Mastering Data Analytics
Copyright Notice & Disclaimer

- *The content contained in this module is provided only for educational 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

MASTERING DATA ANALYTICS

BUSINESS INTELLIGENCE

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

MASTERING DATA ANALYTICS

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

MASTERING DATA ANALYTICS

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

4

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

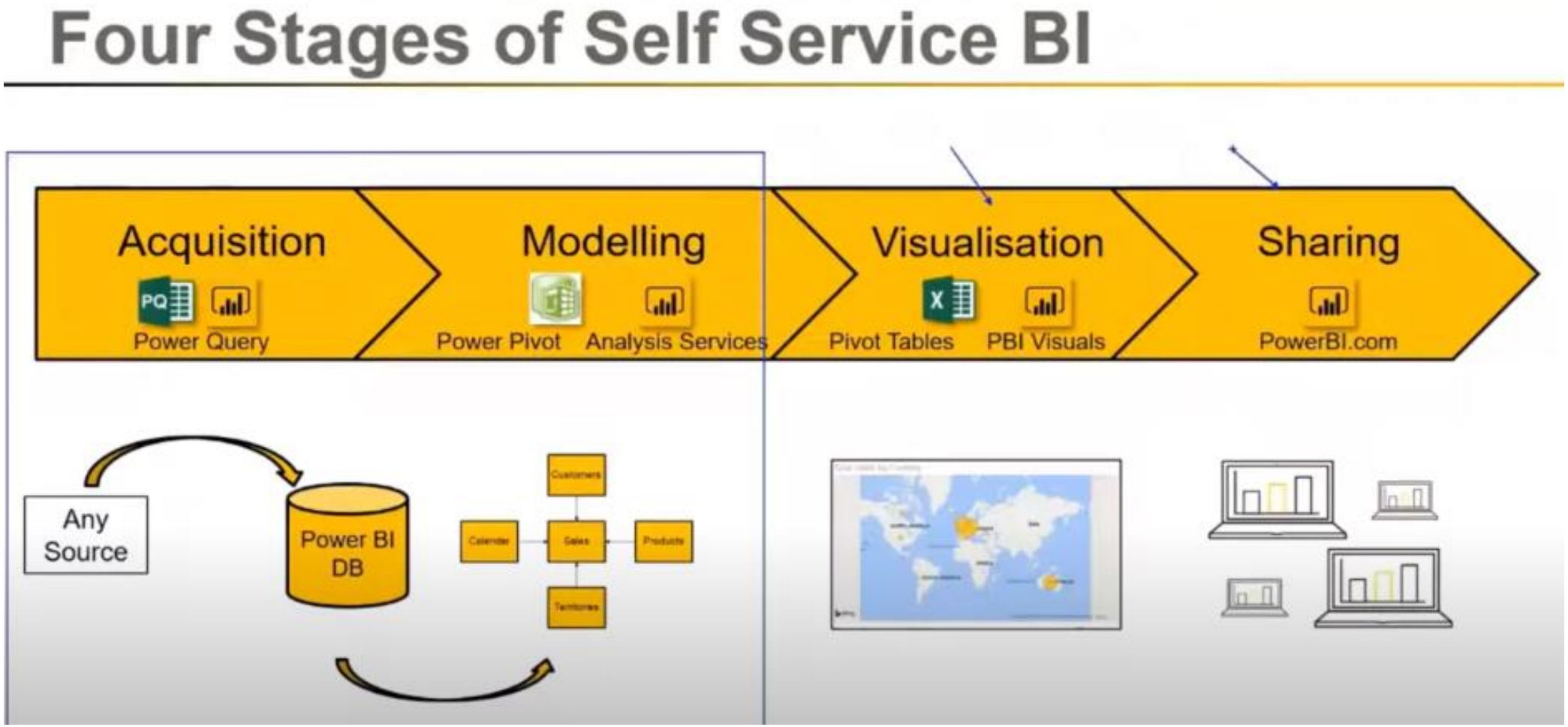
1. Data Model Overview
2. Fact & Dimension
3. Schema
4. Cardinality
5. Cross Filter Direction
6. Hierarchies

