	60200
6	78600	59000
7	81700	68600
8	69200	51900
9	64100	67600
10	81100	78900
11	77000	
12	58800	
13	87800	

DATA PREPARATION

Understand Structure of Data



Team	KPI 1	KPI 2	KPI 3
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31



Team	KPI 1	KPI 2	KPI 3
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31

Wide format

Narrow format

Team	KPI 1	KPI 2	KPI 3
A	KPI 1	88	
A	KPI 2	12	
A	KPI 3	22	
B	KPI 1	91	
B	KPI 2	17	
B	KPI 3	28	
C	KPI 1	99	
C	KPI 2	24	
C	KPI 3	30	
D	KPI 1	94	
D	KPI 2	28	
D	KPI 3	31	

Team	KPI Name	Value
A	KPI 1	88
A	KPI 2	12
A	KPI 3	22
B	KPI 1	91
B	KPI 2	17
B	KPI 3	28
C	KPI 1	99
C	KPI 2	24
C	KPI 3	30
D	KPI 1	94
D	KPI 2	28
D	KPI 3	31

DATA PREPARATION – Data Issue

Data Shape Formatting



- How to identify **when your data needs to be formatted.**
- How to **transform data into the correct format**
- How to **aggregate it to the form required**

C:\Users\Admins\Desktop\MDA\K38\Day 6\Exercise Create Data Model
from Multiple Flat Files

1. **Transpose Table**
2. **Cross Tabulation**
Pivot + Unpivot
3. **Aggregation (Group by)**

DATA PREPARATION – Data Issue

Data Shape Formatting - Transpose Table

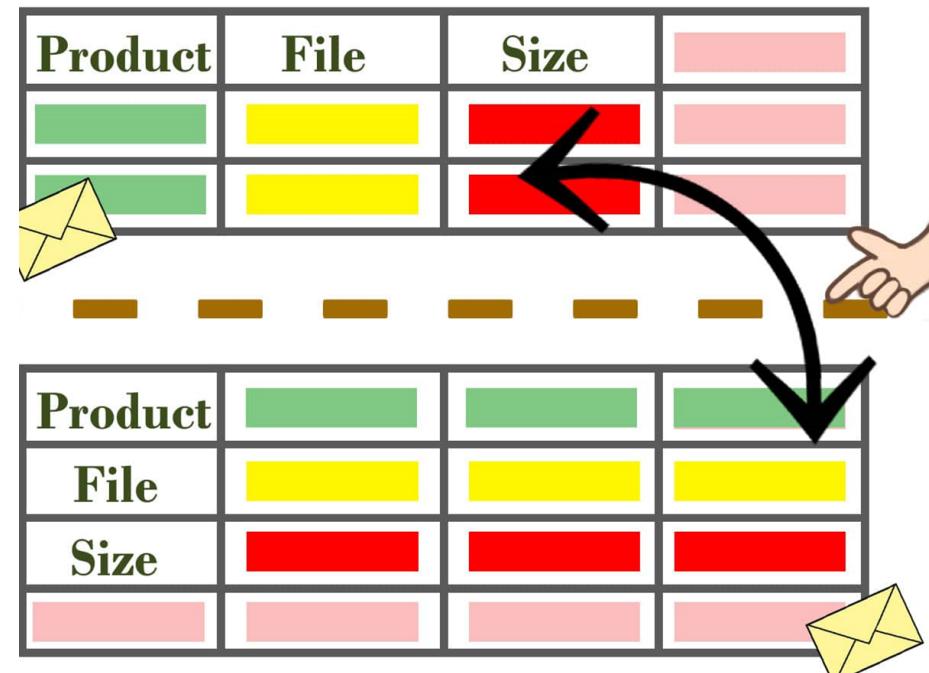


TRANSPOSE TABLE

Region	East	East	East	East	MidWest	MidWest	MidWest	MidWest	South	South	South	South	West	West	West	West
Area	East 1	East 2	East 3	East 4	MidWest 1	MidWest 2	MidWest 3	MidWest 4	South 1	South 2	South 3	South 4	West 1	West 2	West 3	West 4
Aspen	10				74				14				67			
Beaut		16			30					51				20		
Bellen		77			47						61					51
Carlota			39		57				89							56
Doublers			7				16			3				42		
FlatTop				4				34				5				8
Quad			87		112						102		77			
Sunbell				44		47					20		38			
Sunset				23				18		25				83		
Sunshine	76			60							52		49			
V-Rang		8				12		13								8
Yanaki		42						26		20						77



Region	Area	Aspen	Beaut	Bellen	Carlota	Doublers	FlatTop	Quad	Sunbell	Sunset	Sunshine	V-Rang	Yanaki		
East	East 1	10									76				
East	East 2		16	77			7						8	42	
East	East 3				39				87						
East	East 4						4		44	23					
MidWest	MidWest 1		30	47	57				112		60				
MidWest	MidWest 2	74							47						
MidWest	MidWest 3					16						12			
MidWest	MidWest 4						34			18			26		
South	South 1					89						13			
South	South 2	14	51				3			25					
South	South 3			61								20			
South	South 4						5	102	20		52				
West	West 1	67													
West	West 2					42		77	38	83	49				
West	West 3		20		56									77	
West	West 4			51			8					8			



DATA PREPARATION – Data Issue

Data Shape Formatting - Cross Tabulation



UNPIVOT & PIVOT

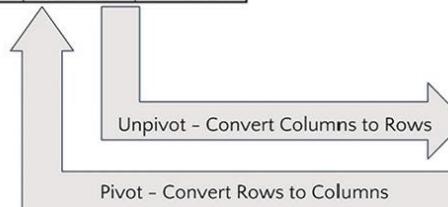
Preparing Data: Pivot and Unpivot

"Wide" Format - aka "Unstacked" Format:

Country	2016	2017	2018
China	1.379B	1.386B	1.393B
India	1.325B	1.339B	1.353B
United States	0.323B	0.325B	0.327B

"Tall" Format - aka "Tidy" data:

Country	Year	Population
China	2016	1.379B
China	2017	1.386B
China	2018	1.393B
India	2016	1.325B
India	2017	1.339B
India	2018	1.353B
United States	2016	0.323B
United States	2017	0.325B
United States	2018	0.327B



ID	ProductName	ProductCost	Wholesale	Retail
1	Widget	10.53	15.25	19.51
2	Thing-a-majig	8.85	12.31	15.41

UNPIVOT

ID	ProductName	PurchaseType	Cost
1	Widget	ProductCost	10.53
1	Widget	Wholesale	15.25
1	Widget	Retail	19.51
2	Thing-a-majig	ProductCost	8.85
2	Thing-a-majig	Wholesale	12.31
2	Thing-a-majig	Retail	15.41

DATA PREPARATION – Data Issue

Data Shape Formatting - Cross Tabulation



Cross Tabulation Pivot

	Product	Category	January	February	March	April	May
1	Graphing Calculators	General	191817	434	70654	166571	99066
2	Office Supplies	General	156628	82183	125043	11205	13089
3	Encyclopedias	Educational	6299	119153	161717	145195	22305
4	Building Blocks	Fun and Games	8313	184270	186021	190255	8211
5	Books about Dinosaurs	Fun and Games	193667	76441	163244	116158	42346
6	Viggo Mortenson DVDs	Fun and Games	181291	178860	144830	35341	13633
7	Clothing	General	6988	71830	69881	137635	16849
8	Frisbee and Frisbee Accessories	Fun and Games	4940	105607	53949	72655	14085
9	Legumes	General	73467	154459	92995	68466	15942
10	Microscopes	Educational	206	324	925	633	487

Cross Tabulation Unpivot

Record	Product	Category	Name	Value
1	Graphing Calculators	General	January	191817
2	Graphing Calculators	General	February	434
3	Graphing Calculators	General	March	70654
4	Graphing Calculators	General	April	166571
5	Graphing Calculators	General	May	99066
6	Graphing Calculators	General	June	64423
7	Graphing Calculators	General	July	72846
8	Graphing Calculators	General	August	52744
9	Graphing Calculators	General	September	16150
10	Graphing Calculators	General	October	98130
11	Graphing Calculators	General	November	42312
12	Graphing Calculators	General	December	133005

DATA PREPARATION – Data Issue

Data Aggregation



Data Aggregation Aggregate by Month

Date	Sales
1/1/2015	89.02
1/2/2015	2257.54
1/3/2015	697.87
1/4/2015	2282.91
1/5/2015	2013.72
1/6/2015	878.43
1/7/2015	542.25
1/8/2015	1362.44



Month	Sales
1	42432.76
2	39165.11
3	43110.24
4	39684.96
5	31870.93
6	30544.65
7	37934.34

DATA PREPARATION – Data Issue

Data Aggregation



Product	Category	Name	Value	Quarter
Graphing Calculators	General	April	166571	Q2
Office Supplies	General	April	11205	Q2
Encyclopedias	Educational	April	145195	Q2
Building Blocks	Fun and Games	April	190255	Q2
Books about Dinosaurs	Fun and Games	April	116158	Q2
Viggo Mortenson DVDs	Fun and Games	April	35341	Q2
Clothing	General	April	137635	Q2
Frisbee and Frisbee Accessories	Fun and Games	April	72655	Q2
Legumes	General	April	68466	Q2
Microscopes	Educational	April	633	Q2
Graphing Calculators	General	August	52744	Q3
Office Supplies	General	August	121453	Q3
Encyclopedias	Educational	August	157398	Q3
Building Blocks	Fun and Games	August	4735	Q3
Books about Dinosaurs	Fun and Games	August	72374	Q3
Viggo Mortenson DVDs	Fun and Games	August	42458	Q3

Data Aggregation Aggregate by Quarter

Quarter	Sum_Value
Q1	2866436
Q2	2805889
Q3	2999634
Q4	2110647

DATA PREPARATION – Data Issue

Dirty Data



Not Parsed Correctly

Name	LastName	FirstName
Smith, John	Smith	John

Extra Characters

Name
"John Smith"

Unexpected Pattern

Email
john.doe@google.com
jmiller@hotmail
schow@yahoo.com

Dirty Data contains some kind of errors in them, or in a format that's unfriendly or unusable

Misspelled Entries

125 Main ~~Str~~reet
↓
125 Main Street

Duplicate Data Records

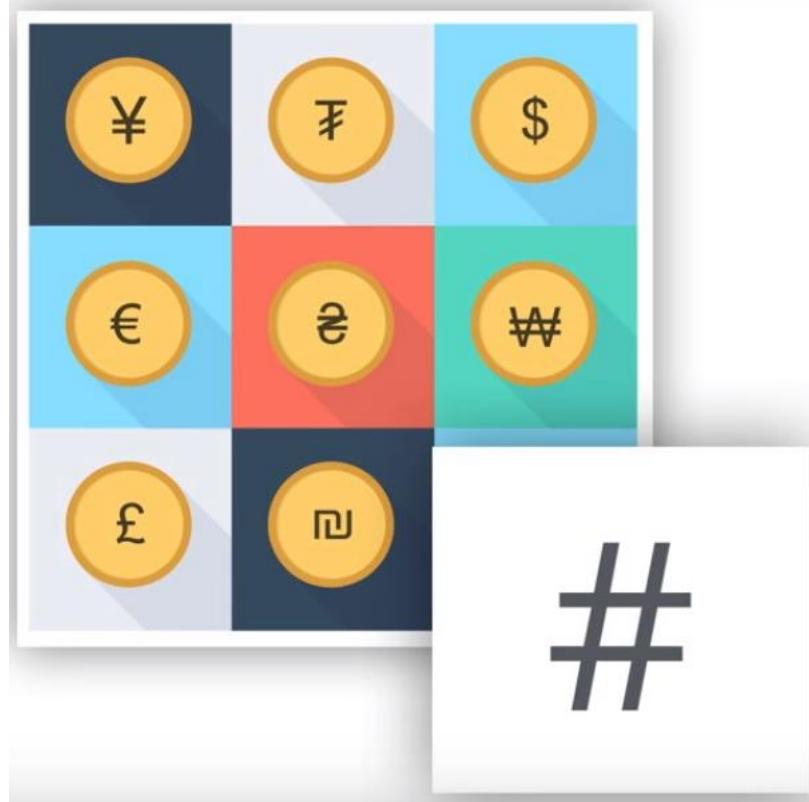
ID	Name
1	John Smith
1	John Smith
2	Jane Doe

Incorrect Data

Date	Sales
2016-05-01	1000
1900-01-01	500
2016-04-28	830
	51

DATA PREPARATION – Data Issue

Dirty Data



Dirty Data

Extra Characters

- Convert Strings to Numbers
- Fields with strange symbols

Extra characters can be currency symbols, number signs... We'd need to remove these before changing between field types

DATA PREPARATION – Data Issue

Dirty Data



Example 1

“Maureen”

“Patrick”

“Ben”

Example 2

19143L

Example 3

\$432

DATA PREPARATION – Data Issue

Dirty Data



No:

Addresses
313 173rd Blvd, Kent, WA 981215
316 66th Blvd, Kent, WA 981244
4358 23rd St, Kent, WA 981225
965 151st St, Kent, WA 981162
7900 173rd Lane, Kent, WA 981266
4047 15th Ave, Kent, WA 981228
4907 13th Ave, Kent, WA 981232
3789 4th Blvd, Seattle, WA 981152
2977 66th Lane, Seattle, WA 981171
3392 23rd St, Seattle, WA 981131

Yes:

Address	City	State	Zip
313 173rd Blvd	Kent	WA	981215
316 66th Blvd	Kent	WA	981244
4358 23rd St	Kent	WA	981225
965 151st St	Kent	WA	981162
7900 173rd Lane	Kent	WA	981266
4047 15th Ave	Kent	WA	981228
4907 13th Ave	Kent	WA	981232
3789 4th Blvd	Seattle	WA	981152
2977 66th Lane	Seattle	WA	981171
3392 23rd St	Seattle	WA	981131

DATA PREPARATION – Data Issue

Missing Data



Missing data: gaps in data

Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3		44000	HS	N
4	55	78000	MA	Y
5	23		HS	N
6	25	42000		N
7	35	121000	PhD	Y
8	51	45000	BA	
9			MS	N
10	67	54000	MA	Y

Missing
Data 

Blank/ Empty cells (CSV)

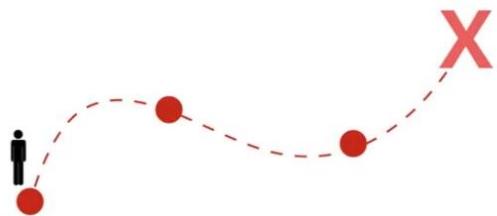
Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3	(Null)	44000	HS	N
4	55	78000	MA	Y
5	23	(Null)	HS	N
6	25	42000	(Null)	N
7	35	121000	PhD	Y
8	51	45000	BA	(Null)
9	(Null)	(Null)	MS	N
10	67	54000	MA	Y

Missing
Data 

Null value (Database)

Reasons

Some statistical
algorithms won't work



Reasons

Can add **BIAS** to a
model

Value

Estimate

Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3	N/A	44000	HS	N
4	55	78000	MA	Y
5	23	N/A	HS	N
6	25	42000	N/A	N
7	35	121000	PhD	Y
8	51	45000	BA	N/A
9	N/A	N/A	MS	N
10	67	54000	MA	Y

Missing
Data 

N/A (program)

BIAS in statistics refers to the tendency of an analysis to either **over** or **underestimate** the values of that specific field or parameter

Why care about
Missing
Data?

DATA PREPARATION – Data Issue

Missing Data



Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3	70	44000	HS	N
4	55	78000	MA	Y
5	23		HS	N
6	25	42000		N
7	35	121000	PhD	Y
8	51	45000	BA	
9	65		MS	N
10	67	54000	MA	Y

Average age without missing values:
44.5

Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3		44000	HS	N
4	55	78000	MA	Y
5	23		HS	N
6	25	42000		N
7	35	121000	PhD	Y
8	51	45000	BA	
9			MS	N
10	67	54000	MA	Y

Average age without missing values:
44.5

Average age with missing values:
40

Real Data

Downward BIAS

DATA PREPARATION – Data Issue

Missing Data



SOLUTIONS

1. Deleting Missing Data
2. Imputation



DATA PREPARATION – Data Issue

Missing Data



Deleting Missing Data

Deleting missing data is often the **default method** because it's **simplicity**. No decisions that need to be made that might confuse the data. You just get rid of records where there are missing values.

However, you should **make sure** that deleting missing data **doesn't have adverse effects on your analysis**. For example, if a particular demographic tended to leave a response blank in a survey, then removing records with blank entries will mean that a part of the population is underrepresented.

One of the downsides is that eliminating missing data reduces the size of the dataset (Ex: cost).

Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3		44000	HS	N
4	55	78000	MA	Y
5	23		HS	N
6	25	42000		N
7	35	121000	PhD	Y
8	51	45000	BA	
9			HS	N
10	67	54000	MA	Y



Respondent ID	Age	Income	Education	Homeowner
1	27	36000	HS	N
2	37	64000	BA	Y
4	55	78000	MA	Y
7	35	121000	PhD	Y
10	67	54000	MA	Y

DATA PREPARATION – Data Issue

Missing Data

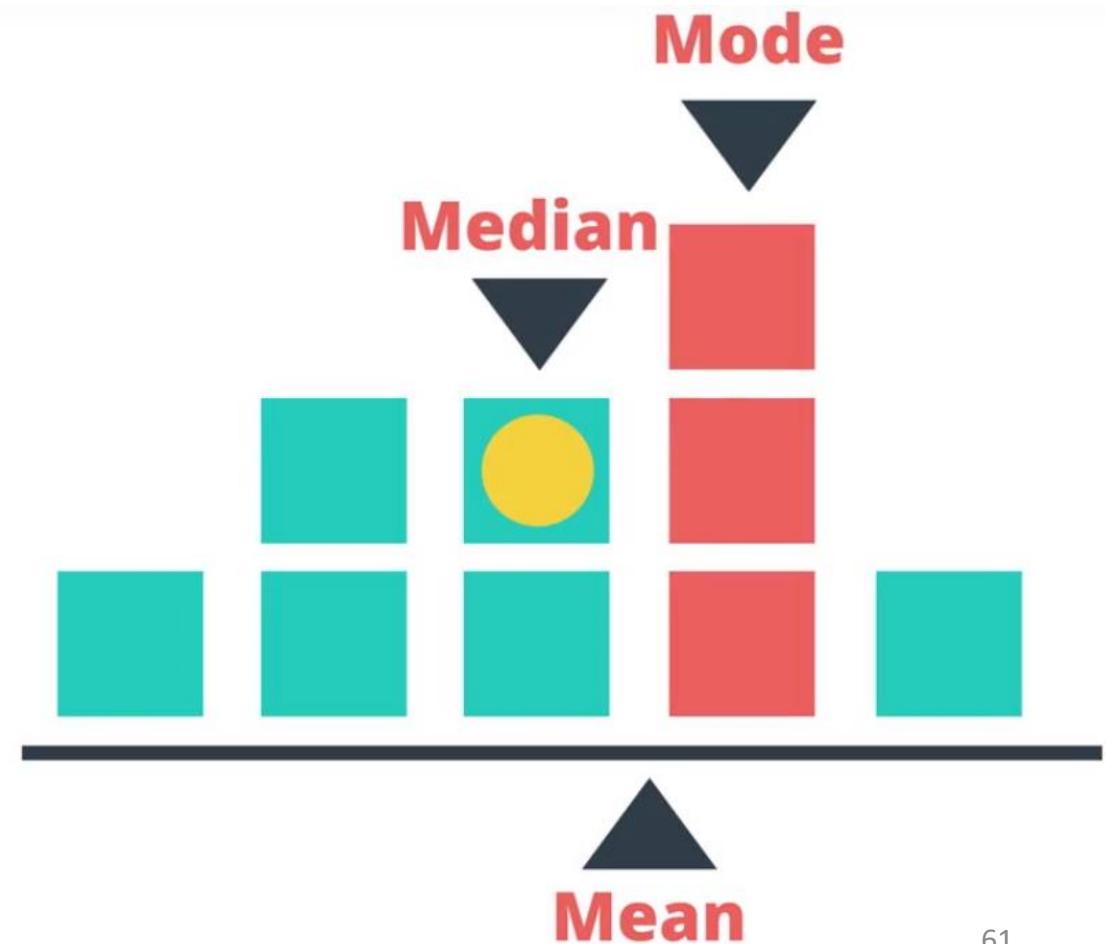


Imputation

In statistics, Imputation is the process of substituting values in the data where the value are missing (we impute values, we are making them up). We are **creating fake data** in order to develop a model that makes sense and is as close to reality as we can get it

**Replace
Missing
Values**

**Mean
Median
Mode**



DATA PREPARATION – Data Issue

Missing Data – Select the method



What methodology might be the best approach

1. **How much data is really missing? (>=20%)**
2. **How the missing data is distributed across the dataset? (2/10 predictor variables missed)**
3. **Whether those specific variables are actually significant to our analysis and model making process**
4. **The missing data is numeric or categorical**

Numeric or Categorical ?

Are they **SIGNIFICANT?**

A	B	C	D	E	F	G	H	I	J
Missing	Missing								
		Missing							
Missing									

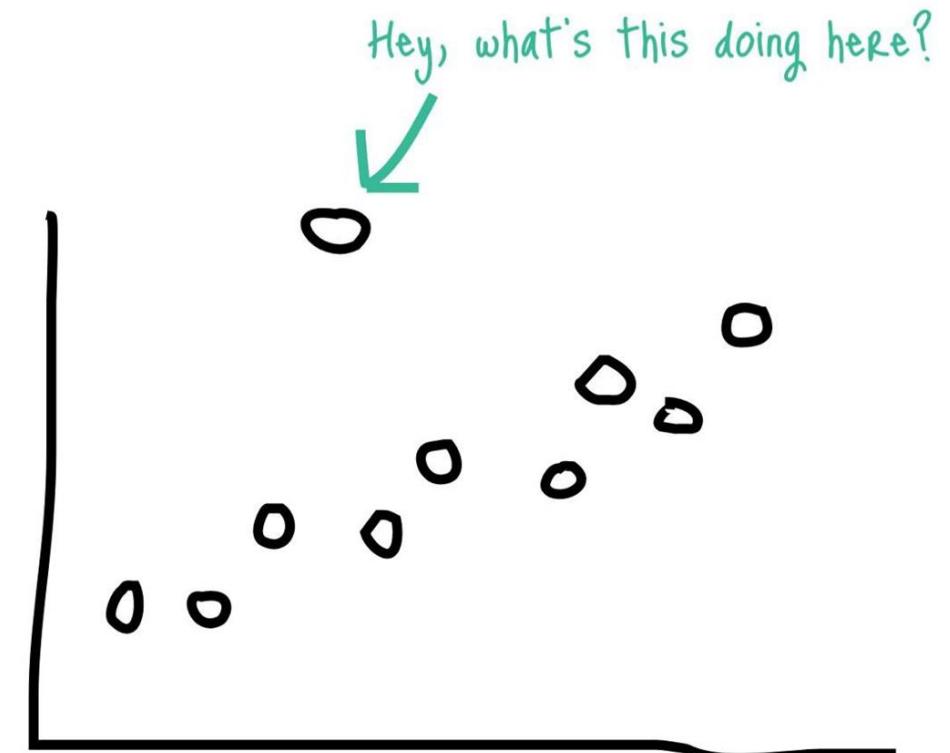
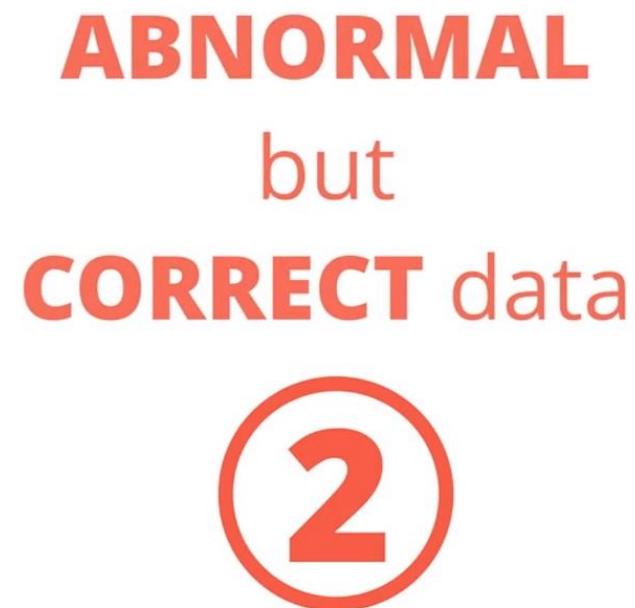
Respondent ID	Age	Income	Education
1	27	32000	HS
2	37	64000	BA
3		44000	HS
4	55	78000	MA
5	23		HS
6	25	42000	
7	35	121000	PhD
8	51	45000	BA
9			MS
10	67	54000	MA

DATA PREPARATION – Data Issue

Outliers



Identifying outliers in the data helps us understand **how vulnerable** our model would be to a **small set of observations**.