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

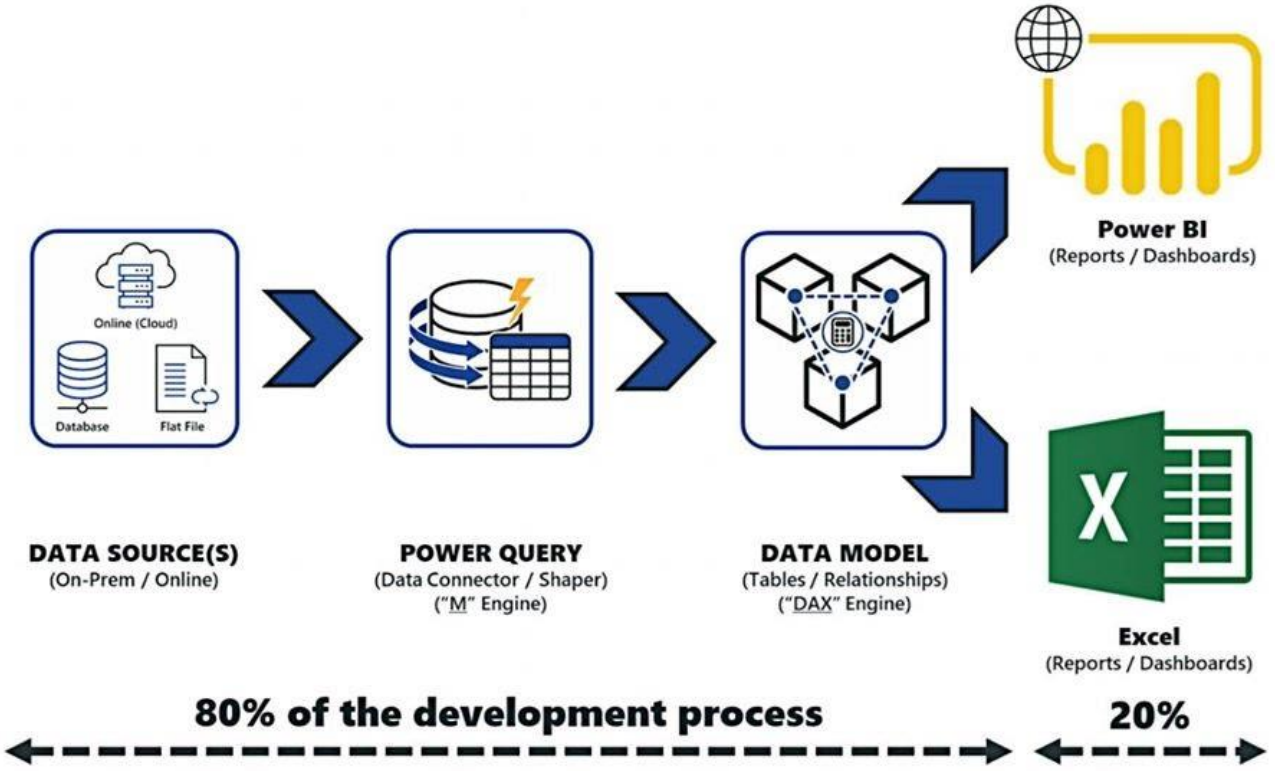
DATA PREPARATION
Power Query Overview - Self-Service BI