DATA PREPARATION – Data Issue

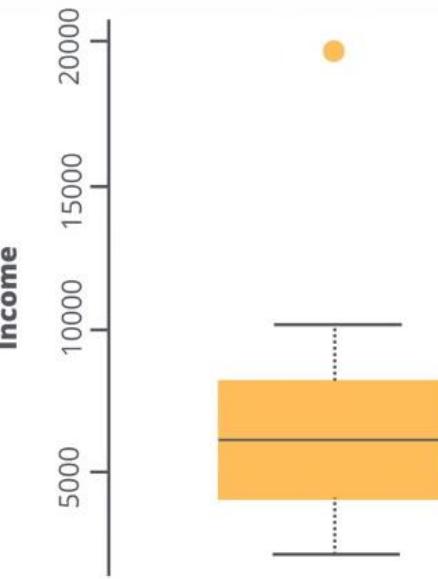
Outliers



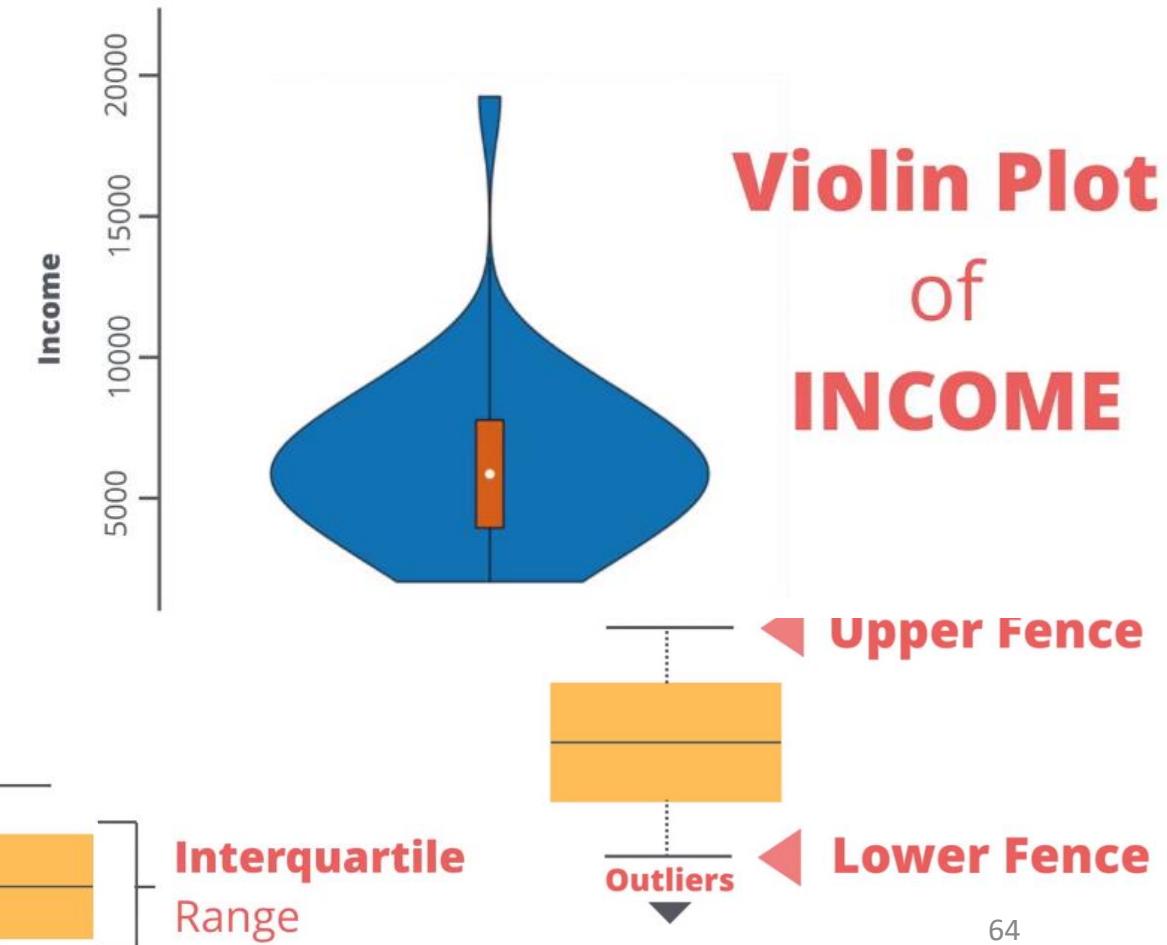
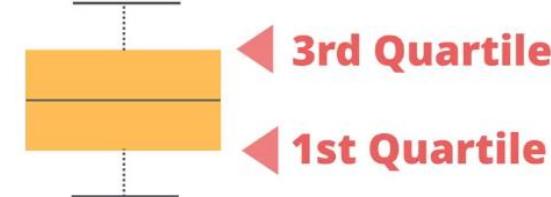
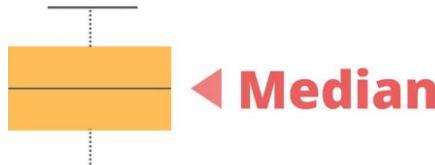
Identifying outliers more **methodically** rather than simply eyeballing them

Violin Plot: shows the volume of the distribution

Others: z-scores or standard deviations



**BOX
and
WHISKER
plot of
INCOME**



DATA PREPARATION – Data Issue

Outliers

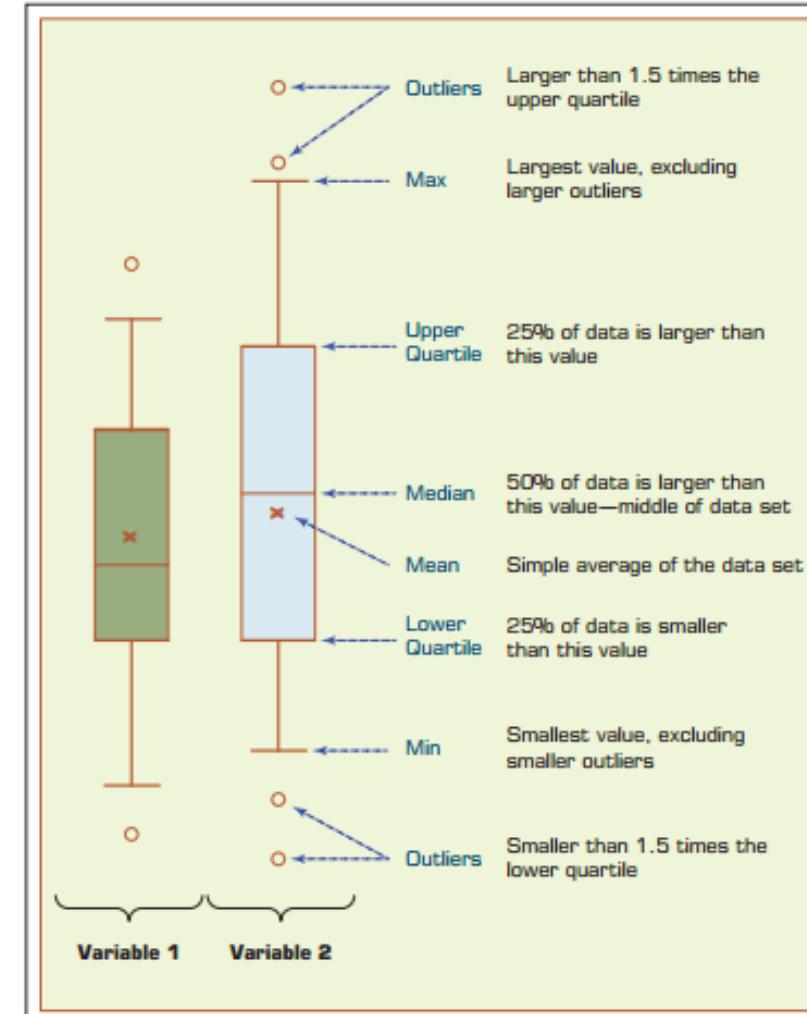
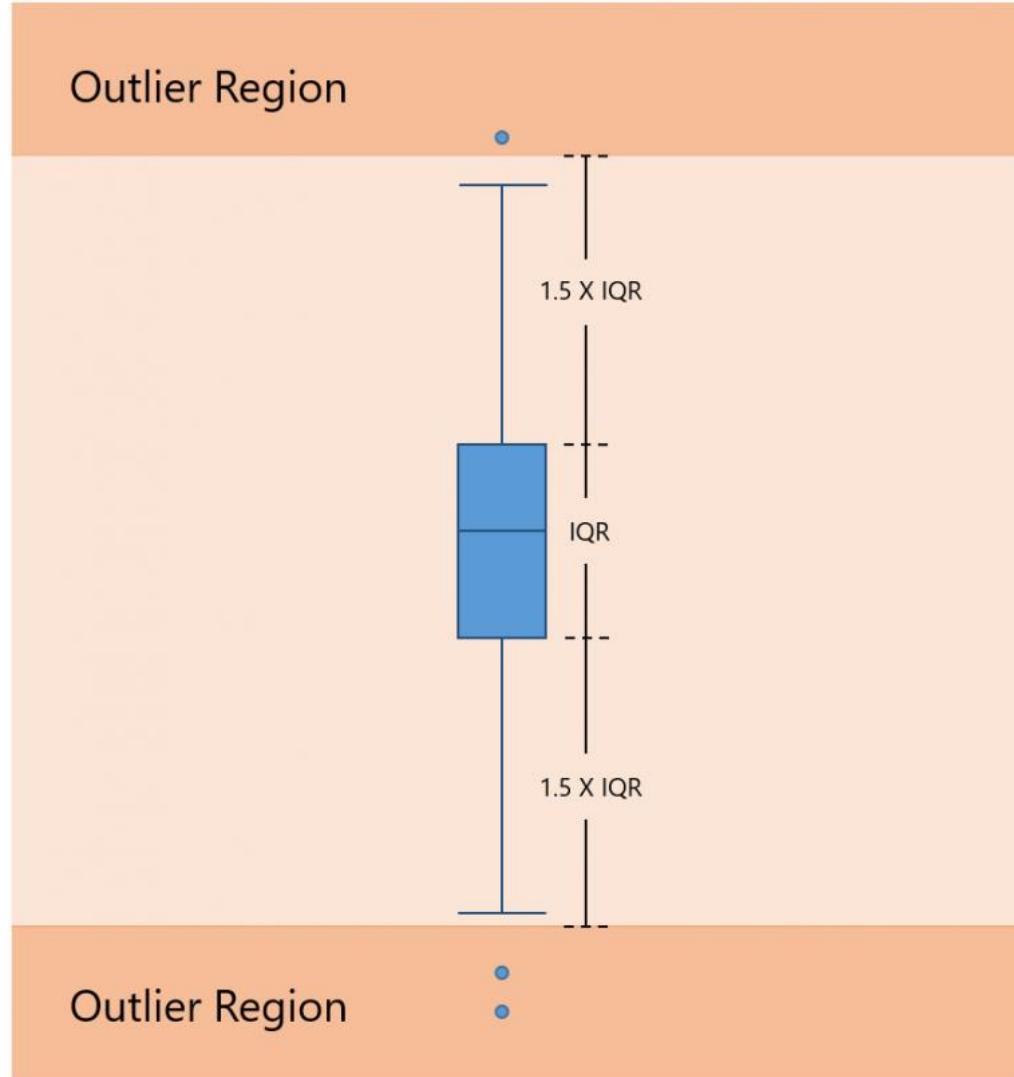


FIGURE 2.8 Understanding the Specifics about Box-and-Whiskers Plots.

DATA PREPARATION – Data Issue

Outliers



Steps

To calculate the upper fence and the lower fence, here are the exact steps:

- 1 . Calculate 1st quartile Q_1 and 3rd quartile Q_3 of the dataset. You can use the Excel function `QUARTILE.INC` or `QUARTILE.EXC`
- 2 . Calculate the Interquartile Range: $IQR = Q_3 - Q_1$
- 3 . Add $1.5 * IQR$ to Q_3 to get the upper fence: $\text{Upper Fence} = Q_3 + 1.5 * IQR$
- 4 . Subtract $1.5 * IQR$ to Q_1 to get the lower fence: $\text{Lower Fence} = Q_1 - 1.5 * IQR$

If a value is **1.5 times** the INTERQUARTILE RANGE of a data set, then it can be considered an OUTLIER

DATA PREPARATION – Data Issue

Outliers



1 & 2/ ERRORS

1. Try to go back to the original source to determine the correct data
2. Delete the record from the dataset



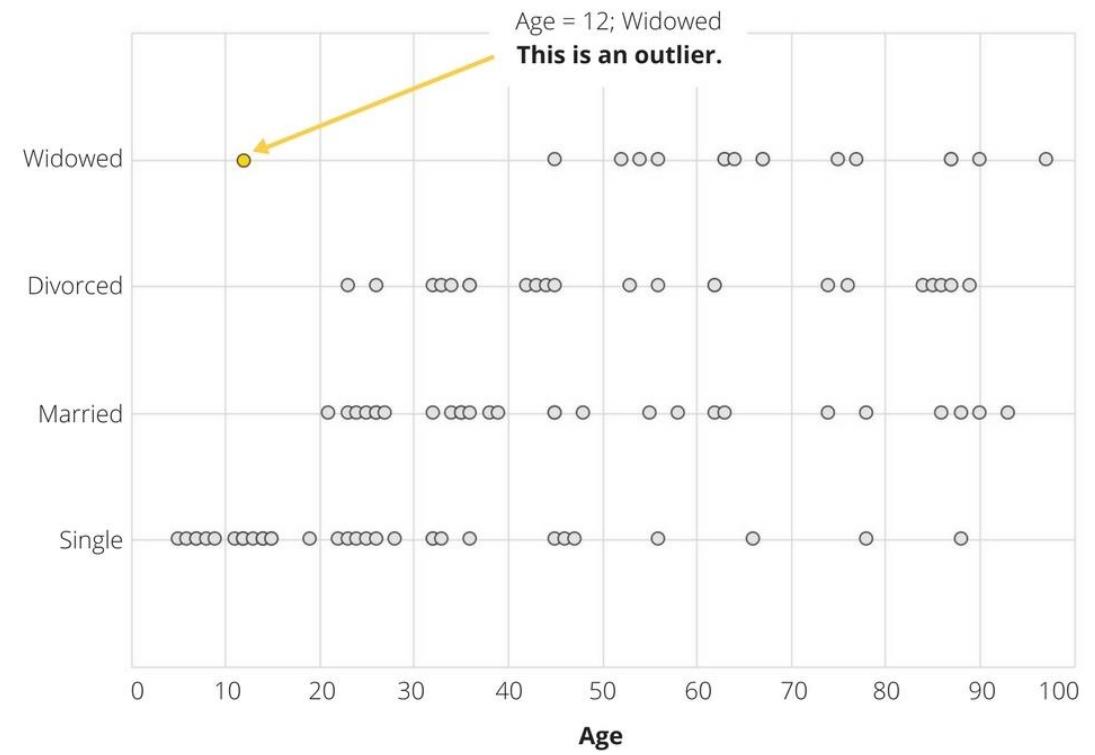
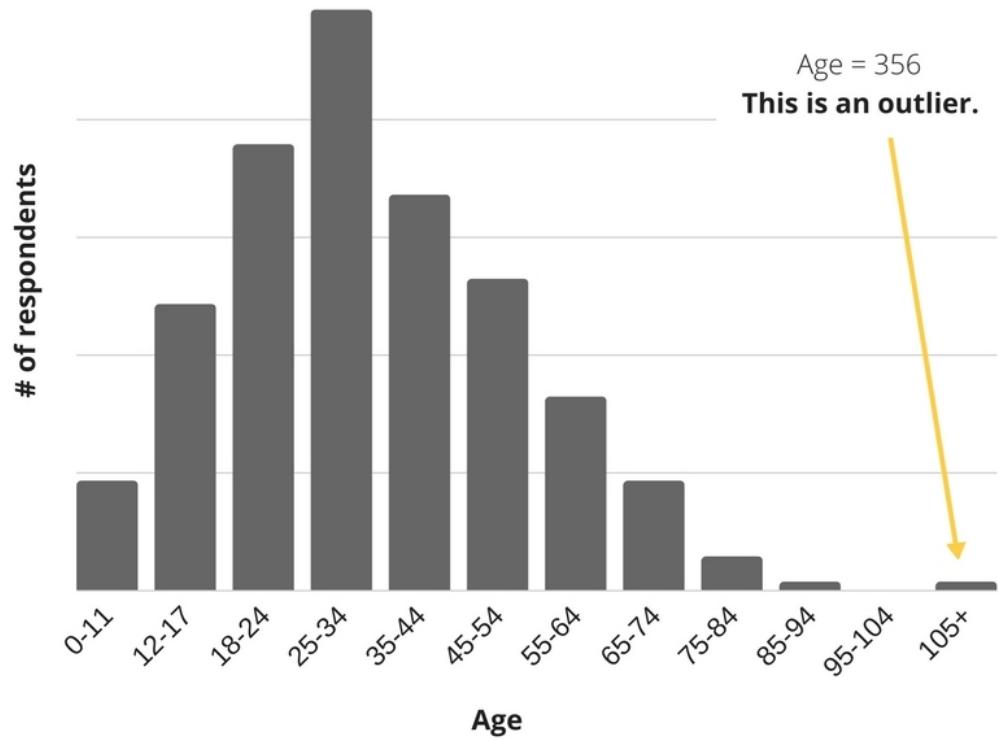
3. Don't have obvious errors, but we aren't certain whether the data is accurate or not

**REMOVE
Insignificant
Outliers**

Ex: Age: 299

DATA PREPARATION – Data Issue

Outliers

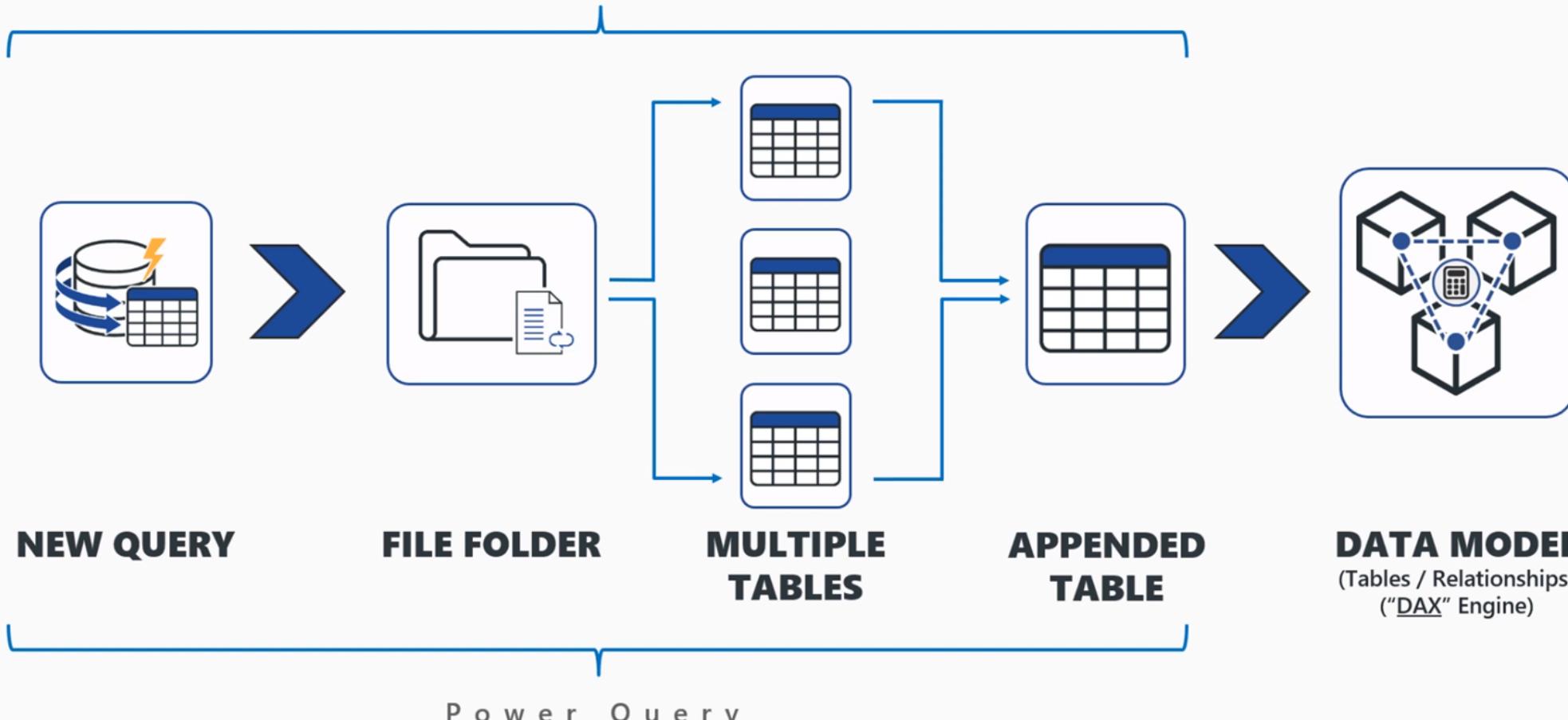


DATA PREPARATION – Data Issue

Combine Data from Folder



Combining Flat Files From A Folder



DATA PREPARATION - Data Blending



Data may come from different places, and as a results, it'll all need to be stitched together into one data file

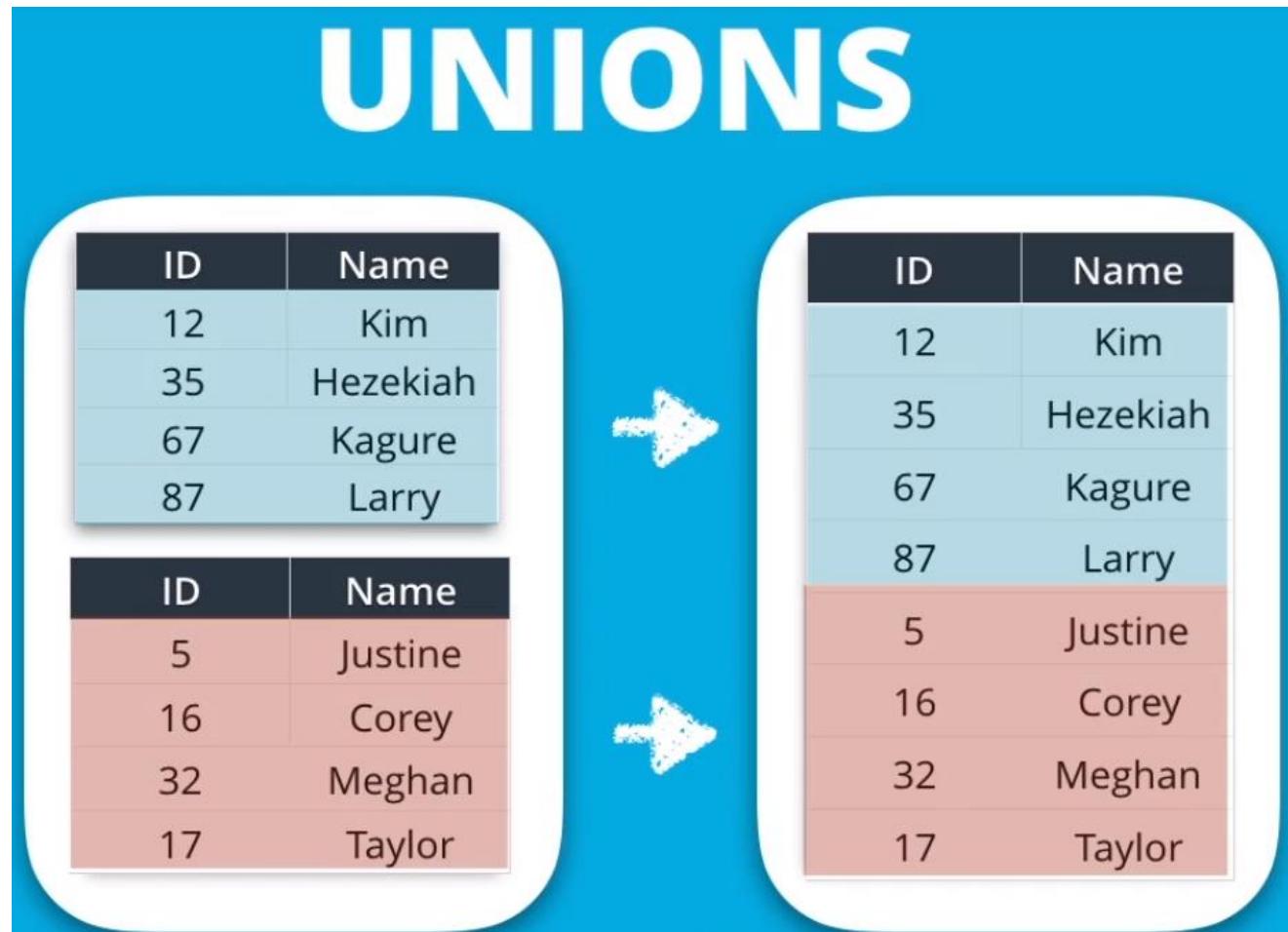
- ## Data Blending
- Unions
 - Joins
 - Fuzzy Matching
 - Spatial Matching

DATA PREPARATION - Data Blending

Unions / Append Queries



Union allows you to take multiple datasets and deal with them as one

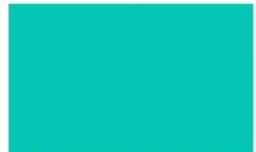


DATA PREPARATION - Data Blending

Join / Merge Query



Union



+

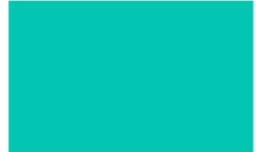


Table 1

ID	Transaction	Spend
5	5	500
16	10	750
32	1	200
17	2	50

?

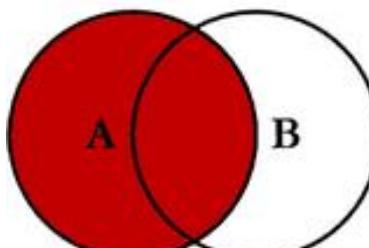


Table 2

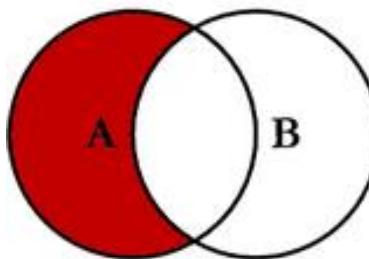
ID	Name	2010-01-15	2010-02-15
5	Mark Jones	TRUE	FALSE
16	Alan Young	FALSE	FALSE
32	Janet Chow	TRUE	FALSE
17	Roy Byron	TRUE	TRUE

Table 3

ID	Address	City	ST	ZIP	Distance
5	539 S PAYNE AVE	CASPER	WY	82509	5.5
16	706 E. LONGMONT	GILLETTE	WY	82716	10.7
32	1275 NORTH 11TH	LARAMIE	WY	82072	15.4
17	1203 RUSSELL	LARAMIE	WY	82070	16.5

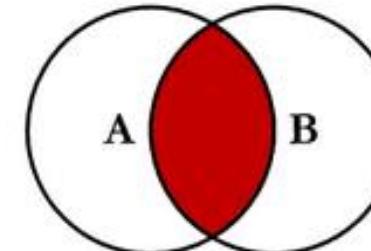


```
SELECT <select_list>  
FROM TableA A  
LEFT JOIN TableB B  
ON A.Key = B.Key
```

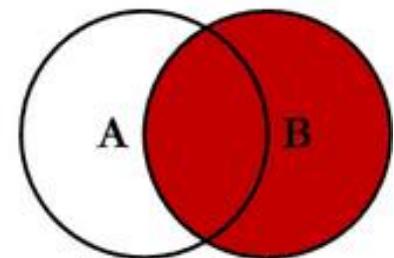


```
SELECT <select_list>  
FROM TableA A  
LEFT JOIN TableB B  
ON A.Key = B.Key  
WHERE B.Key IS NULL
```

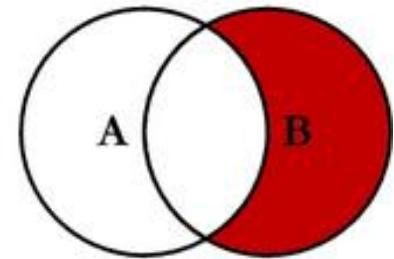
SQL JOINS



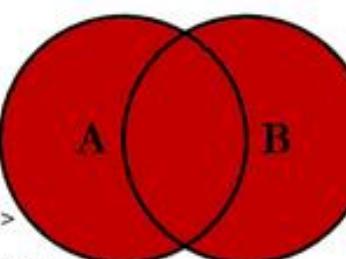
```
SELECT <select_list>  
FROM TableA A  
INNER JOIN TableB B  
ON A.Key = B.Key
```



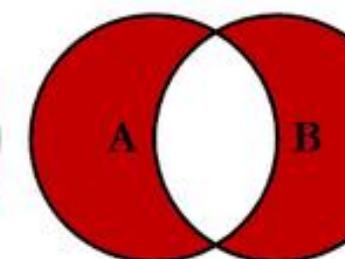
```
SELECT <select_list>  
FROM TableA A  
RIGHT JOIN TableB B  
ON A.Key = B.Key
```



```
SELECT <select_list>  
FROM TableA A  
RIGHT JOIN TableB B  
ON A.Key = B.Key  
WHERE A.Key IS NULL
```



```
SELECT <select_list>  
FROM TableA A  
FULL OUTER JOIN TableB B  
ON A.Key = B.Key
```



© C.L. Moffatt, 2008

```
SELECT <select_list>  
FROM TableA A  
FULL OUTER JOIN TableB B  
ON A.Key = B.Key  
WHERE A.Key IS NULL  
OR B.Key IS NULL
```

DATA PREPARATION - Data Blending

Union Query



- **Online Sales:** Sales made through an online channel.

	Channel Name	Date	CustomerID	Units
1	Online	6/2/2020	US-187982	20
2	Online	6/2/2020	US-235789	30
3	Online	6/5/2020	US-187982	15
4	Online	6/5/2020	US-235789	40

- **Store Sales:** Sales made through the company's physical locations.

	Date	Units	Referer	CustomerID	Channel Name
1	6/4/2020	200	www.powerbi.com	e-78956	Store A
2	6/4/2020	300	www.powerquery.c...	e-78899	Store B
3	6/7/2020	100	www.xbox.com	e-75214	Store A
4	6/7/2020	70	www.microsoft.com	e-23658	Store B

Append

Concatenate rows from three or more tables into a single table.

Two tables Three or more tables

Available table(s)

Online Sales
Store Sales
Wholesale Sales

Add

Tables to append

Online Sales
Store Sales
Wholesale Sales

OK Cancel

DATA PREPARATION - Data Blending

Merge Query



Merge

Select a table and matching columns to create a merged table.

Product

description	ProductSubcategoryKey	Manufacturer	Brand	Class	Style	Color
P3 and WMA		1 Contoso, Ltd	Contoso	Economy	Product0101001	Silver
P3 and WMA		1 Contoso, Ltd	Contoso	Economy	Product0101002	Blue
B driver plays MP3 and WMA		1 Contoso, Ltd	Contoso	Economy	Product0101003	White
splay, plays MP3 and WMA		1 Contoso, Ltd	Contoso	Economy	Product0101004	Silver
splay, plays MP3 and WMA		1 Contoso, Ltd	Contoso	Economy	Product0101005	Red

Product Subcategory

ProductSubcategoryKey	ProductSubcategoryLabel	ProductSubcategoryName	ProductSubcategoryDescription
1 0101		MP4&MP3	MP4&MP3
2 0102		Recorder	Recorder
3 0103		Radio	Radio
4 0104		Recording Pen	Recording Pen
5 0105		Headphones	Headphones

Join Kind

Left Outer (all from first, matching from second)

Use fuzzy matching to perform the merge

Fuzzy merge options

✓ The selection matches 1690 of 1690 rows from the first table.

OK Cancel

DATA PREPARATION - Data Blending

Fuzzy Matching



Fuzzy Matching will enable you to join 2 data sets together where a **regular join may fail**. The Fuzzy Match **identifies records with similar string values** in specified fields.

Fuzzy Matching uses **algorithms** to score how similar 2 words or phrases are.

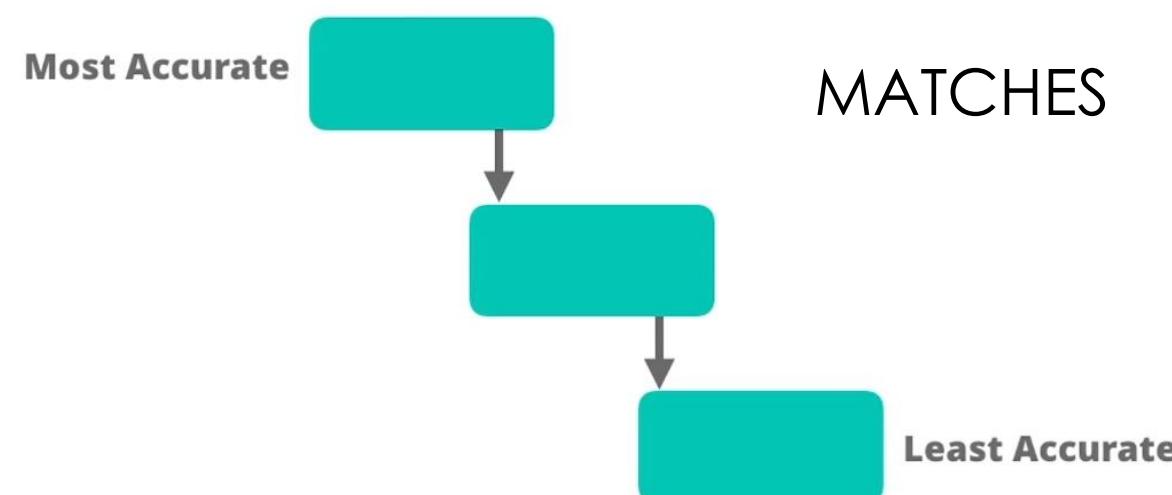
Fuzzy Matching Algorithms

Jaro: The Jaro algorithm is a measure of characters in common, being no more than half the length of the longer string in distance, with consideration for transpositions.

Levenshtein: The Levenshtein algorithm counts the number of edits (insertions, deletions, or substitutions) needed to convert one string to the other.

Definition:

Identifies **NON-IDENTICAL** duplicates of a dataset by **SPECIFYING PARAMETERS** to match on.



DATA PREPARATION - Data Blending

Fuzzy Matching - Example



Andrew Main, 25 State St



Andy Main, 25 State Street

Cot
Coat

It looks at these words and **calculate a closeness** of match score based on the similarity of these words.

The match threshold is the minimum score achieved by the fuzzy matching for it to be considered to be a match

acct_holder	acct_holder2
WOLFMAN CO	WOLFMAN CO LLC
WOLFMAN CO	WOLFMAN CO LLC
CO TEXTILE AND VARNISH	COLORADO TEXTILE AND VARNISH
COLORADO DEPT OF THE TREASURY	COLORADO DEPT OF THE TREASURY
BROWNSVILLE DEPARTMENT OF REVENUE	BROWNSVILLE DEPT OF REVENUE
BROWNSVILLE DEPARTMENT OF REVENUE	BROWNSVILLE DEPT OF REVENUE
CO TEXTILE AND VARNISH	COLORADO TEXTILE AND VARNISH
COLORADO DEPT OF THE TREASURY	COLORADO DEPT OF THE TREASURY
COLORADO DEPT OF THE TREASURY	COLORADO DEPT OF THE TREASURY

DATA PREPARATION - Data Blending

Spatial Matching

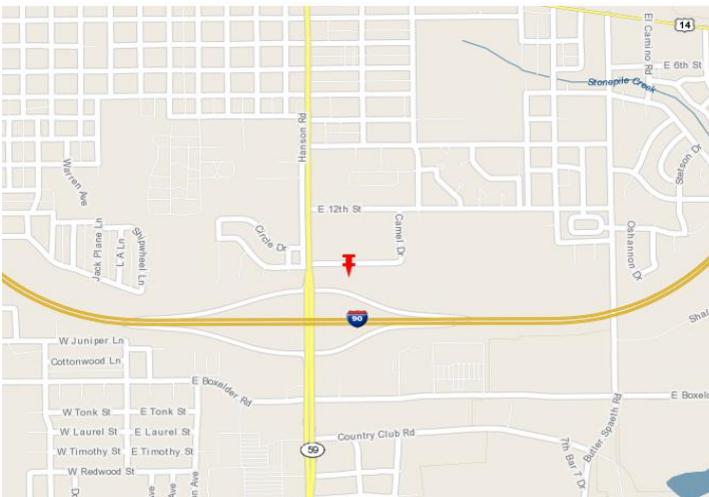


Types of Spatial Data

All of these location data examples are represented by **points, lines, or polygons**

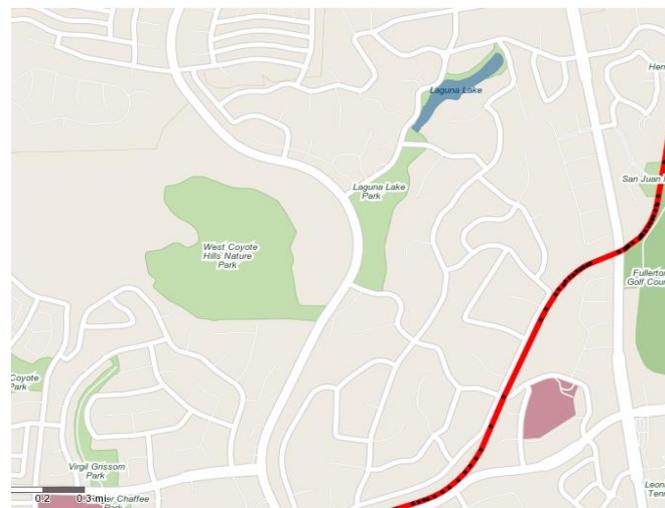
Points

A point, also referred to as a centroid, is in the form of a **latitude and longitude** which we use to **pinpoint** its exact location.



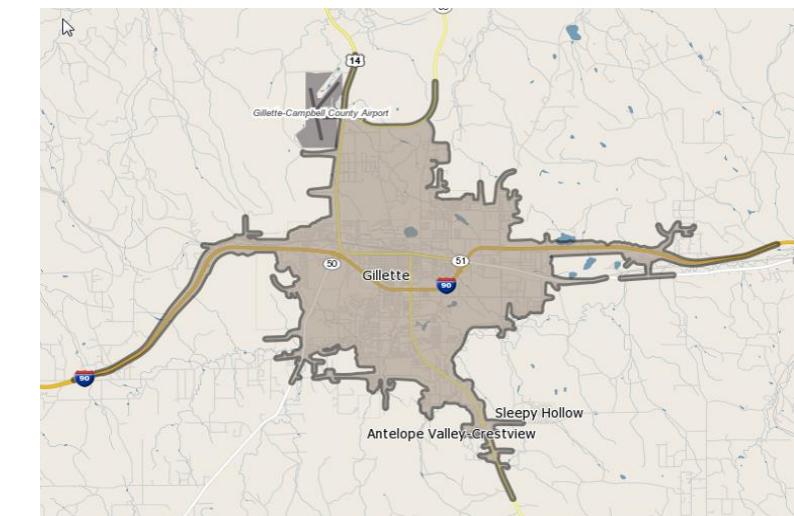
Lines

A line is **a string of latitudes and longitude** locations.



Polygons

Polygons are made up of a **series of longitude and latitude coordinates** defining all of the vertices of a region.



DATA PREPARATION - Data Blending

Spatial Matching



Spatial Blending



There **aren't fields** that can be used to join them together

Gray area: How many customers fall within a store trade area is to match them and assign a store number to them

DATA PREPARATION – Transform Data

Why Combine Queries



- Why combine queries
- Append
- Merge

Why Combine Queries?

- Combining queries means:
 - Connecting to two or more data sources
 - Shaping them as needed
 - Consolidating them into one useful query
- There are two primary ways of combining queries: Append and Merge

Append

- Append the query when you have additional rows of data, that you would like to add to an existing query
- The result set will have same column, as column names are exactly the same as in the queries to be appended
- Append queries do not remove duplicates

Merge

- Merge the queries once you have one or additional columns that you would like to add to another query
- There ought to be a joining or matching criteria between two queries
- The number of rows depends on matching criteria between queries
- The number of columns depends on columns selected within the result set

DATA PREPARATION – Check List



CREATING AN ANALYTICAL DATASET		Issues	1st Fix-date	2nd Fix-date
Data Source	Enough Data Up to Date			
Data Types	Data Types correctly			
Data Issues	Dirty Data Not Parsed Correctly Extra characters Unexpected Pattern Incorrect Data Duplicate Data Records Misspelled Entries			
	Missing Data Deleting Missing Data Imputation Advanced methods			
	Outliers Errors: Cross-check & fix Errors: Detect No certainty: Remove if Insignificant Certainty: Truncation			
Data Formatting	Transposing Aggregating Data Cross Tabulation			
Data Blending	Unions Joins Fuzzy Matching Spatial Matching			

“Garbage in, garbage out”



Your analysis is as good as your data.



Business Intelligence in Corporates

Summary



Know Role of Power Query & How to get data in Power Query?

PQ Basic Transformation. What is Profiling Data?

Data Issue – Bad Shape (Transpose + Unpivot/Pivot + Aggregation)

Data Issue – Missing Data + Outliers

Load from File vs Load from Folder

Data Blending. Why need to Merge / Append Queries?

2. End-to-End Business Intelligence Workflow in Power BI

01

Data Preparation

1. Power Query Overview
2. Get Data
3. PQ – Basic Transform Data
4. Profiling Data
5. Data Issues
 - 4a. Bad Shape + Dirty Data
 - 4b. Missing Data + Outliers
5. Combine Data from Folder
6. Blending Data
7. Checklist

02

Data Modelling

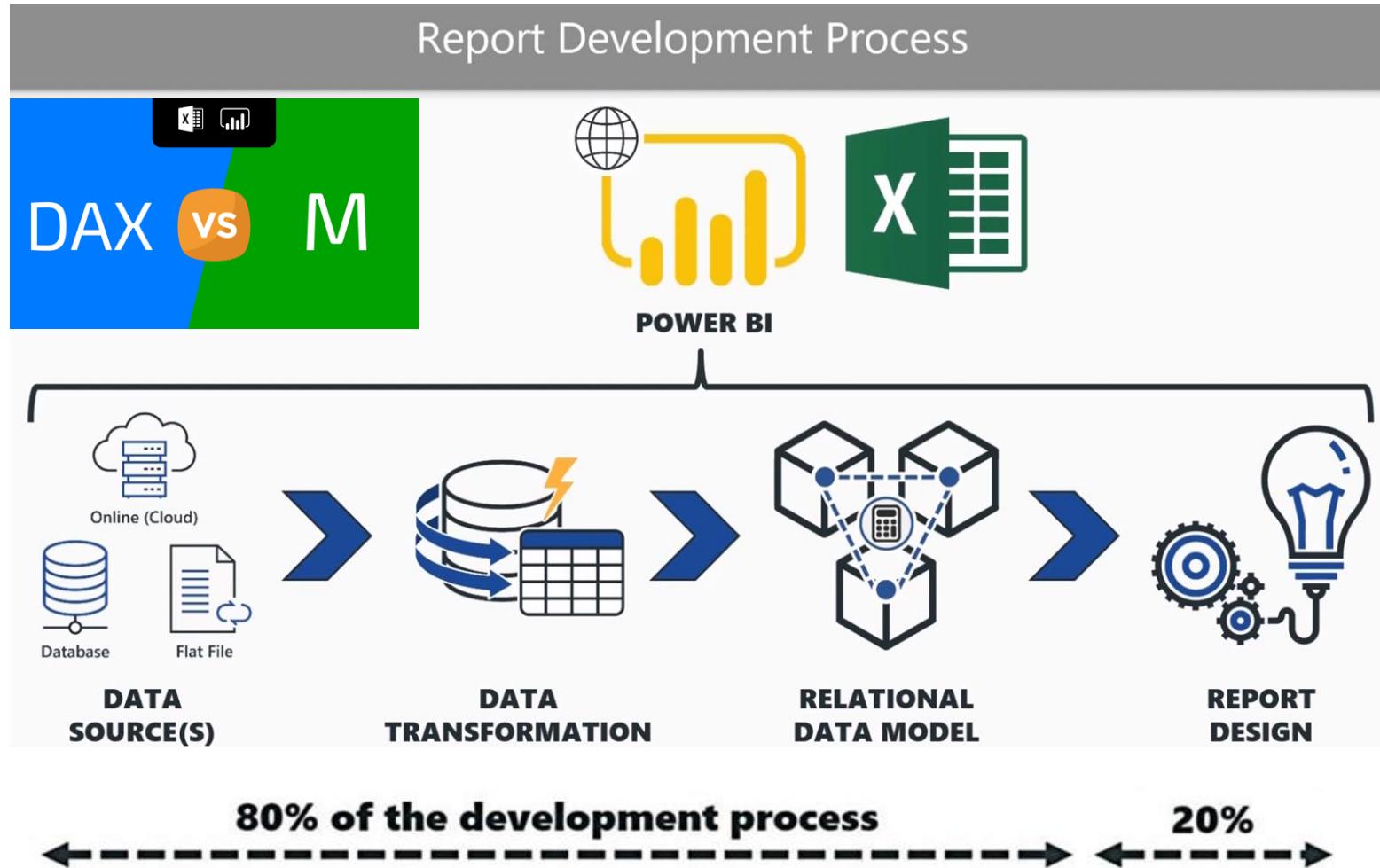
1. Data Model Overview
2. Fact & Dimension
3. Schema
4. Cardinality
5. Cross Filter Direction
6. Hierarchies

03

End-to-End in Power BI Cloud

1. Introduce PBI Ecosystem (PBI Service)
2. Fact & Dimension
3. Prep Data (On Pro & Premium)
4. Data Modeling
5. Report and Dashboard
6. Refresh Scorecard & Metrics
7. Sharing, Collaboration, (PBI Mobile)
8. Deployment Pipelines

Data Modeling





What is a Data model?



A Power BI **Data Model** is a **collection of tables with relationships** which enable your business users to easily understand and explore their data to get business insights.

Why is it important to have a Good Data model?

- Improves understandability of the data
- Increases performance of dependent processes and systems
- Increases resilience to change



A Framework to Help You Think About Your Data

Who

What

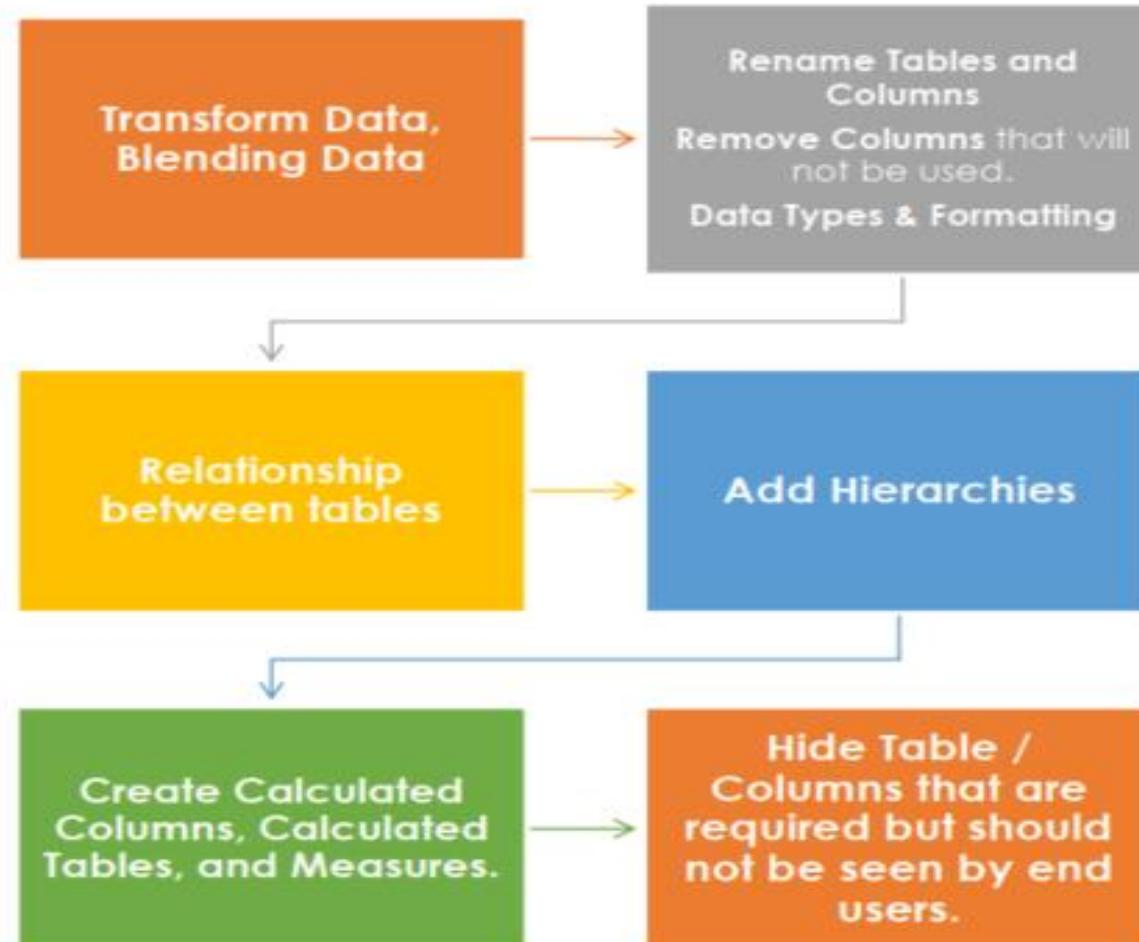
When

Where

1. Start with your transactions
2. Ask
 1. Who
 2. What
 3. When
 4. Where
3. Single key column for each lookup table
4. Load and join the tables



Data Modeling





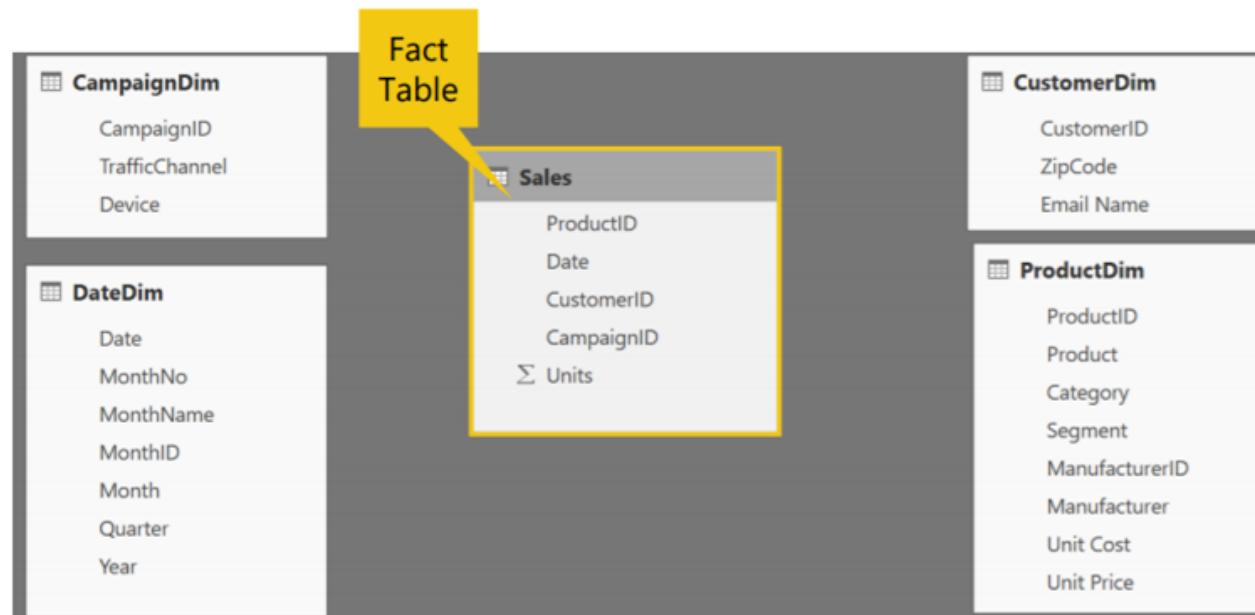
- **Fact Tables and Dimension Tables**
- Schema
- Relationships
- Cardinality
- Cross filter direction
- Hierarchies

DATA MODEL

Fact Tables and Dimension Tables



Components of a data model – Fact Table



Fact Table

- Contains Measures (or items to be aggregated) of a business process

Examples:

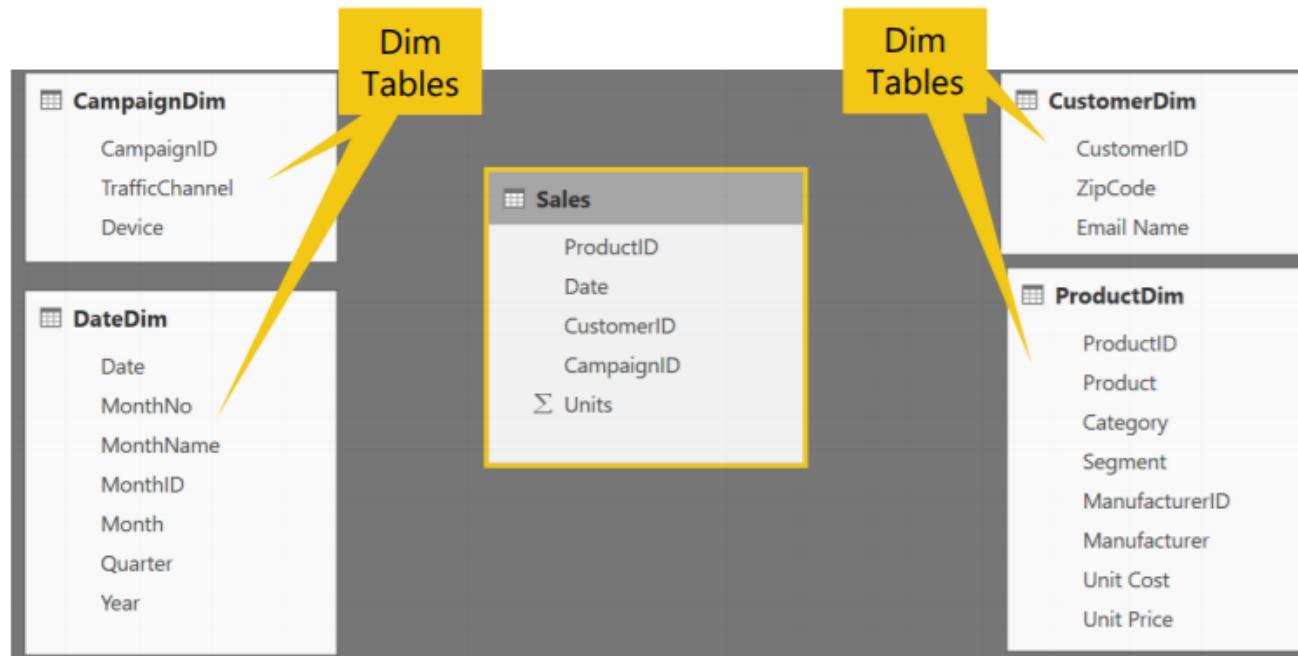
- Transactions
- Sales Revenue
- Units
- Cost

- Measures are usually sliceable.

Examples: By Month,
By Customer



Components of a data model – Dim Table



Dim Table

A Dim (or Dimension) table contains descriptive attributes that define how a fact should roll up.

Examples:

By month, By Customer,
By Geo



Table Types In Data Models



DATA TABLE

- Typically **TALL** (Many Rows)
- **MOSTLY** Dates and Numbers 
- Do **MATH** against it (E.g. SUM, AVERAGE, MIN, MAX, etc...) 
- **Fast** changing (Updated Often)
- **MAY** contain Time Dimensions (E.g. Order Date, Record Time, etc...)



LOOKUP TABLE

- Typically **WIDE** (Many Columns)
- **MOSTLY** Text 
- **LOOKUP** Information (E.g. Name, Address, Description, etc...) 
- **Slow** Changing (Updated Less Often)
- Does **NOT** typically contain Time Dimensions



Duplicate Information On Single Tables

Order ID	ProductID	Product Name	Product Sub Category	Product Category	Amount
1	1	Road Bike 150	Road Bikes	Bikes	\$364.78
1	2	Mountain Bike 370	Mountain Bikes	Bikes	\$519.85
1	3	2L Water Bottle	Bike Gear	Accessories	\$6.94
2	1	Road Bike 150	Road Bikes	Bikes	\$364.78
2	2	Mountain Bike 370	Mountain Bikes	Bikes	\$519.85
2	3	2L Water Bottle	Bike Gear	Accessories	\$6.94
3	1	Road Bike 150	Road Bikes	Bikes	\$364.78
3	2	Mountain Bike 370	Mountain Bikes	Bikes	\$519.85
3	3	2L Water Bottle	Bike Gear	Accessories	\$6.94
4	1	Road Bike 150	Road Bikes	Bikes	\$364.78
4	2	Mountain Bike 370	Mountain Bikes	Bikes	\$519.85
4	3	2L Water Bottle	Bike Gear	Accessories	\$6.94

DATA MODEL

Fact Tables and Dimension Tables



Duplicate Information On Single Tables

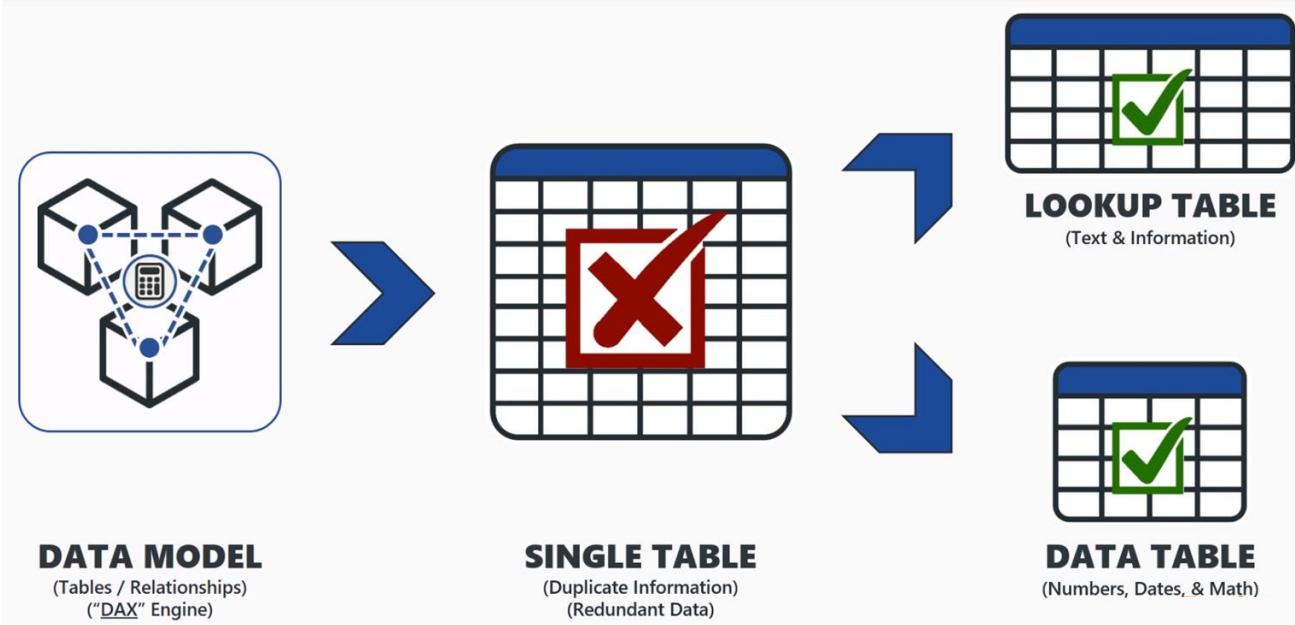
Order ID	ProductID	Product Name	Product Sub Category	Product Category	Amount
1	1	Road Bike 150	Road Bikes	Bikes	\$364.78
1	2	Mountain Bike 370	Mountain Bikes	Bikes	\$519.85
1	3	2L Water Bottle	Bike Gear	Accessories	\$6.94
2	1	Road Bike 150	Road Bikes	Bikes	\$364.78
2	2	Mountain Bike 370	Mountain Bikes	Bikes	\$519.85
2	3	2L Water Bottle	Bike Gear	Accessories	\$6.94
3	1	Road Bike 150	Road Bikes	Bikes	\$364.78
3	2	Mountain Bike 370	Mountain Bikes	Bikes	\$519.85
3	3	2L Water Bottle	Bike Gear	Accessories	\$6.94
4	1	Road Bike 150	Road Bikes	Bikes	\$364.78
4	2	Mountain Bike 370	Mountain Bikes	Bikes	\$519.85
4	3	2L Water Bottle	Bike Gear	Accessories	\$6.94

DATA MODEL

Fact Tables and Dimension Tables



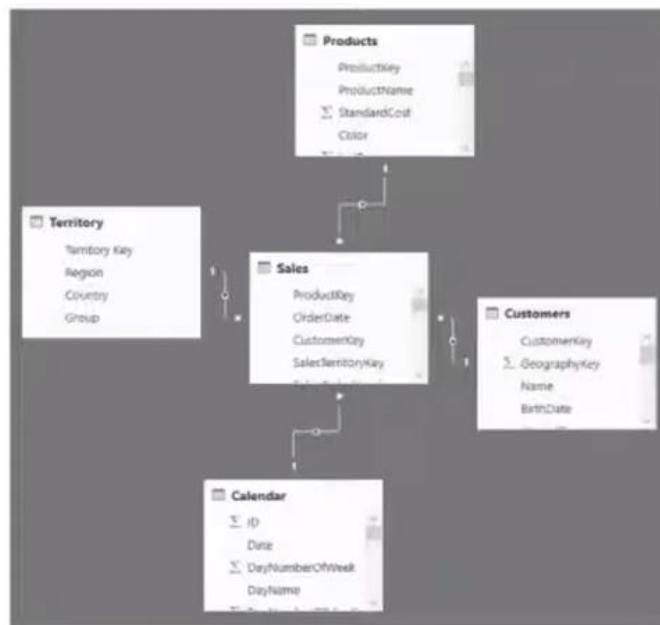
Multiple Tables In Data Models



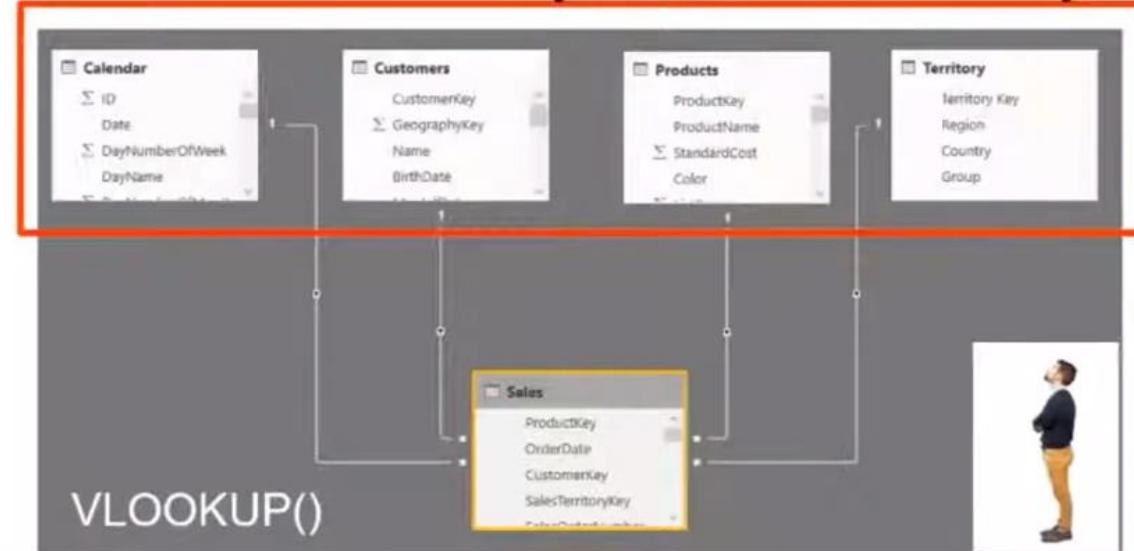


Star Schema is the Optimal Structure

One Entity, One Table



Lookup Tables at the Top



- The arrows are the most important part of the relationship – they indicate filter propagation
- You must understand filter propagation to be good at Power BI and DAX.

DATA MODEL

Fact Tables and Dimension Tables



Separating Tables by Data Type

Order ID	ProductID	Amount
1	1	\$364.78
1	2	\$519.85
1	3	\$6.94
2	1	\$364.78
2	2	\$519.85
2	3	\$6.94
3	1	\$364.78
3	2	\$519.85
3	3	\$6.94
4	1	\$364.78
4	2	\$519.85
4	3	\$6.94



ProductID	Product Name	Product Sub Category	Product Category
1	Road Bike 150	Road Bikes	Bikes
2	Mountain Bike 370	Mountain Bikes	Bikes
3	2L Water Bottle	Bike Gear	Accessories



DATA TABLE

(Numbers, Dates, Math)

LOOKUP TABLE

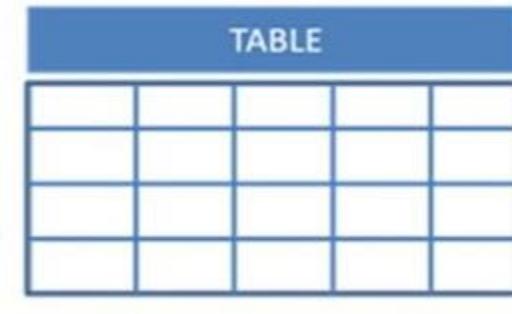
(Text & Information)

DATA MODEL

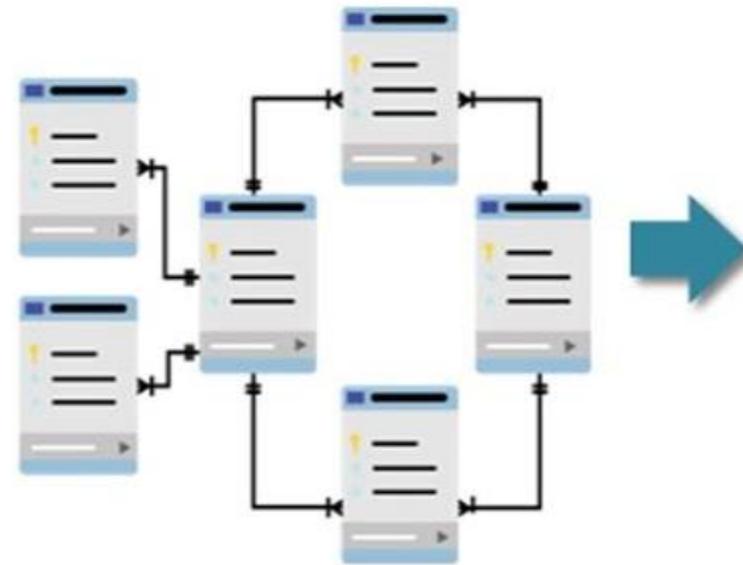
Fact Tables and Dimension Tables



Database
Normalization



1NF 2NF 3NF





Facts vs Dimensions

Facts get Aggregated

Sum of Sales

Count of Products

Dimensions slice Facts

By Account Group

By Customer Class

By Year, Month, Day

Easy to identify based on where they live

Department Name

- Rentals
- Restaurant
- Landscaping
- General & Admin

Dimension

Date Range

- All Periods
- YEARS
- 2007 2008 2009

Dimension

Drag fields between areas below:

Filters

Columns

- Month_Short
- Σ Values

Dimension Dimension

Rows

- AccGroup
- AccSubGroup

Dimension Dimension

Σ Values

- Sum of Amount
- Count of Customers

It's a Fact!



If you want to aggregate it to get a single data point from multiple records, it's a Fact



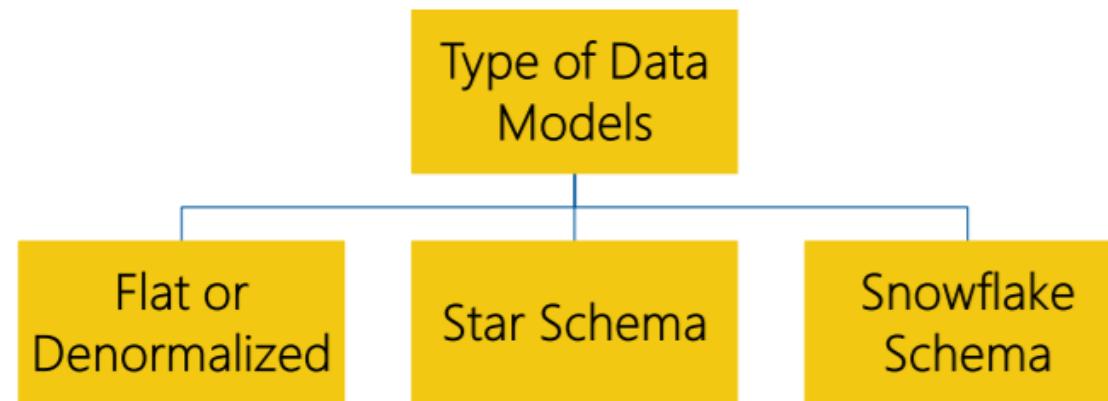
If it won't be aggregated, it's a Dimension



- Fact tables and dimension tables
- **Schema**
- Relationships
- Cardinality
- Cross filter direction
- Hierarchies



Data Model Brings Facts and Dimensions Together





Flat or Denormalized Schema

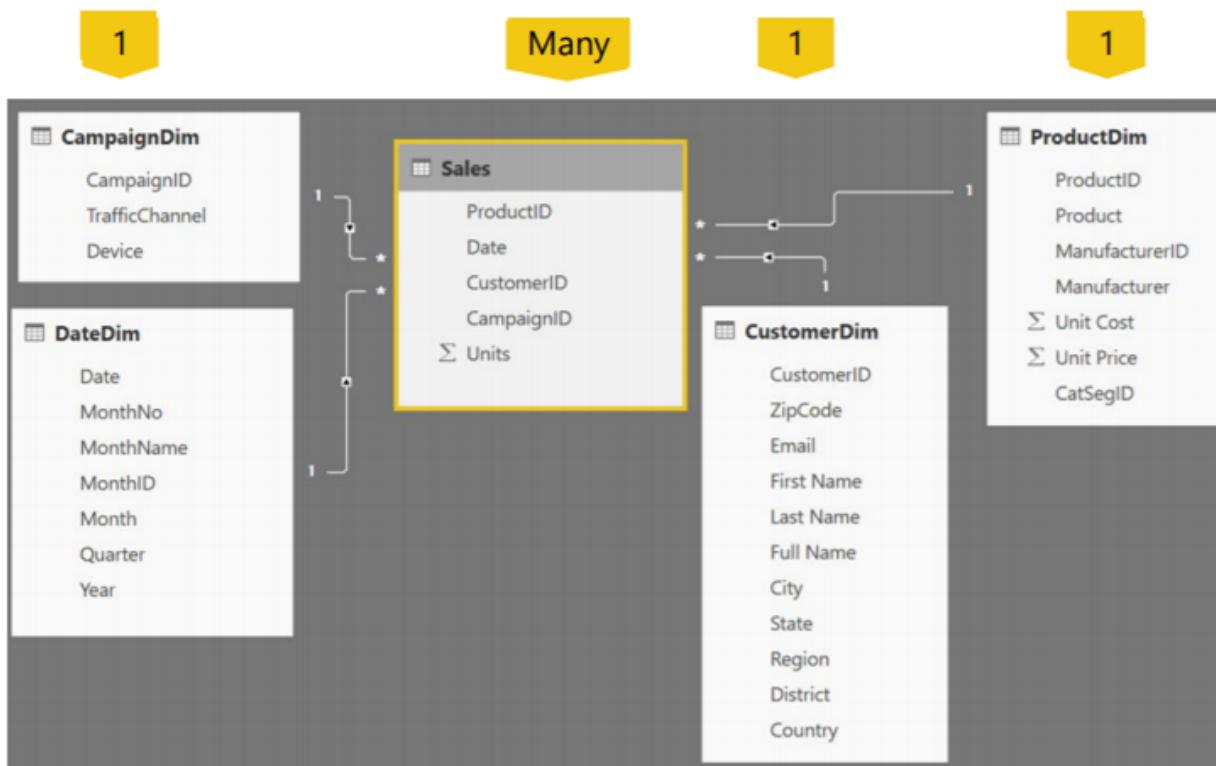


	ProductID	Product	Date	CustomerID	Email	Last Name	First Name	Full Name	CampaignID	Units	CategoryID
1	670	Maximus UD-41	8/25/2011	70283	Farrash.HenryReyna.com	Rent	Farrash	Farrash Rent	22	1	10
2	685	Maximus UD-30	3/24/2014	70283	Farrash.HenryReyna.com	Rent	Farrash	Farrash Rent	18	1	10
3	685	Maximus UD-30	11/28/2014	128234	Martha.McCainReyna...	McCain	Martha	Martha McCain	8	1	10
4	685	Maximus UD-30	4/21/2018	27183	Heads.McIntoshReyna...	McIntosh	Heads	Heads McIntosh	22	1	10
5	695	Maximus UD-30	1/6/2013	230930	Lunes.WalkerReyna.com	Walker	Lunes	Lunes Walker	21	1	10
6	695	Maximus UD-30	3/23/2013	182241	Upton.PageReyna.com	Page	Upton	Upton Page	17	1	10
7	695	Maximus UD-34	8/25/2011	196385	Drake.WellsReyna.com	Wells	Drake	Drake Wells	22	1	4
8	695	Maximus UD-34	8/30/2014	140000	Wallace.BenderReyna...	Bender	Wallace	Wallace Bender	17	1	4
9	695	Maximus UD-34	8/12/2014	110381	Aetna.EricksonReyna...	Erickson	Aetna	Aetna Erickson	20	1	4
10	695	Maximus UD-34	4/16/2014	48327	John.BradleyReyna.com	Bradley	John	John Bradley	7	1	4
11	695	Maximus UD-34	2/28/2013	65982	Toku.GrossReyna.com	Gross	Toku	Toku Gross	17	1	4
12	695	Maximus UD-34	6/6/2013	87	Yoshi.HenryReyna.com	Grant	Yoshi	Yoshi Grant	18	1	4
13	695	Maximus UD-34	5/14/2013	54797	Silvia.CarrilloReyna...	Carrillo	Silvia	Silvia Carrillo	10	1	4
14	695	Maximus UD-34	4/9/2013	248715	Mark.HewittReyna.com	Hewitt	Mark	Mark Hewitt	19	1	4
15	695	Maximus UD-34	4/28/2013	248715	Mark.HewittReyna.com	Hewitt	Mark	Mark Hewitt	8	1	4
16	695	Maximus UD-34	3/29/2014	240031	Oscar.AvilaReyna.com	Avila	Oscar	Oscar Avila	18	1	4
17	695	Maximus UD-34	2/26/2014	201004	Dunoon.McIntoshReyna...	McIntosh	Dunoon	Dunoon McIntosh	19	1	4
18	615	Maximus UD-80	5/4/2012	212645	Jacob.RantiagoReyna...	Ramiro	Jacob	Jacob Santiago	22	1	10
19	615	Maximus UD-80	5/4/2012	70486	Hillary.CollerReyna...	Coller	Hillary	Hillary Coller	22	1	10
20	615	Maximus UD-80	5/4/2012	124489	Chester.MitchellReyn...	Mitchell	Chester	Chester Mitchell	22	1	10
21	615	Maximus UD-80	5/4/2012	221470	Sage.YangReyna.com	Yang	Sage	Sage Yang	22	1	10
22	615	Maximus UD-80	6/5/2012	168009	Wallace.BenderReyna...	Bender	Wallace	Wallace Bender	22	1	10
23	615	Maximus UD-80	6/3/2012	154426	Liliana.BonlapReyna...	Bonlap	Liliana	Liliana Bonlap	22	1	10
24	615	Maximus UD-80	6/4/2012	191291	Joselle.LeeReyna.com	Lee	Joselle	Joselle Lee	22	1	10

- All attributes for model exist in a single table
- Highly inefficient
- Model has extra copies of data > slow performance
- Size of a flat table can grow and quickly "blow up" as data model becomes complex



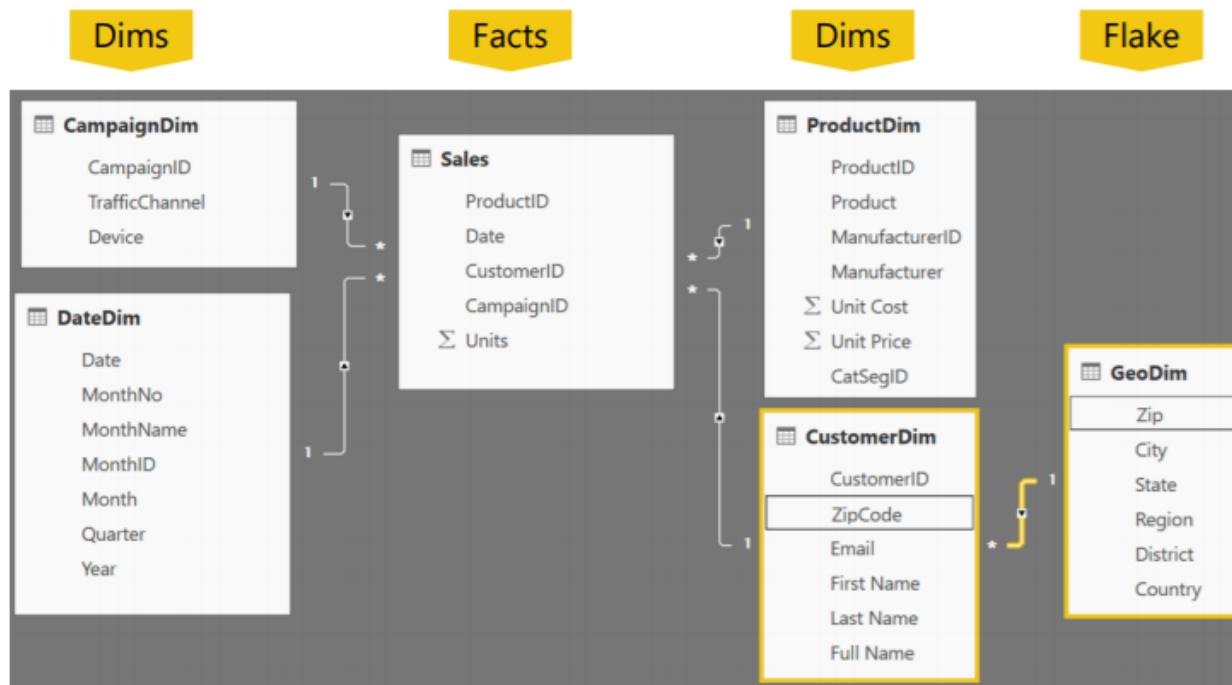
Star Schema



- Fact table in the middle
- Surrounded by Dims
- Looks like a 'Star'
- Fact table is the "Many" side of the (one to many) relationship



Snowflake Schema



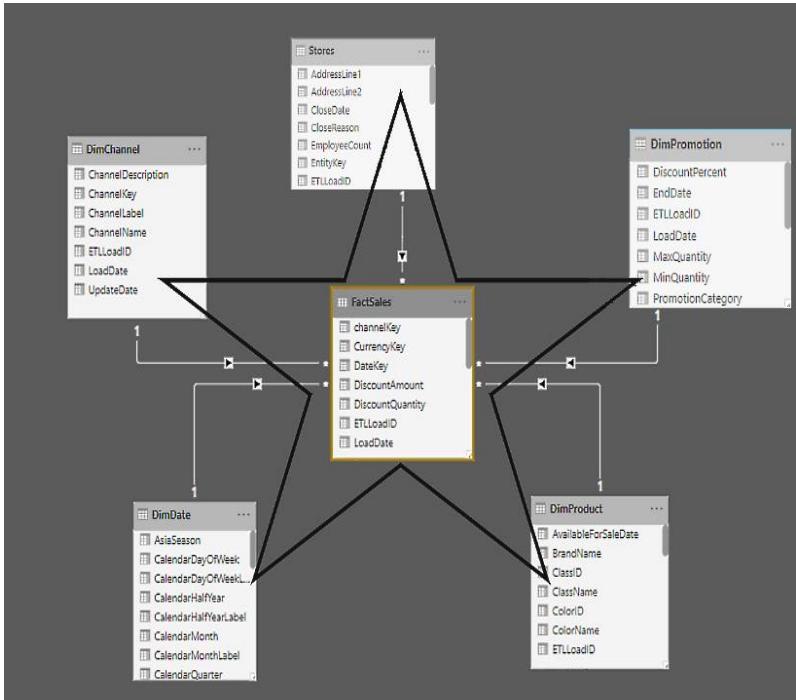
- Center is a Star schema
- Fact table in middle
- Surrounded by Dims
- Dims “snowflake” off of other Dims
- If you have many, it looks like a ‘Snowflake’
- Dim or Fact tables can be the “Many” side of the relationship

DATA MODEL

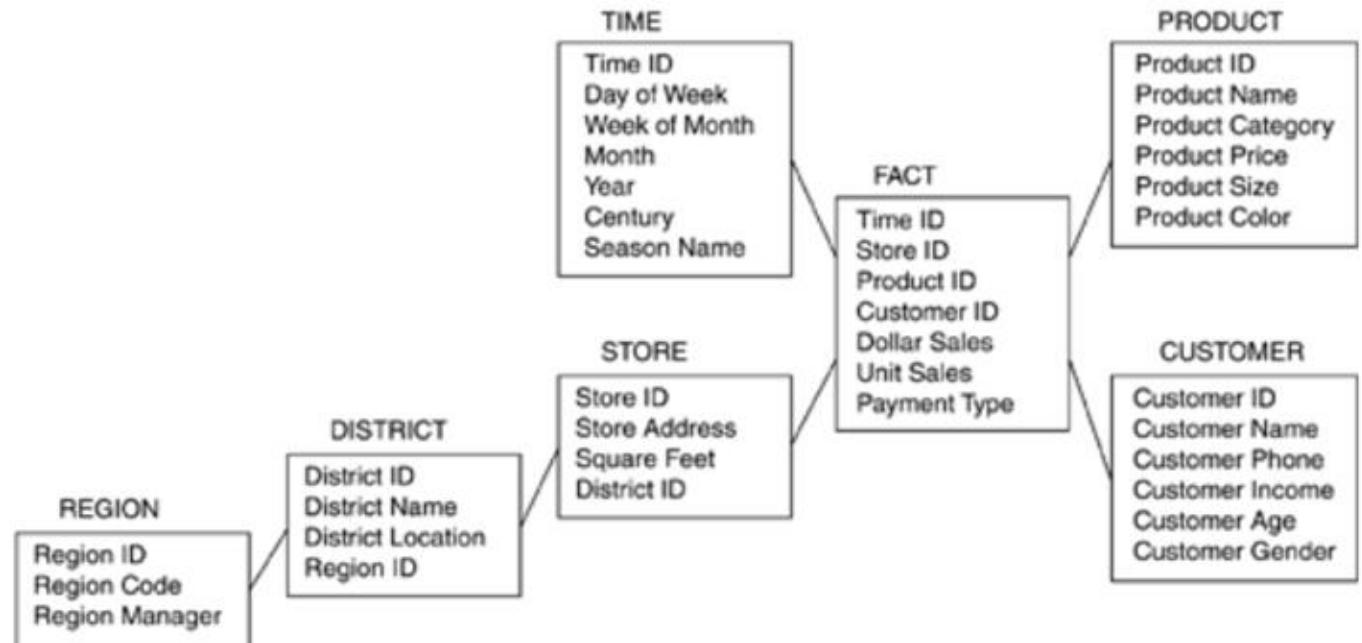
Schema



Star Schema



Snowflake Schema



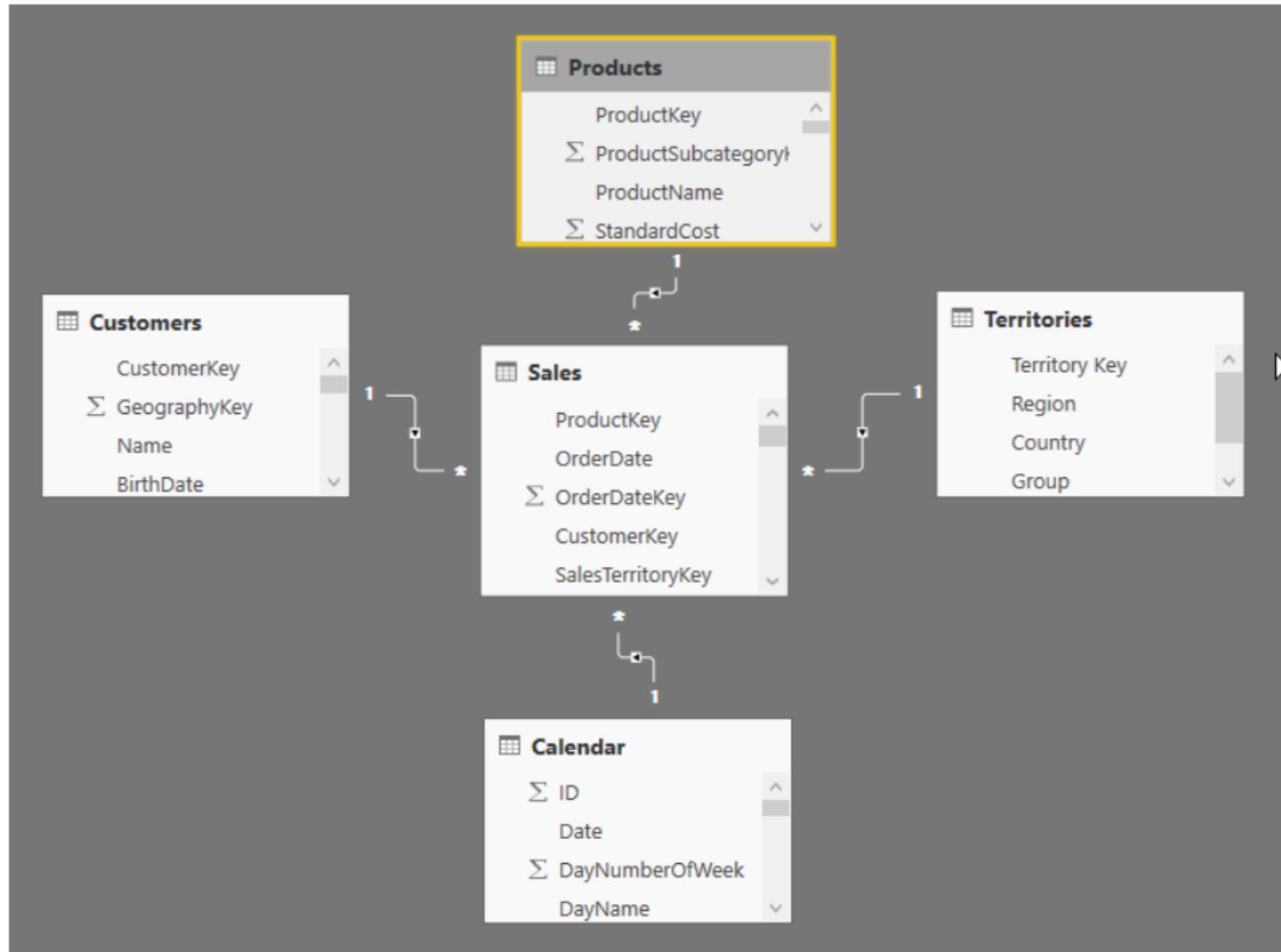
DATA MODEL

Schema

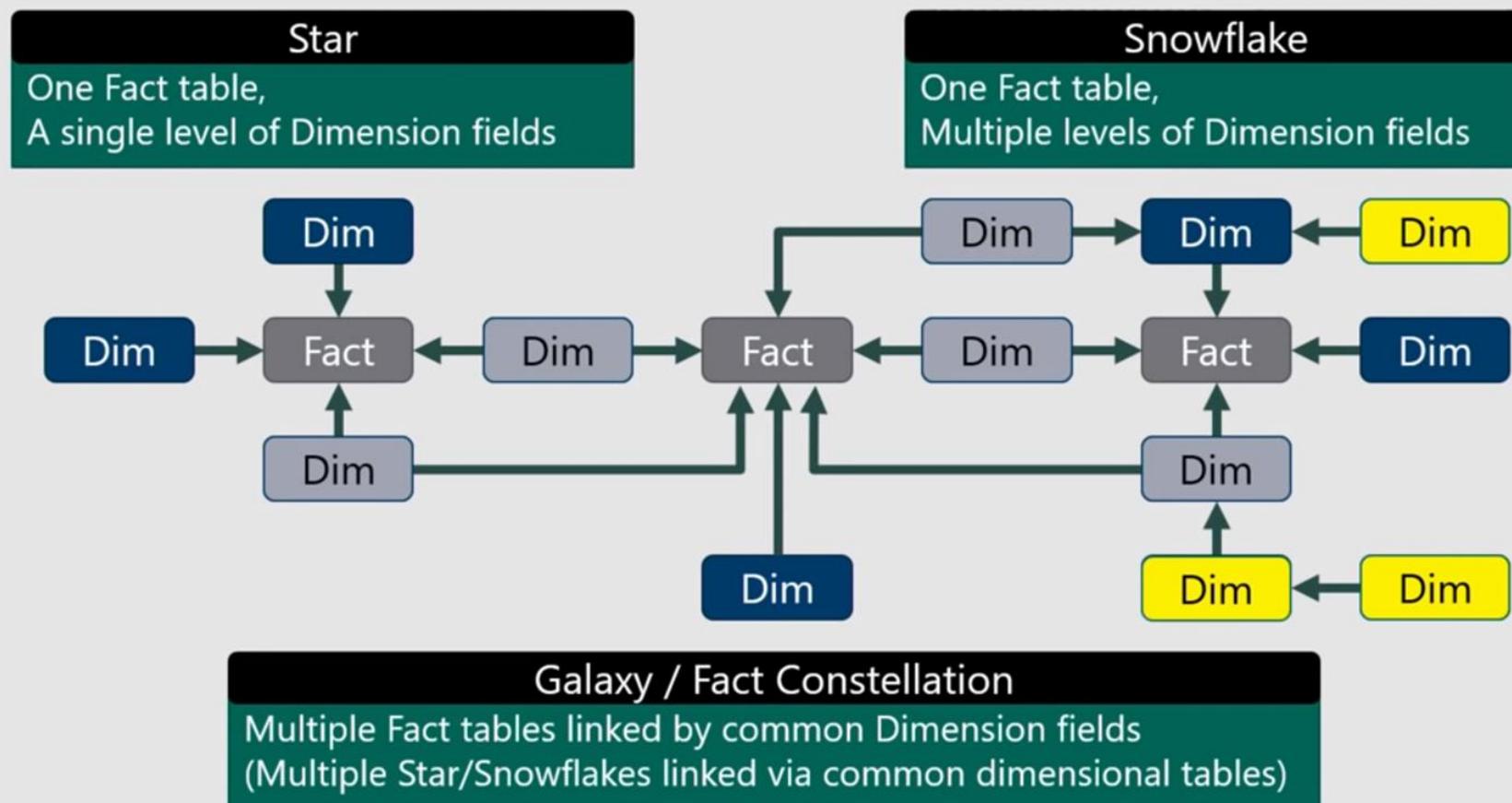


The Star Schema is the Optimal Shape

The generally accepted approach to bringing data into Power Pivot is to bring in your data in a “Star Schema” format. This is a technical term coming from the Kimball methodology (also known as dimensional modelling) which describes the logical way data should be structured for optimal reporting performance.

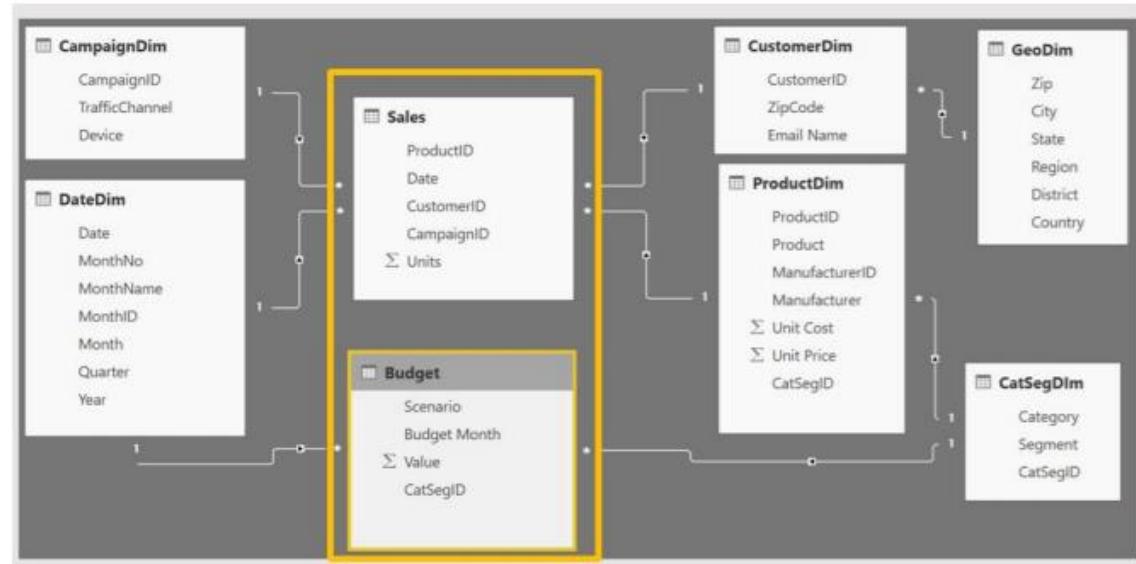


Relationship Schemas





Granularity & Multiple Fact Tables



Sales (Daily by Product)

Budget (Monthly by Product Category & Product Segment)

- Grain (**granularity**) measures the level of detail in a table
Example:
One row per order or per Item
Daily or Monthly date grain
- If your facts have very different granularities, split them into Multiple Fact tables & connect them to shared dimensions at the lowest common granularity.



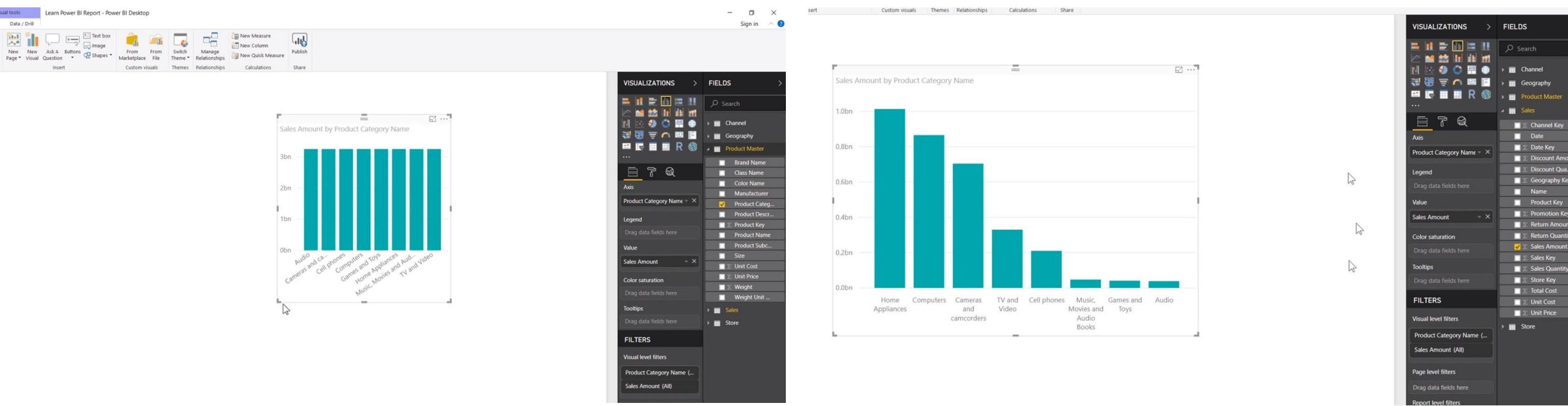
- Fact tables and dimension tables
- Schema
- **Relationships**
- Cardinality
- Cross filter direction
- Hierarchies

DATA MODEL

Relationship



No Relationships



DATA MODEL

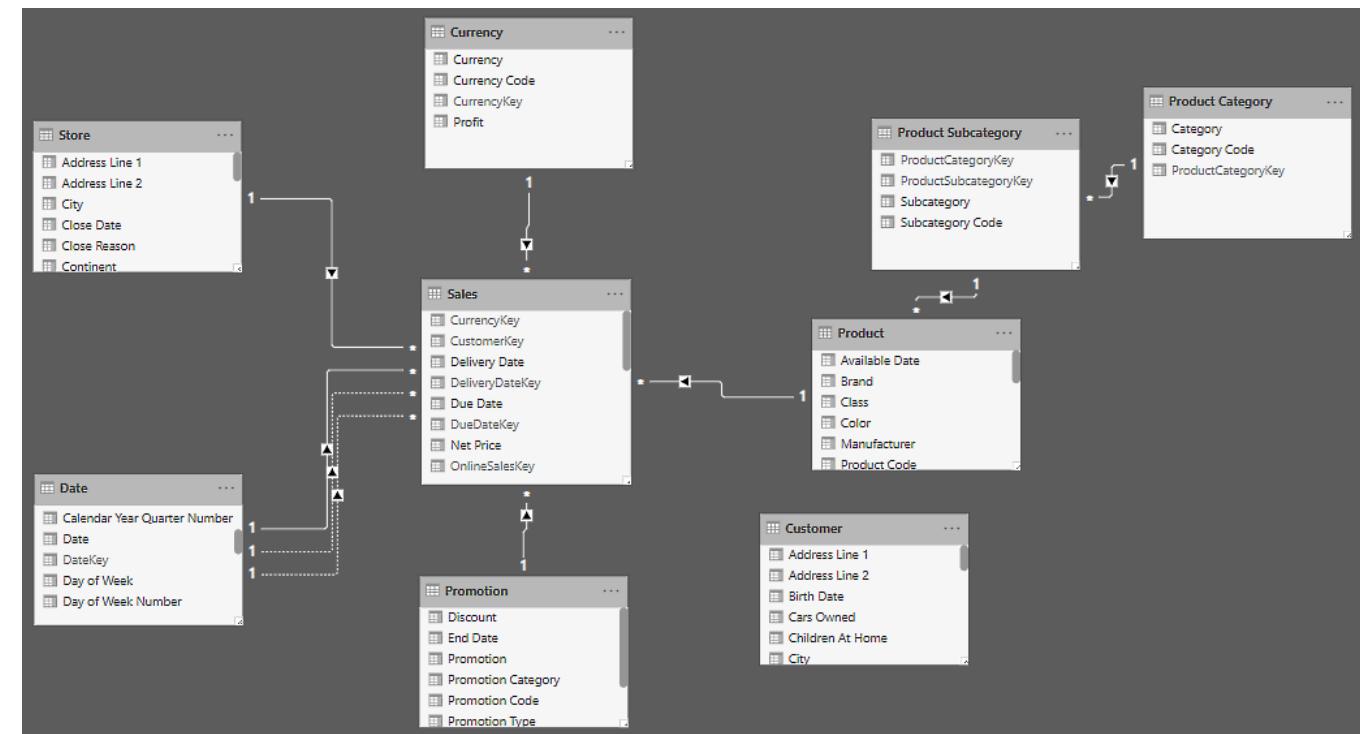
Relationship



Once Data imported, the next step is **creating relationships** among tables.

If data is imported from database where primary and foreign keys have already been defined, these should be imported along with the tables.

A relationship is defined by a **single column** from each table. You **cannot use multiple columns to define a relationship**, but you can create new columns in each table that consist of multiple columns concatenated together. You can then use these to create the relationships.

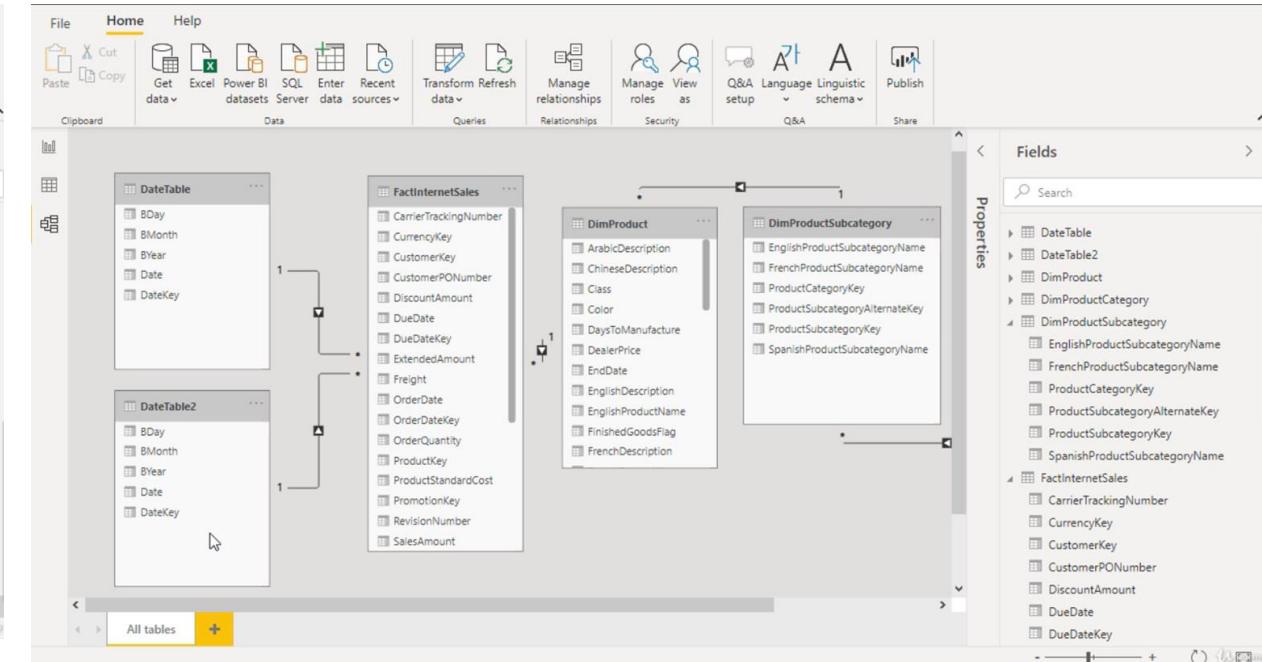
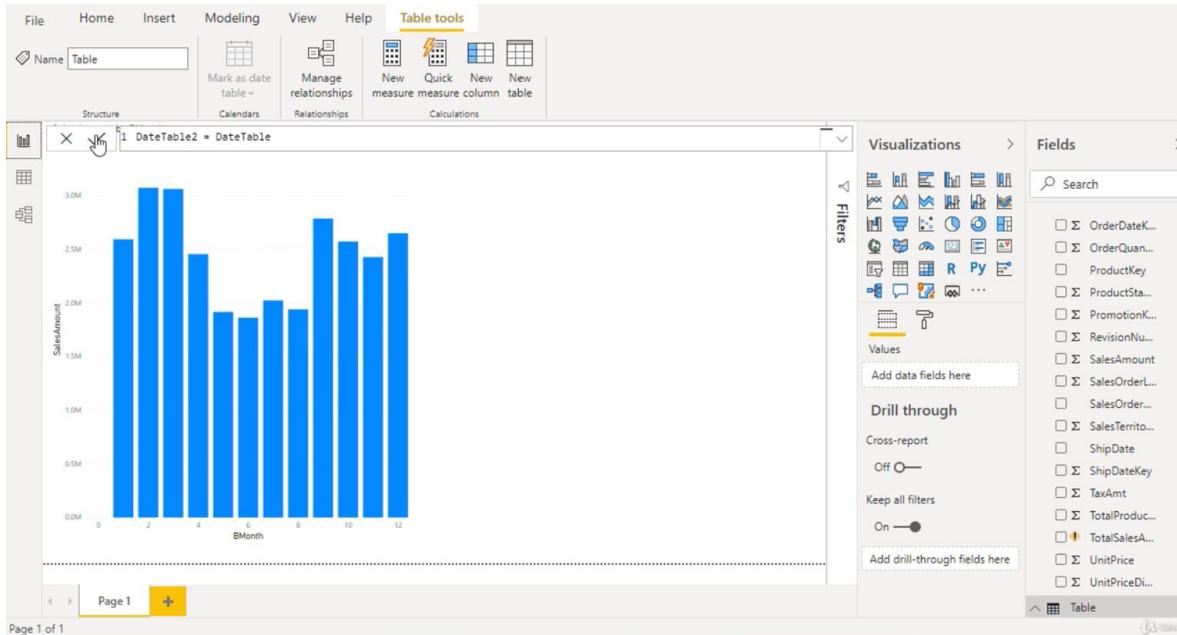


DATA MODEL

Cardinality



Fact Table has multiple Date? (Order date, Shipped date)





Creating Relationships

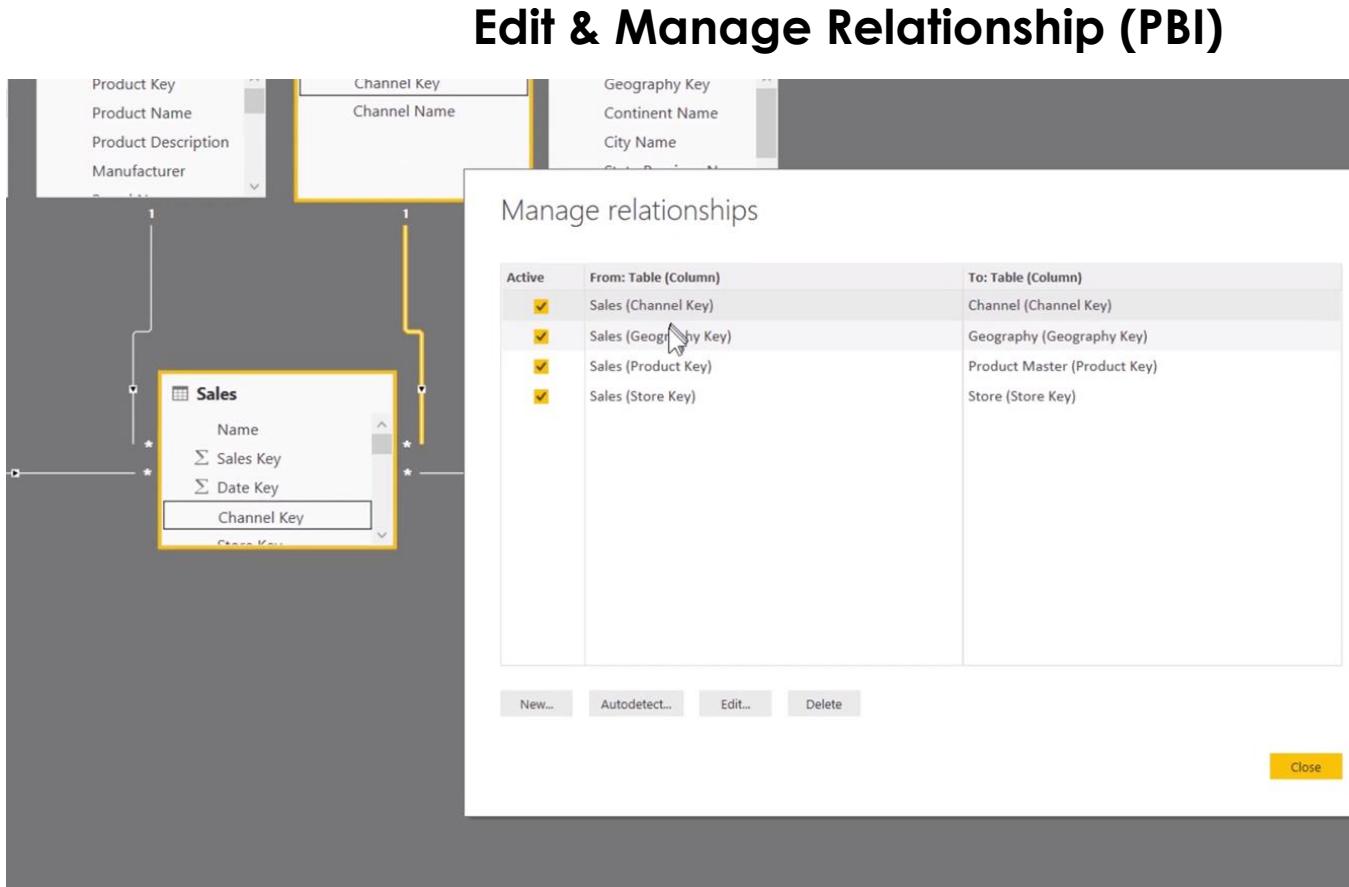
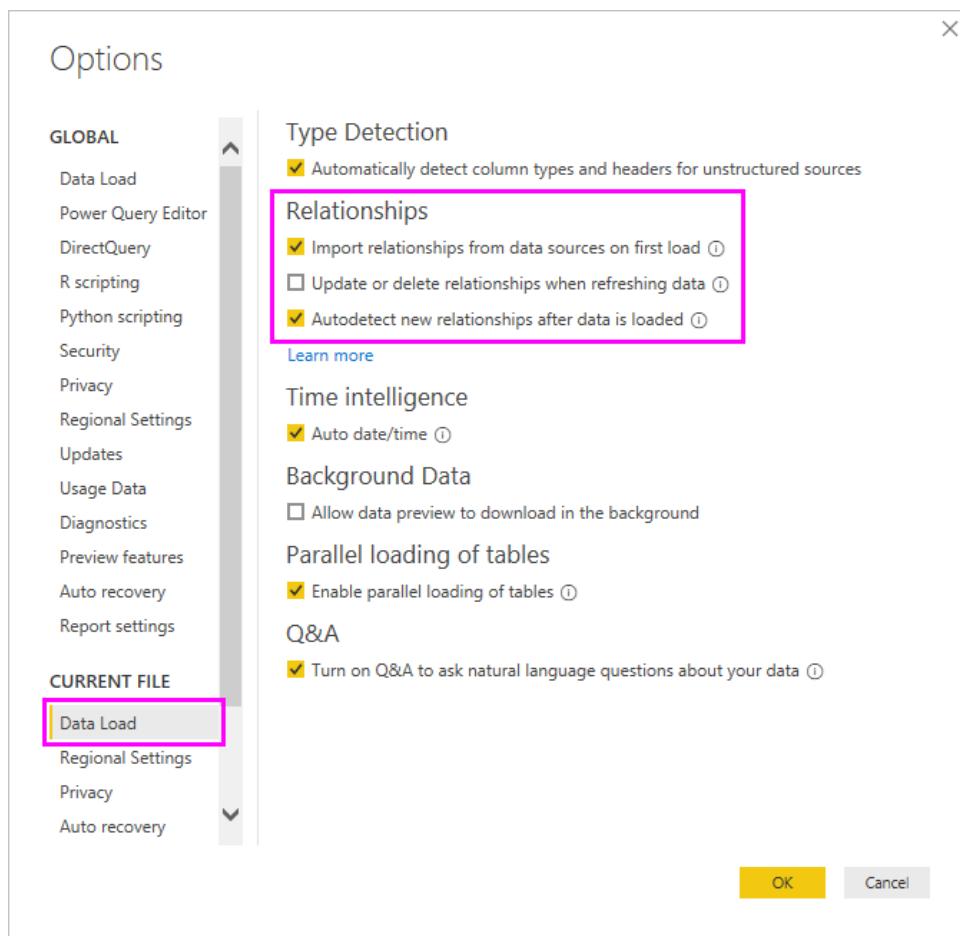
- Power BI Desktop looks at column names in the tables you are querying to determine if there are any potential relationships
- Two methods to create relationships:
- **Use Autodetect:**
 - It detects relationships in most cases
- **Creating it manually:**
 - Defining your own relationships is also possible

DATA MODEL

Relationship



Autodetect new relationship

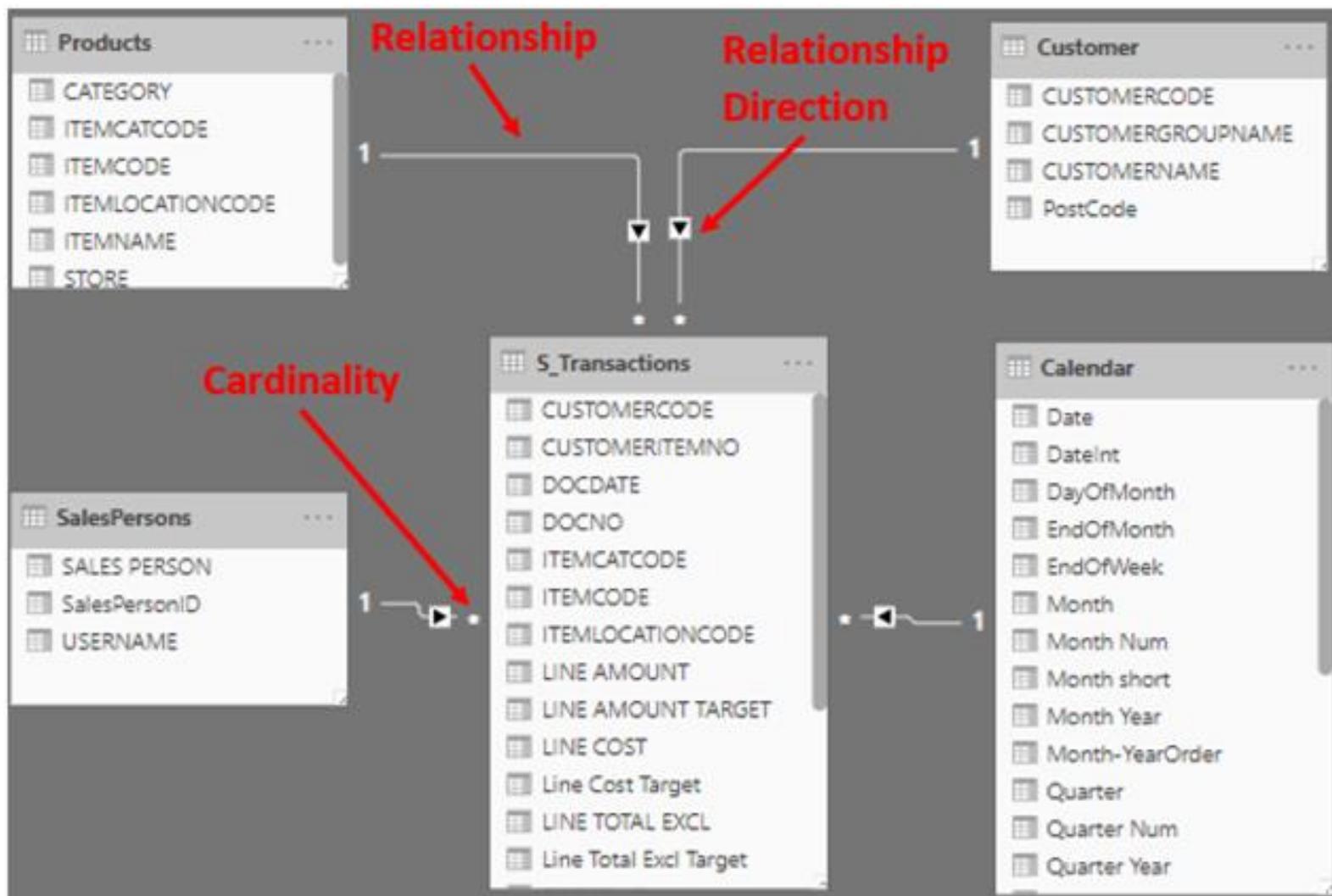




- Fact tables and dimension tables
- Schema
- Relationships
- **Cardinality**
- Cross filter direction
- Hierarchies

DATA MODEL

Relationship

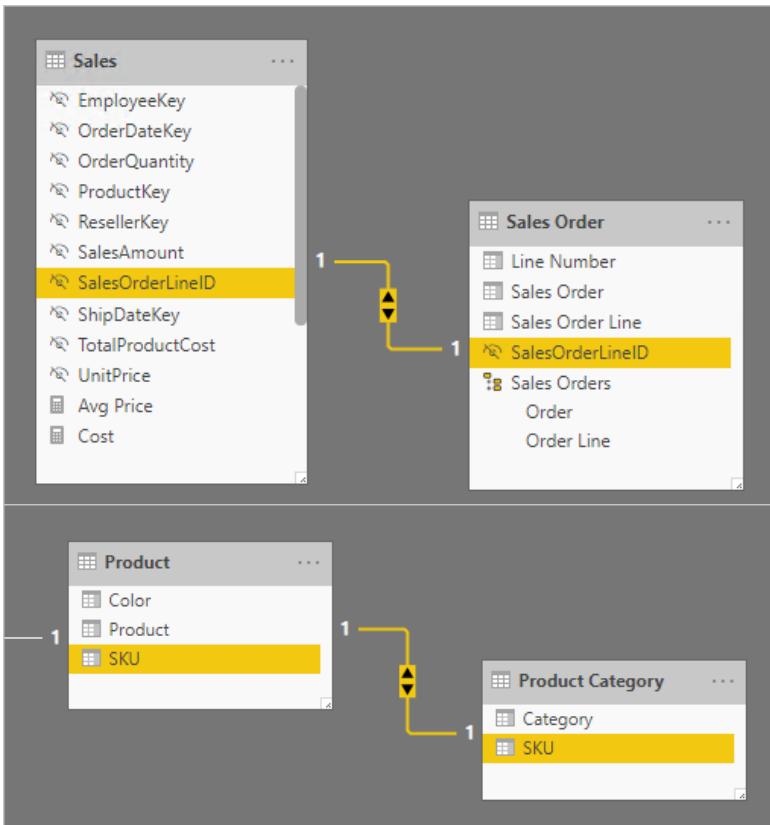


DATA MODEL

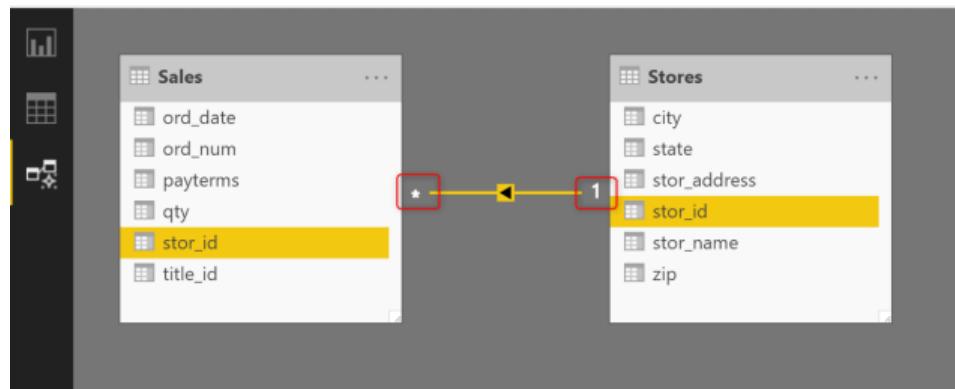
Cardinality



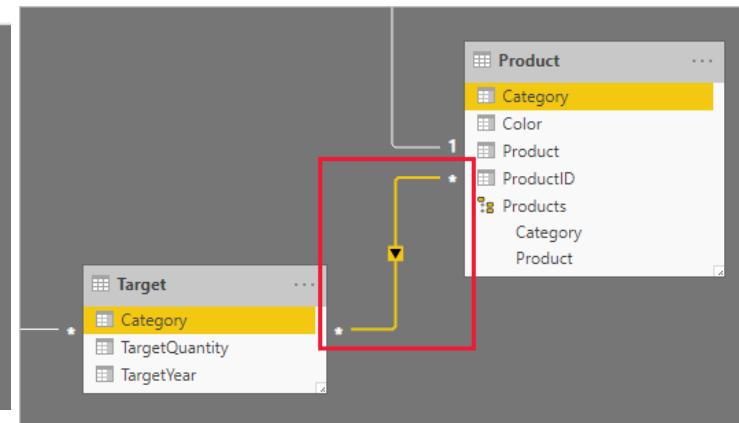
One-to-One



One-to-Many

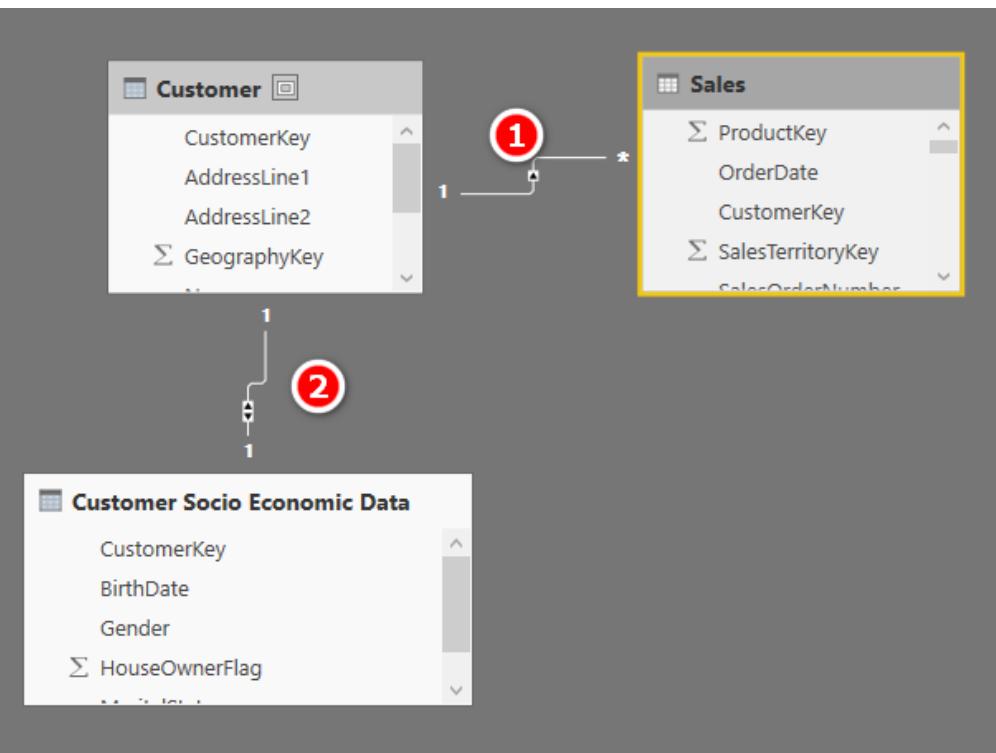


Many-to-Many





One to One Relationships



- (1) is a **One-to-Many** relationship between the Customer table (Lookup table) and the Sales table (Data table).
- (2) The Customer Socio Economic Data table is joined to the Customer table via a 1 to 1 relationship. If there is a benefit (to the user of reports) of splitting this Socio-Economic Data into a separate table then do it. Otherwise, highly recommend to combine all the data one table using PQ.

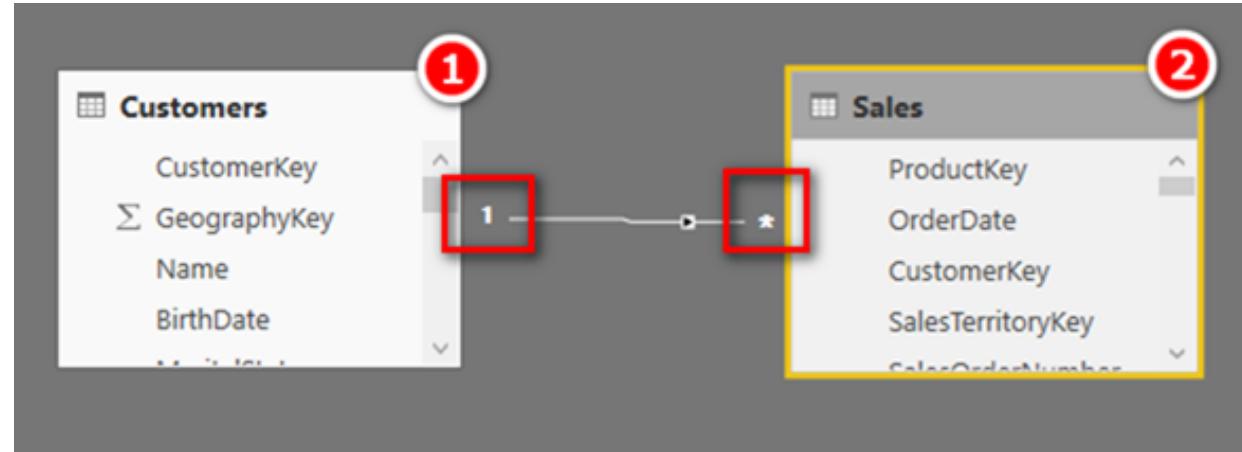
Every relationship has a “Cost”. The performance impact may not be noticeable for simple models but may become an issue with very complex models.

Don’t automatically accept the table structure coming from your source data.

As a data modeler, making decisions on the best way to load data. Your source system is probably not optimized for reporting (unless it is a reporting DataMart).



One to Many Relationships



The One-to-Many Relationship is the foundation of Power Pivot.

The “Customer” Table is on the 1 side of the relationship and the “Sales” Table is on the many side of the relationship. These tables are linked using a common field/column called “Customer Key”. Customer Key is a code that uniquely identifies each customer. **There can be no duplicates of the customer key in the customer table.** Conversely the **customer can purchase as many times as needed and hence the customer key can appear in the Sales table as many times as necessary.** This is where the name “one to many” comes from – the customer key occurs once and only once in the Customers table but can appear many times in the Sales table. Tables on the one side of the relationship are called Dimension tables (I call them Lookup tables) and the tables on the many side of the relationship are called Fact tables (I call them Data tables).



- Fact tables and dimension tables
- Schema
- Relationships
- Cardinality
- **Cross Filter Direction**
- Hierarchies



CROSS FILTER DIRECTION

When you create a relationship between two tables you have the choice of bidirectional cross filters or single direction cross filters.

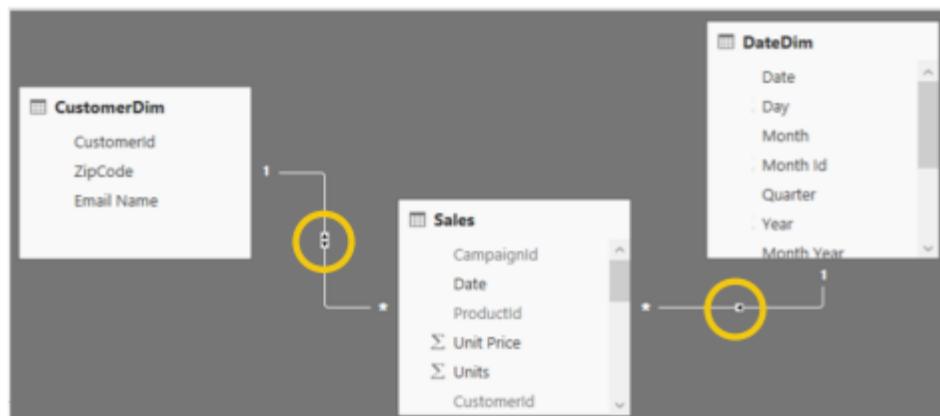
- **Single direction:** If you filter records in the table on the one side of a relationship, the filtering choices are carried through to the table on the many side of the relationship. However, if you filter records in the table on the many side of the relationship, these are not carried through to the table on the one side of the relationship.
- **Both:** Unlike single direction cross filters, these filters flow in both directions. So, if you filter records in the table on either side of the relationship, they will be carried across to the table on the other side of the relationship. For filtering purposes, both tables in the relationship are treated like they are a single table. However, with **bidirectional cross filters**, it is possible to create an ambiguous set of relationships, especially when you have a complex pattern of tables. Because of this, you **should avoid using bidirectional filters where possible**.



Filter Context and Multiple Tables



Filter Context and Multiple Tables

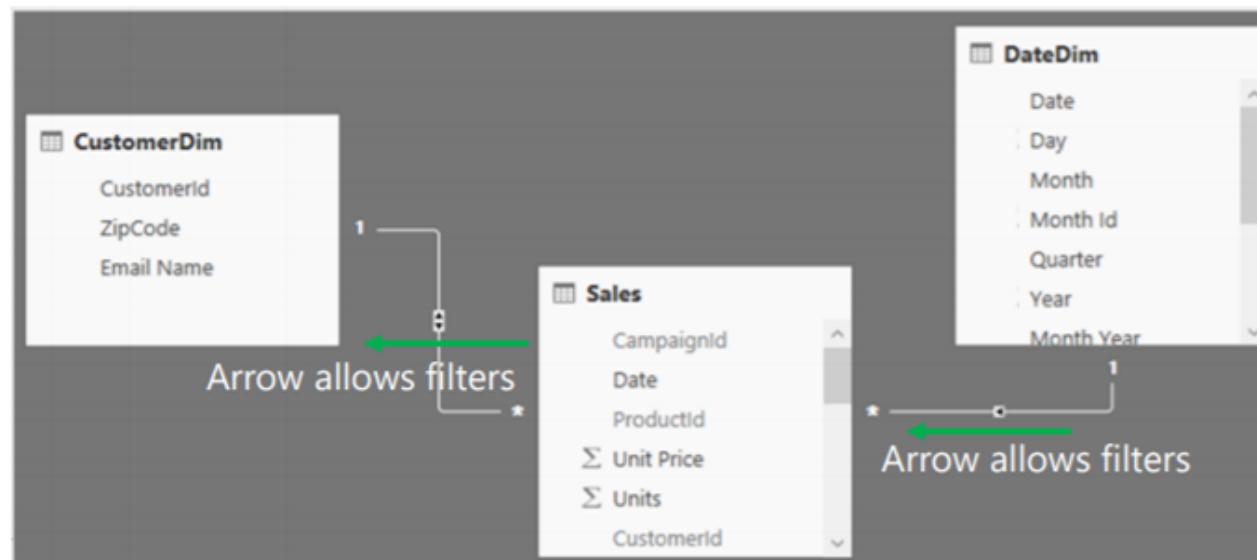


- Filters (Filter context) automatically propagate based on direction of arrows in relationships
- Examples
 - Filter goes from DateDim to CustomerDim
 - Filter does not go from CustomerDim to DateDim

Filter Context and Multiple Tables



Filter Context and Multiple Tables – Right Arrow Direction



Cross filtering works properly

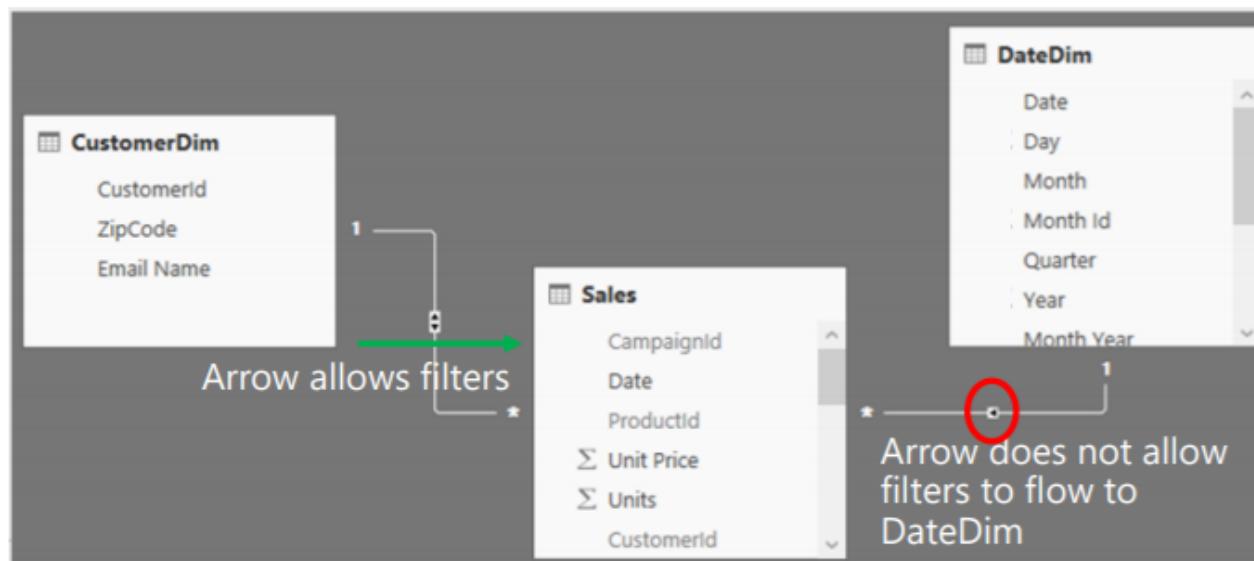
Month	Total Sales M	Count of CustomerId
Jan	\$1,673,394.03	7132
Feb	\$431,531.13	2820
Mar	\$690,671.10	4017
Apr	\$852,018.76	4629
May	\$972,018.47	5185
Jun	\$907,703.04	4854
Jul	\$608,678.35	3680
Aug	\$1,355,530.22	6242
Sep	\$720,851.83	4186
Oct	\$1,117,087.73	5728
Nov	\$2,372,763.71	8242
Dec	\$2,003,261.11	7683
Total	\$13,705,509.48	10000

- Filter goes from DateDim to CustomerDim
- This is why the above Pivot table works

Filter Context and Multiple Tables



Filter Context and Multiple Tables – Wrong Arrow Direction



Cross filtering
does not work

CustomerId	Total Sales M	Count of Month
	\$1,985.76	12
00001	\$438.34	12
00002	\$840.08	12
00003	\$1,246.69	12
00004	\$706.23	12
00005	\$1,653.97	12
00006	\$2,170.10	12
00007	\$2,308.44	12
00008	\$1,517.34	12
00009	\$1,184.11	12
00010	\$2,221.02	12
00011	\$1,646.48	12
Total	\$13,705,509.48	12

- Filter goes from DateDim to CustomerDim
- This is why the Count of Month in the table above is incorrect



Be careful with bi-directional relationships

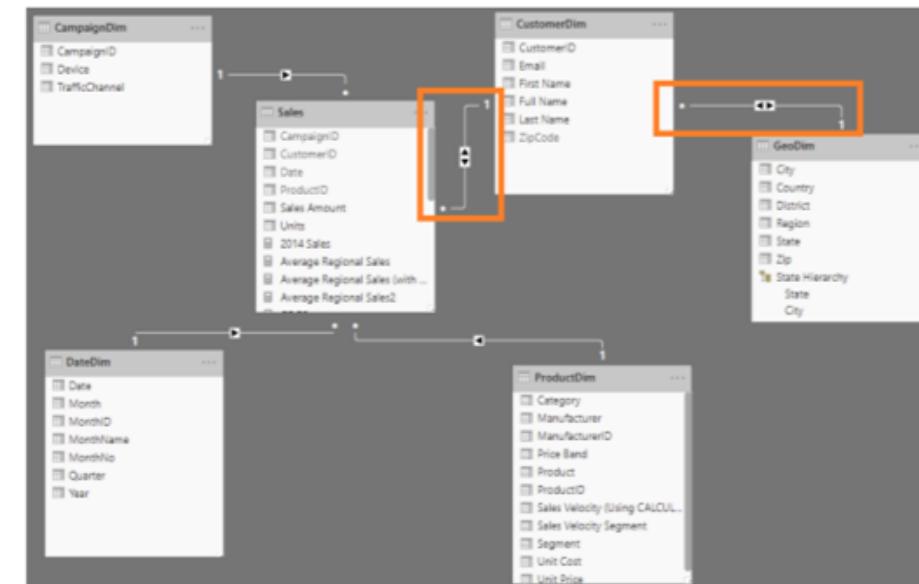


Scenario

- Most relationships in the model are set to bi-directional

Why is it undesired?

- Applying filters/slicers traverses many relationships and can be slower
- Some filter chains unlikely to add business value



Proposed Solution

- Only use bi-di where the business scenario requires it



- Fact tables and dimension tables
- Schema
- Relationships
- Cardinality
- Cross filter direction
- **Hierarchies**



Hierarchies

A hierarchy is a set of **nested columns** that are **grouped together** in a way that allows you to drill up and down a report visual using a single object from the field list. A **typical example** of a hierarchy is usually found in a **date table**, where a date hierarchy might consist of the year, month, week, and day fields. This could then be used with a report visual, where it would allow you to aggregate data by these values, giving you the ability to drill up and down by them.

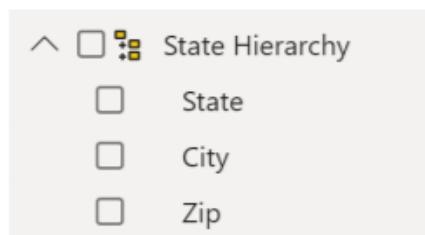
Hierarchies in Power BI

- Power BI lets you create your own hierarchies
- Hierarchies can drill up and down on your data points
- Various operations can be performed
- Dates are a unique type of hierarchy



Hierarchies

- Power BI generates Date hierarchies when dates are added to visuals, this allows the end user to drill from Year, Quarter, Month & Day.
- Users can also create custom hierarchies in the model by dragging a lower level field onto the parent.

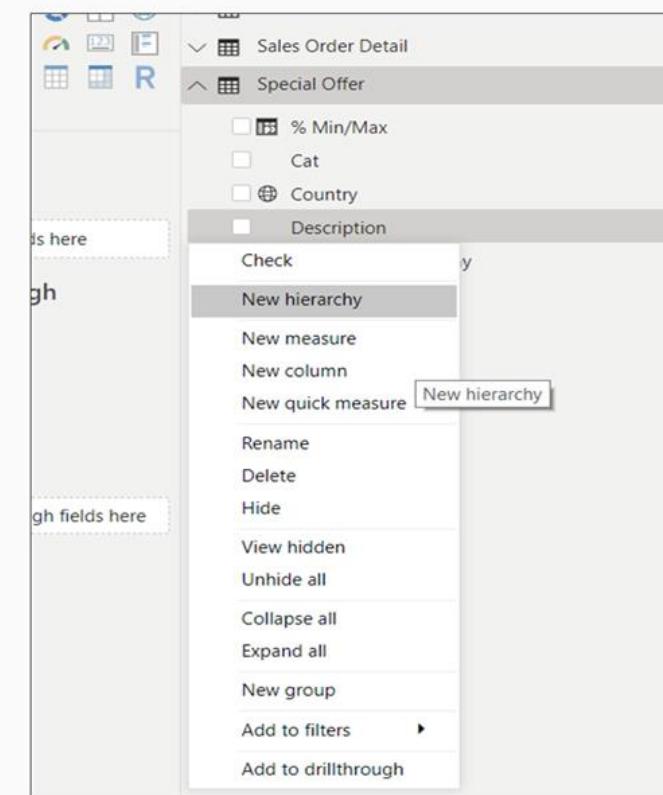


Date	Month	Σ MonthID	MonthName	Σ MonthNo	Quarter	Σ Year
Year						
Quarter						
Month						
Day						



Creating Your Own Hierarchy

- There are two ways to create a hierarchy:
 - Drag and drop one field on the another
 - Select the field that you want to add and right-click on it; this will open the context menu, where you can either create a new hierarchy or add to the existing hierarchy



DATA MODEL

Hierarchies



Continent Name	Total Sales	Total Expenses	Net Profit	Net Quantity	Order Count	Average Order Amount
Asia	\$703,151,234	\$322,629,302	\$380,521,932	3,107,004	173,110	\$4,062
Armenia	\$9,556,092	\$4,396,610	\$5,159,482	39,548	2,731	\$3,499
Armenia	\$9,556,092	\$4,396,610	\$5,159,482	39,548	2,731	\$3,499
Australia	\$29,467,554	\$13,485,077	\$15,982,477	126,150	8,226	\$3,582
Molonglo	\$9,384,501	\$4,330,834	\$5,053,666	42,746	2,712	\$3,460
New South Wales	\$20,000,000	\$10,000,000	\$10,000,000	20,000	10,000	\$3,642
Bhutan					45	\$3,602
Thimphu District					45	\$3,602
China					83	\$4,464
Beijing					15	\$4,619
GuangDong					29	\$3,967
Hong Kong					35	\$3,578
Shanghai					57	\$3,471
Xinjiang					47	\$3,620
India					60	\$3,552
Maharashtra					98	\$3,603
National Capital Territory					75	\$3,559
West Bengal						
Iran	\$19,507,129	\$8,930,136	\$10,576,993	80,534	5,496	\$3,549
Tehran	\$19,507,129	\$8,930,136	\$10,576,993	80,534	5,496	\$3,549
Japan	\$61,569,438	\$28,198,201	\$33,371,237	274,662	17,353	\$3,548
Chubu	\$4,554,969	\$2,098,597	\$2,456,372	22,868	1,235	\$3,688
Hokkaido	\$8,418,490	\$3,880,760	\$4,537,730	34,939	2,242	\$3,755
Kansai	\$19,956,875	\$9,069,686	\$10,887,189	87,819	5,654	\$3,530

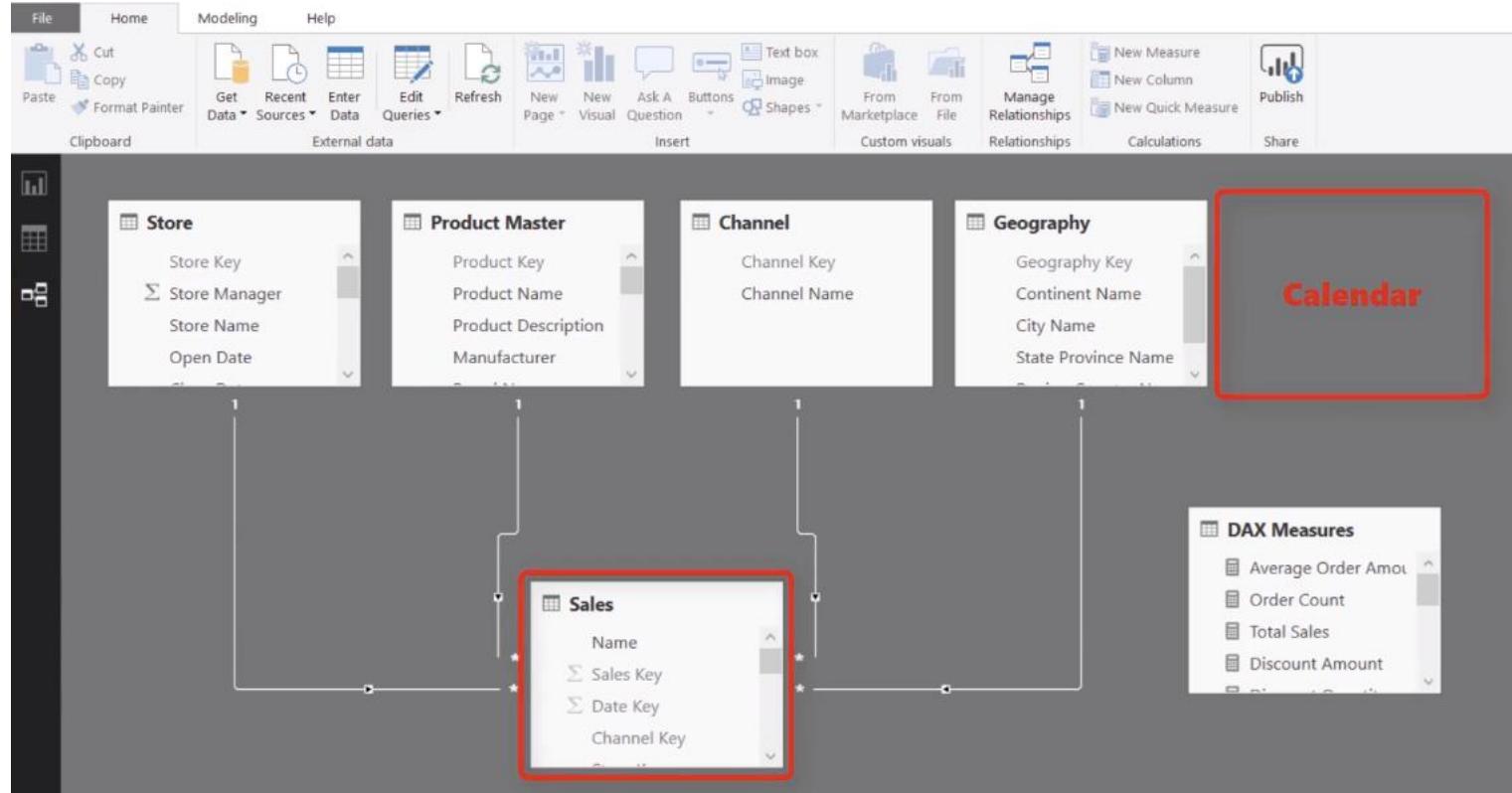
Power BI updated their hierarchy buttons in 2018. Current hierarchy buttons look like this. Same functionality as before, but a new square button design, and slightly re-ordered.

The screenshot shows the Power BI 'Visualizations' pane on the left and the 'Fields' pane on the right. In the 'Fields' pane, the 'Style' section is open, displaying various options like 'Default', 'Minimal', 'None', 'Bold header', etc. A red arrow points to the 'Minimal' option, which is highlighted with a yellow background. The 'Fields' pane also lists several DAX measures and fields, with checkboxes indicating their current state.

Field	Type	Status
Average Order A...	DAX Measure	Selected
Discount Amount	DAX Measure	Unselected
Discount Quantity	DAX Measure	Unselected
Net Profit	DAX Measure	Selected
Net Quantity	DAX Measure	Selected
Order Count	DAX Measure	Selected
Return Amount	DAX Measure	Unselected
Return Quantity	DAX Measure	Unselected
Sales Quantity	DAX Measure	Unselected
Total Cost	DAX Measure	Unselected
Total Expenses	DAX Measure	Selected
Total Sales	DAX Measure	Selected
Channel	Geography	Unselected
Geography	Geography	Unselected
Product Master	Product Master	Unselected
Sales	Sales	Unselected
Store	Store	Unselected
Address Line 1	Address	Unselected
Address Line 2	Address	Unselected
Close Date	Date	Unselected
Employee Count	Count	Unselected
Manager Full Na...	Name	Unselected
Open Date	Date	Unselected
Selling Area Size	Size	Unselected
Store Manager	Name	Unselected
Store Name	Name	Selected

DATA MODEL

Duplicate vs Reference Tab
New Table Calendar in Model



DAX can help us create a Calendar table that we'll anchor to our sales data that will be able to grab the start date for sales and the end date for sales from our Sales table, and then create a Calendar table for exactly that range.

DATA MODEL

Duplicate vs Reference Tab
New Table Calendar in Model



Screenshot of the Power BI Modeling tab ribbon.

Clipboard content: `Calendar = CALENDAR(MIN(Sales[Date]), MAX(Sales[Date]))`

Toolbox dropdown (highlighted): `MAX(ColumnOrScalar1, [Scalar2])`
Description: Returns the largest numeric value in a column, or the larger value between two scalar expressions. Ignores logical values and text.

Toolbox items visible:

- [Date]
- [Day]
- [Month]
- [MonthNo]
- [Quarter]
- [QuarterNo]
- [Year]

DATA MODEL

Duplicate vs Reference Tab
New Table Calendar in Model



The screenshot shows the Power BI Data Model ribbon. The 'Add Column' tab is highlighted with a red box and a number '1'. Below it, the 'From Selection' button is also highlighted with a red box and a number '3'. A green arrow points from the 'Reference' option in the context menu of the 'Queries [1]' pane to the 'From Selection' button.

Queries [1]

- DB Connection
- Copy
- Paste
- Delete
- Rename
- Enable load
- Include in report refresh
- Duplicate
- Reference** (highlighted with a green arrow)
- Move To Group
- Move Up
- Move Down
- Create Function...
- Convert To Parameter
- Advanced Editor
- Properties...

File Home Transform Add Column View Tools Help

Conditional Column Index Column Duplicate Column

Column From Examples Custom Invoke Custom Function

From All Columns General From Selection

Merge Columns ABC Extract Statistics Standard Scientific Information

Format ABC Parse From Text From Number

Trigonomer 10² .00 Rounding

123 Parse

10² .00 Rounding

From Text From Number

From Number

transactions Table

	1 ² ₃ trans_id	1 ² ₃ date
1	695247	930101
2	171812	930101
3	207264	930101
4	1117247	930101
5	579373	930102
6	771035	930102
7	452728	930103
8	725751	930103

DATA MODEL

Duplicate vs Reference Tab
New Table Calendar in Model



The 3rd way: M Language

Query Settings

```
= List.Dates(#"Calendar Start Date", Number.From(#"Calendar End Date" - #"Calendar Start Date")+1, #duration(1,0,0,0))
```

PROPERTIES

- Name: DIM Calendar
- All Properties

APPLIED STEPS

- ListDates
- Converted to Table
- Renamed Columns
- Changed to Date
- ...Added Calendar Column...
- Inserted Year
- Added Year Short
- Inserted Month
- Inserted Month Name
- Inserted Month Name Short
- Inserted Month & Year
- Inserted Month & Year Sh...
- Inserted Quarter
- Inserted Quarter Name
- ...Added Fiscal Columns...
- Added Fiscal Month
- Added Fiscal Month Name
- Added Fiscal Quarter
- Added Fiscal Quarter Name
- Added Fiscal Year
- Added Fiscal Year Name
- ...Added OFFSET Columns...
- Inserted Day Offset
- Inserted Month Offset
- Inserted Week Offset
- Inserted Quarter Offset
- Inserted Year Offset
- Inserted Fiscal Year Offset

Query Settings

```
= Table.AddColumn(#"Inserted Day", "Day Name", each Date.DayOfWeekName([Date]), type text)
```

PROPERTIES

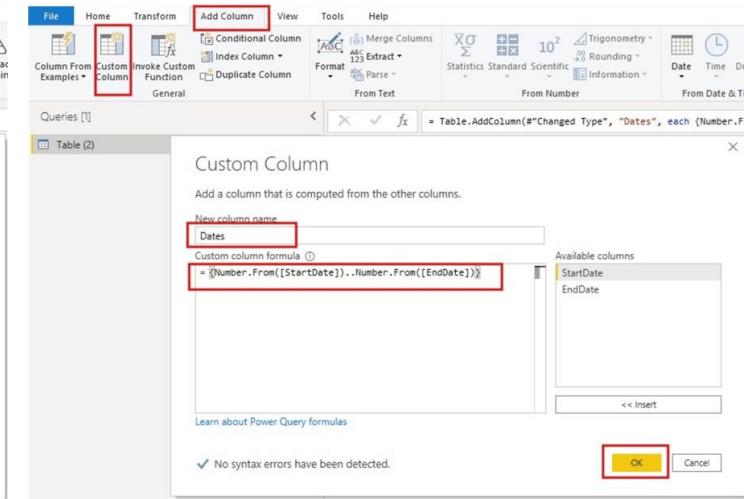
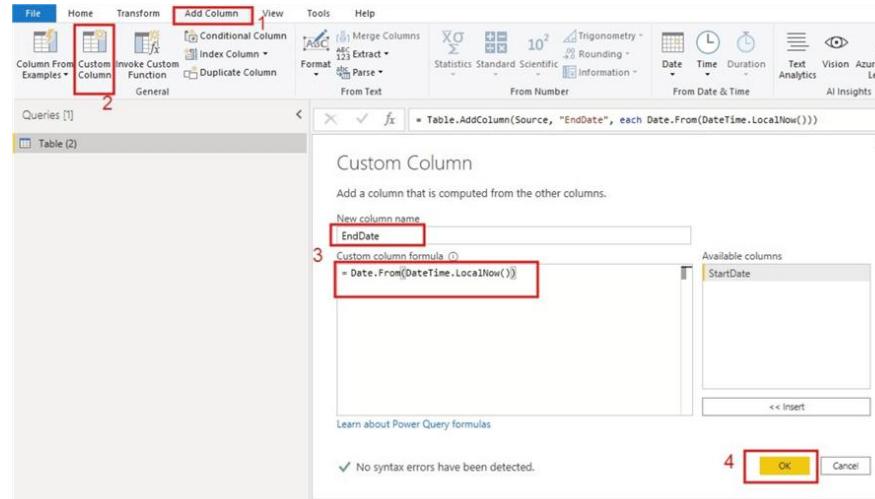
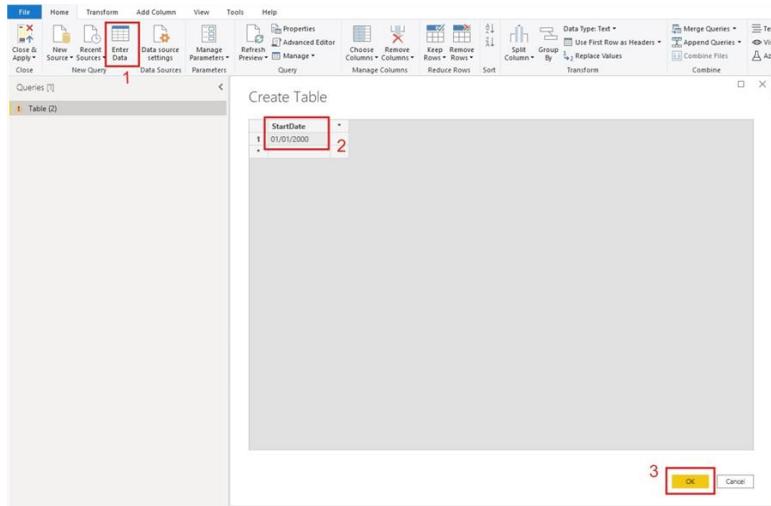
- Name: DIM Calendar
- All Properties

APPLIED STEPS

- Inserted Month Name
- Inserted Month Name Short
- Inserted Month & Year
- Inserted Month & Year Sh...
- Inserted Quarter
- Inserted Quarter Name
- ...Added Fiscal Columns...
- Added Fiscal Month
- Added Fiscal Month Name
- Added Fiscal Quarter
- Added Fiscal Quarter Name
- Added Fiscal Year
- Added Fiscal Year Name
- ...Added OFFSET Columns...
- Inserted Day Offset
- Inserted Month Offset
- Inserted Week Offset
- Inserted Quarter Offset
- Inserted Year Offset
- Inserted Fiscal Year Offset
- ...Added General Column...
- Inserted Day
- Inserted Day Name
- Inserted Day of Week
- Inserted Weekend Flag
- Inserted End of Week
- Inserted Future/Past Flag
- Filtered Rows

DATA MODEL

Duplicate vs Reference Tab New Table Calendar in Model



	StartDate	EndDate	Dates
1	01/01/2000	04/01/2021	36526
2	01/01/2000	04/01/2021	36527
3	01/01/2000	04/01/2021	36528
4	01/01/2000	04/01/2021	36529
5	01/01/2000	04/01/2021	36530
6	01/01/2000	04/01/2021	36531
7	01/01/2000	04/01/2021	36532
8	01/01/2000	04/01/2021	36533
9	01/01/2000	04/01/2021	36534
10	01/01/2000	04/01/2021	36535
11	01/01/2000	04/01/2021	36536
12	01/01/2000	04/01/2021	36537
13	01/01/2000	04/01/2021	36538
14	01/01/2000	04/01/2021	36539
15	01/01/2000	04/01/2021	36540
16	01/01/2000	04/01/2021	36541
17	01/01/2000	04/01/2021	36542
18	01/01/2000	04/01/2021	36543
19	01/01/2000	04/01/2021	36544
20	01/01/2000	04/01/2021	36545
21	01/01/2000	04/01/2021	36546
22	01/01/2000	04/01/2021	36547

	StartDate	EndDate	Dates
1	01/01/2000	04/01/2021	01/01/2000
2	01/01/2000	04/01/2021	02/01/2000
3	01/01/2000	04/01/2021	03/01/2000
4	01/01/2000	04/01/2021	04/01/2000
5	01/01/2000	04/01/2021	05/01/2000
6	01/01/2000	04/01/2021	06/01/2000
7	01/01/2000	04/01/2021	07/01/2000
8	01/01/2000	04/01/2021	08/01/2000
9	01/01/2000	04/01/2021	09/01/2000
10	01/01/2000	04/01/2021	10/01/2000
11	01/01/2000	04/01/2021	11/01/2000
12	01/01/2000	04/01/2021	12/01/2000
13	01/01/2000	04/01/2021	13/01/2000
14	01/01/2000	04/01/2021	14/01/2000
15	01/01/2000	04/01/2021	15/01/2000
16	01/01/2000	04/01/2021	16/01/2000
17	01/01/2000	04/01/2021	17/01/2000
18	01/01/2000	04/01/2021	18/01/2000
19	01/01/2000	04/01/2021	19/01/2000
20	01/01/2000	04/01/2021	20/01/2000



Hide Columns and Tables

- IDs and surrogate keys are needed for relationships, but are not useful for reporting
- Hiding columns and tables that are not useful for reporting simplifies the data model
- Avoid importing a column or a table if it is not needed in calculation, sorting, or defining a relationship
- This is extremely simple, but necessary

DATA MODEL

Search Tab



Hide all key Field in the relationship and that's going to give us a nice shortened list of columns to hide all of our key columns for relationships. So, we are going to quickly just run through and hide all of these.

The screenshot shows a data modeling interface with two main panels. On the left is the 'FIELDS' search tab, which includes sections for 'VISUALIZATIONS', 'FILTERS', 'DRILLTHROUGH', and a search bar. On the right is a data grid displaying a table with columns: Return Quantity, Return Amount, Discount Quantity, Discount Amount, Total Cost, Sales Amount, Geography Key, and Date. The data grid contains 20 rows of sample data. To the right of the data grid is a 'FIELDS' sidebar listing various dimensions and their keys, such as Channel, Geography, Product Master, Sales, and Store, with the 'Geography Key' highlighted in yellow.

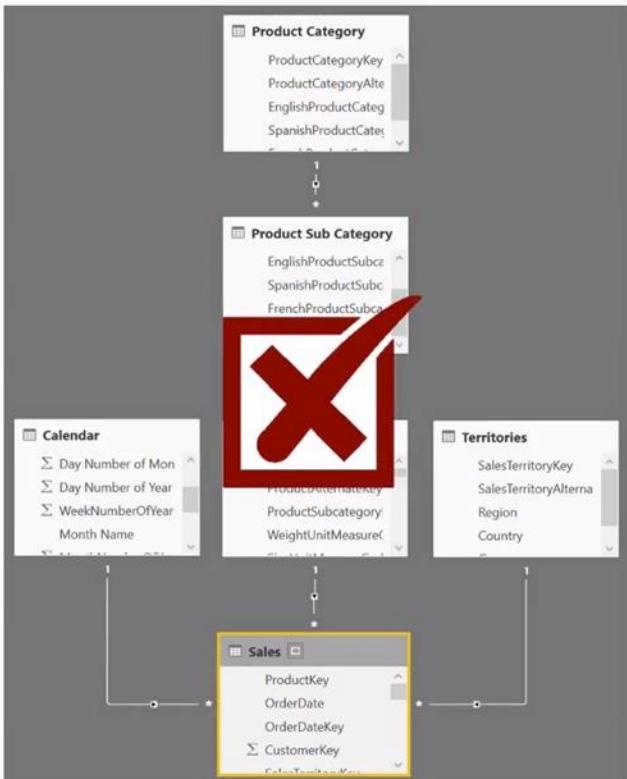
Return Quantity	Return Amount	Discount Quantity	Discount Amount	Total Cost	Sales Amount	Geography Key	Date
0	0	0	0	2107.2	6359.9	804	4/1/2007
0	0	0	0	2107.2	6359.9	946	4/1/2007
0	0	0	0	2107.2	6359.9	789	4/1/2007
0	0	0	0	2107.2	6359.9	944	4/2/2007
0	0	0	0	2107.2	6359.9	941	4/2/2007
0	0	0	0	2107.2	6359.9	810	4/3/2007
0	0	0	0	2107.2	6359.9	839	4/3/2007
0	0	0	0	2107.2	6359.9	892	4/4/2007
0	0	0	0	2107.2	6359.9	551	4/5/2007
0	0	0	0	2107.2	6359.9	875	4/5/2007
0	0	0	0	2107.2	6359.9	934	4/5/2007
0	0	0	0	2107.2	6359.9	846	4/6/2007
0	0	0	0	2107.2	6359.9	938	4/7/2007
0	0	0	0	2107.2	6359.9	836	4/7/2007
0	0	0	0	2107.2	6359.9	897	4/7/2007
0	0	0	0	2107.2	6359.9	920	4/7/2007
0	0	0	0	2107.2	6359.9	949	4/7/2007
0	0	0	0	2107.2	6359.9	840	4/8/2007
0	0	0	0	2107.2	6359.9	951	4/9/2007
0	0	0	0	2107.2	6359.9	922	4/9/2007
0	0	0	0	2107.2	6359.9	829	4/9/2007
0	0	0	0	2107.2	6359.9	877	4/9/2007
0	0	0	0	2107.2	6359.9	693	4/10/2007
0	0	0	0	2107.2	6359.9	825	4/11/2007
0	0	0	0	2107.2	6359.9	862	4/12/2007



- **Merge Table in Power Query**
- Naming Convention
- VertiPaq Engine
- Optimize Your File tips



Why do we Want to Merge Tables?



Snowflake Schema



Star Schema

DATA MODEL

Merge



"}],"Product Sub Category",JoinKind.LeftOuter)

ABC 123 WeightUnitMeasureID ABC 123 UnitCost ABC 123 UnitPrice Product Sub Category

	ABC 123 WeightUnitMeasureID	ABC 123 UnitCost	ABC 123 UnitPrice	Product Sub Category
4.5	ounces			
5.6	ounces			
7.4	ounces			
11	ounces			
1	pounds			
11	ounces			
2	ounces			
8.8	ounces			
2.9	pounds			
5	ounces			
1	pounds			
30	grams			

Search Columns to Expand A Z

Expand Aggregate

(Select All Columns)
 Product Subcategory Key
 Product Subcategory Name
 Product Category Key

Use original column name as prefix

OK Cancel

/6.45 149.95 Table

QUERY SETTINGS

PROPERTIES

Name

Product

All Properties

APPLIED STEPS

Source

Navigation

Removed Other Columns

Merged w/ Product Sub Cate...

Renamed Columns

Changed Type

DATA MODEL

Merge



Combine

: Category", {"Product Category Key"}, "Product Category", JoinKind.LeftOuter)

MeasureID	ABC 123	UnitCost	ABC 123	UnitPrice	A ^B C Product Subcategory Name	1 ² 3 Product Category Key	Product Category
	11	21.57	MP4&MP3				
	30.58	59.99	MP4&MP3				
	35.72	77.68	MP4&MP3				
	50.56	109.95	MP4&MP3				
	61.62	134	MP4&MP3				
	91.93	199.9	MP4&MP3				
	91.93	199.9	MP4&MP3				
	84.49	255	MP4&MP3				
	48.92	95.95	MP4&MP3				
	99.14	299.23	MP4&MP3				
	106.69	232	MP4&MP3				
	76.45	149.95	Recording Pen				
	91.95	199.95	Recording Pen				
	98.07	296	Recording Pen				
	79.53	156	Recording Pen				
	83.24	181	Recording Pen				
	13.1	25.69	Bluetooth Headphones				
	22.05	47.95	Bluetooth Headphones				
	17.45	37.95	Bluetooth Headphones				
	18.65	40.55	Bluetooth Headphones				
	45.98	99.99	Bluetooth Headphones				
	49.69	149.99	Bluetooth Headphones				
	49.69	149.99	Bluetooth Headphones				
	24.26	67.4	Bluetooth Headphones				

Search Columns to Expand

Expand Aggregate

(Select All Columns)
 Product Category Key
 Product Category Name

Use original column name as prefix

OK Cancel

QUERY SETTINGS

PROPERTIES

Name: Product

APPLIED STEPS

- Source
- Navigation
- Removed Other Columns
- Merged w/ Product Sub Cate...
- Expanded Product Sub Categ...
- Merged w/ Product Category
- Renamed Columns
- Changed Type



- Merge in Power Query
- **Naming Convention**
- VertiPaq Engine
- Optimize Your File tips



Rename Columns

- You might need names that are more meaningful when designing the report
- When you load the data from a data source, you get the column names from that data source
- Use the Power BI rename column option to rename a column

Rename Tables

- When you import data from a data source, you get the default table names of the data source
- You might need table names that are more meaningful and easier to understand
- Use the Power BI rename table option to rename a table

DATA MODEL

Naming Convention



The screenshot shows the 'Choose Columns' dialog box. At the top, there's a toolbar with various icons for managing columns and rows. A red box highlights the 'Choose Columns' icon. Below the toolbar, the title 'Choose Columns' is displayed, followed by the instruction 'Choose the columns to keep'. A search bar labeled 'Search Columns' is present. A list of column names is shown with checkboxes next to them. Most checkboxes are checked, except for 'Product Code', 'Subcategory Key', 'Subcategory Name', 'Category Key', 'Category Code', 'Product Description', 'Brand', 'Class', and 'Stock Type'. At the bottom right of the dialog are 'OK' and 'Cancel' buttons.

Column Naming

Rename all columns that will be visible in the data model using concise, self-describing, meaningful and user-friendly names.

Your data model should be **designed for users** and not for developers to design reports and consume. Even if technical professionals are designing reports, field names are going to show up as labels and titles.

Friendly Column Name	Technical Name
Customer Number	customerNumber
Customer Nbr	CustomerNbr
Customer #	CUST_NBR
Product Brand Name	productBrandName
	ProdBrandNm
	PRD_BRAND_NM



- Merge in Power Query
- Naming Convention
- **VertiPaq Engine**
- Optimize Your File Tips



Compression



Practical Example of Compression

Dashboard in a Day Class Data

Sales Fact	420.0 MB
Dimensions	4.4 MB
Int'l Sales	32.4 MB
Total Data	456.8 MB

Queries ONLY – No Data Loaded

Query Metadata	113 KB
----------------	--------

DIAD Complete Data Model

Data Model	59.4 MB
------------	---------

Almost 8X
Compression!!



Use integer surrogate keys, pre-sort them



- Power BI compresses rows in segments of millions of rows
- Integers use Run Length Encoding
- Sorting will maximize compression when encoded as it reduces the range of values per segment



- Merge in Power Query
- Naming Convention
- VertiPaq Engine
- **Optimize Your File Tips**

Options

CURRENT FILE

Data Load

Time intelligence

 Auto date/time [Learn more](#) Delivery Date Key

Name

 01-simple-dimensional-model.pbix

Size

5.216 KB

Time intelligence

 Auto date/time [Learn more](#)  Delivery Date Key

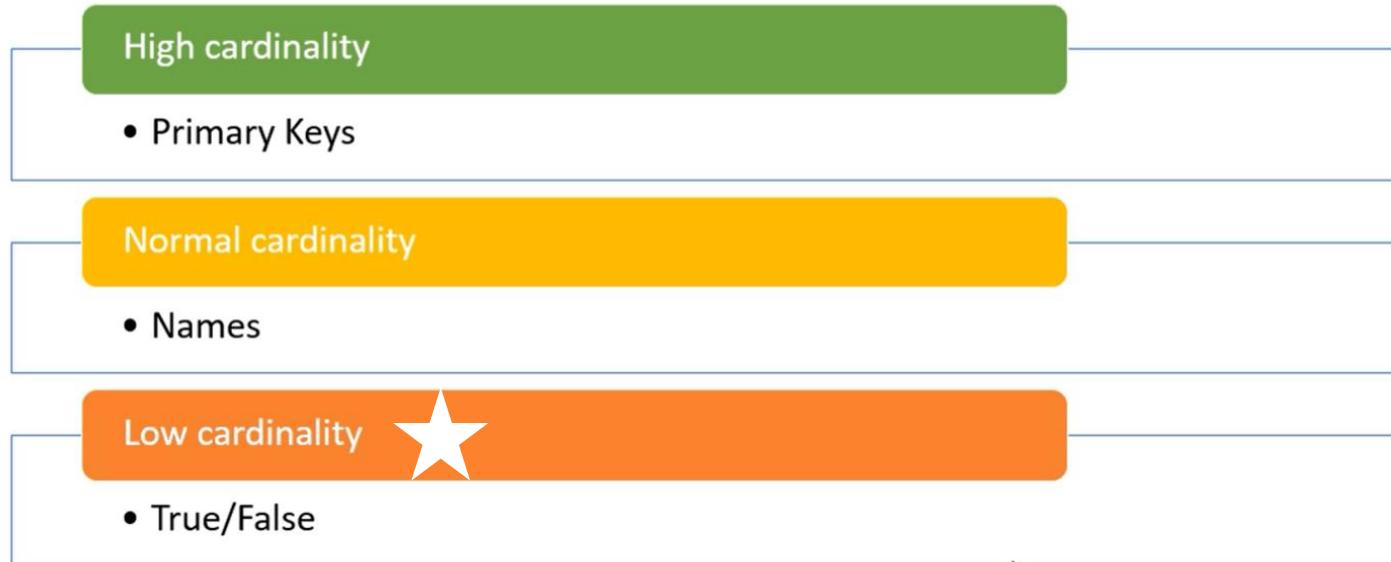
Name

 01-simple-dimensional-model.pbix

Size

31.572 KB

Improve cardinality levels through summarization and by changing data types



The **higher the cardinality** of a column, the more rows will be stored in a dictionary, and the bigger the size. Having a **few unique values**, the size of the dictionary will be much smaller, & VertiPaq will be able to achieve a much higher compression ratio.

DATA MODEL

Optimization - Lower Cardinality whenever possible



Avoid high precision/cardinality columns



Scenario

- Model contains columns at a higher precision than needed for analysis e.g. datetime in milliseconds, weight to 6 decimal places
- Model contains columns that are highly unique

Why is it undesired?

- Less compression with high precision/cardinality
- Increases time to load into memory
- Increases refresh time

Proposed Solution

- Remove if not needed
- Reduce precision
- Split datetime into date and time

DATA MODEL

Optimization - Lower Cardinality whenever possible



Power BI Date Formatting

File Home Modeling Help

Manage Relationships New Measure New Column New Table New Parameter What If Sort

Data type: Date ▾ Format: 14-Mar-01 (dd-MMM-yy) ▾ \$ ▾ % , .00 Auto ▾

Formatting

Date	Call Duration	Cost	Date.1	Time
05-Aug-19	12/31/1899 8:48:53 PM	247	Monday, August 5, 2019	9:41:00 AM
09-Aug-18	12/31/1899 3:45:41 PM	488	Thursday, August 9, 2018	7:59:02 PM
25-May-19	12/31/1899 9:28:01 AM	407	Saturday, May 25, 2019	9:35:51 AM
11-Apr-19	12/31/1899 10:59:32 PM	208	Thursday, April 11, 2019	11:31:51 PM
23-Aug-19	12/31/1899 8:53:56 PM	473	Friday, August 23, 2019	9:52:30 AM
31-Jul-18	12/31/1899 1:35:39 PM	384	Tuesday, July 31, 2018	8:56:07 PM
25-Aug-18	12/31/1899 10:21:06 PM	387	Saturday, August 25, 2018	1:29:57 AM

DATA MODEL

Optimization - Lower Cardinality whenever possible



Use integers instead of strings



Why is it undesired?

- Strings use dictionary encoding, integers use run length encoding which is more efficient

Proposed Solution

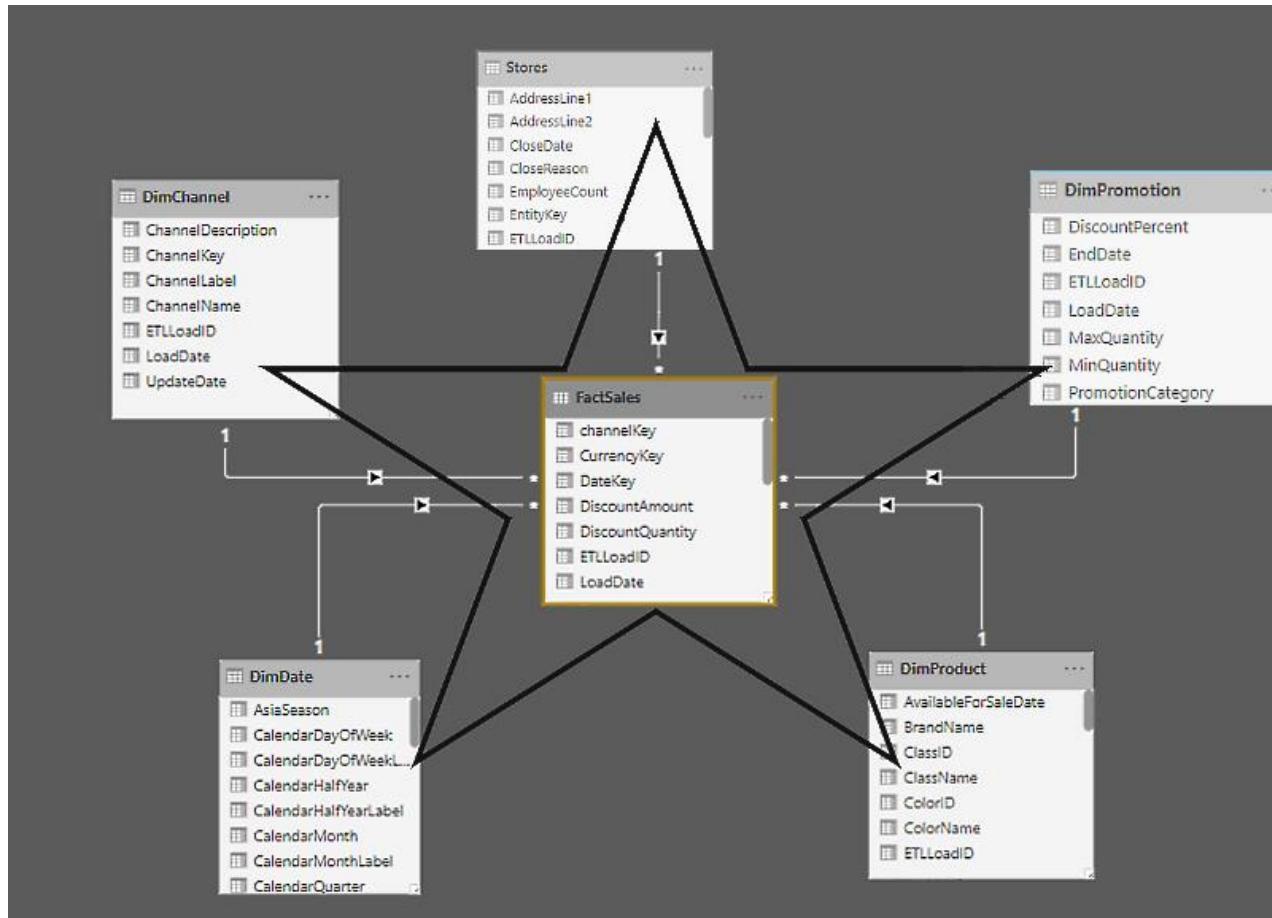
- Check data types and set to integer if known to be numerical

DATA MODEL

Optimize Your File - Optimize Schema



Keeping your data model simple





Consider subsets for very large models



Scenario

- Large model – hundreds of tables and tens of GB
- Large high grain fact tables – millions to billions

Why is it undesired?

- Aggregating/measures across large facts can affect performance
- Large models become harder to maintain and use ad-hoc

Proposed Solution

- Consider aggregations and composite models features
- Build manual summary tables with smart measures
- Create smaller models for the most common business cases



In-Memory Database



PBI – In-Memory Database

- Data stored in **RAM (in memory)**
- RAM is all electronic – **Read/Write is fast**
- Laptops have smaller **RAM space (~8GB)**

Power BI compresses data to conserve space in RAM



Data Mode Types in Power BI



How can I tell what Data Model Type I have?

- Live Connect to SQL Analysis Services (SSAS) tabular
 - Report view only available
- DirectQuery to SQL or other relational source
 - Report & Relationship views available
- Import data into Power BI (creates a copy of the data)
 - Report, Data and Relationship views available

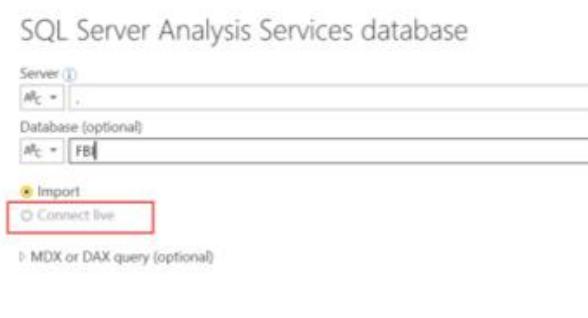




Connection: Live Connect



- Live Connect to Multidimensional or Tabular
 - On Premise or Azure
- Only a single connection will be made and all modeling is done in the cube
- You can not add relationships or additional data source
- If allowed, you can add DAX measures





Connection: DirectQuery to Relational Source



- Direct Query to SQL or other relational source
 - On Premise or Azure
- Composite modeling is possible where some data sources are in Direct Query mode and a few are in Import mode
- You can add relationships and DAX

SQL Server database

Server i
APC

Database (optional)
APC

Data Connectivity mode i
 Import
 DirectQuery

Advanced options



Import Mode



What is unique about Power BI Desktop in Import Mode?

- Columnar database
- In-memory database

Let us understand some of the internals of Power BI Desktop !!



Columnar Database



Row Based Database

First Name	Last Name	Sales
John	Smith	\$10
Jane	Doe	\$25
Hardy	B	\$35

PBI - Columnar Database

First Name	Last Name	Sales
John	Smith	\$10
Jane	Doe	\$25
Hardy	B	\$35

- Stores **each row separately** (like a separate file)
- Retrieving multiple columns from a single row is fast
- Retrieving multiple rows from a single column is slower

- Stores **each column separately** (like a separate file)
- Retrieving multiple columns from a single row is slow
- Retrieving multiple rows from a single column is faster
- **Columnar databases are well suited for analytics**



Choosing storage mode: Import vs DirectQuery



- Import is your first choice (all in memory = best speed, no DAX limits)
- When is it inappropriate
 - Extremely large data volumes
 - Need near real-time access to data from source
 - Considerable existing investment in external DW or OLAP (modelled, conformed, cleaned, calcs defined etc). SSAS MD, SAP HANA and BW are common.
 - Regulatory and data sovereignty requirements
- Considerations
 - How much source data, how compressible? Rule of thumb is 5x-10x
 - Is Premium an option? (larger datasets supported there)
 - Will blended architecture suffice? (Composite models, Aggregations for summary data)
 - Some limits on DAX in DirectQuery mode (e.g. time intelligence)



MASTERING DATA ANALYTICS



BUSINESS INTELLIGENCE

Module 2: End-to-End BI Workflow
in Power BI (Part 2)

2. End-to-End Business Intelligence Workflow in Power BI

01

Data Preparation

1. Power Query Overview
2. Get Data
3. PQ – Basic Transform Data
4. Profiling Data
5. Data Issues
 - 4a. Bad Shape + Dirty Data
 - 4b. Missing Data + Outliers
5. Combine Data from Folder
6. Blending Data
7. Checklist

02

Data Modelling

1. Data Model Overview
2. Fact & Dimension
3. Schema
4. Cardinality
5. Cross Filter Direction
6. Hierarchies

03

End-to-End in Power BI Cloud

1. Introduce PBI Ecosystem (PBI Service)
2. Fact & Dimension
3. Prep Data (On Pro & Premium)
4. Data Modeling
5. Report and Dashboard
6. Refresh Scorecard & Metrics
7. Sharing, Collaboration, (PBI Mobile)
8. Deployment Pipelines



THANK YOU