DATA PREPARATION
Power Query Overview

MASTERING DATA ANALYTICS

REPORT DEVELOPMENT PROCESS



POWER QUERY BENEFITS

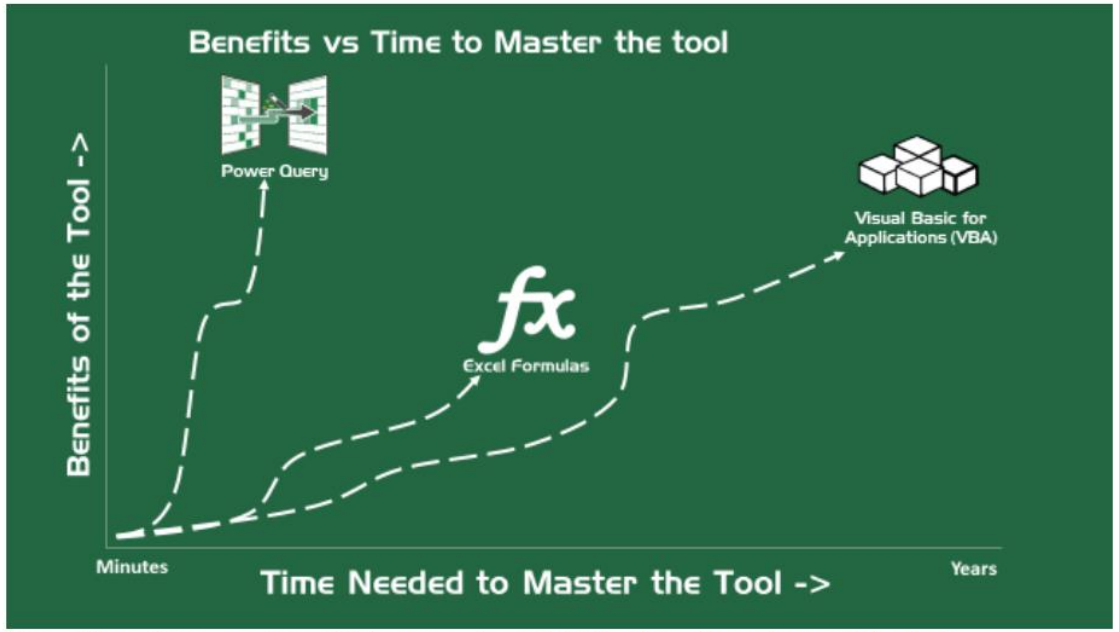
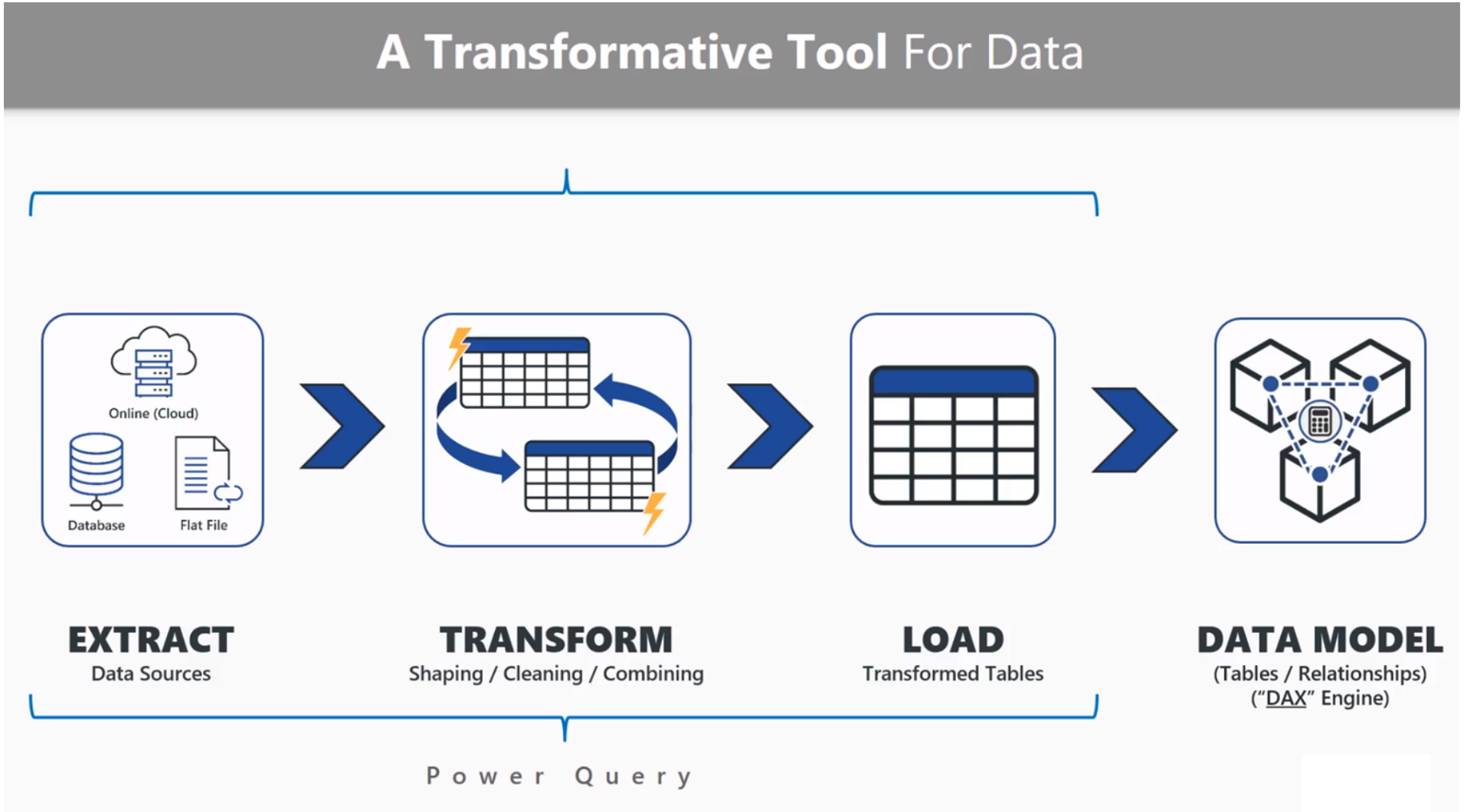


Figure 3 Power Query was designed to be an easy-to-use data transformation and manipulation tool.

DATA PREPARATION
Power Query Overview

MASTERING DATA ANALYTICS



DATA PREPARATION
Power Query Overview

MASTERING DATA ANALYTICS

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

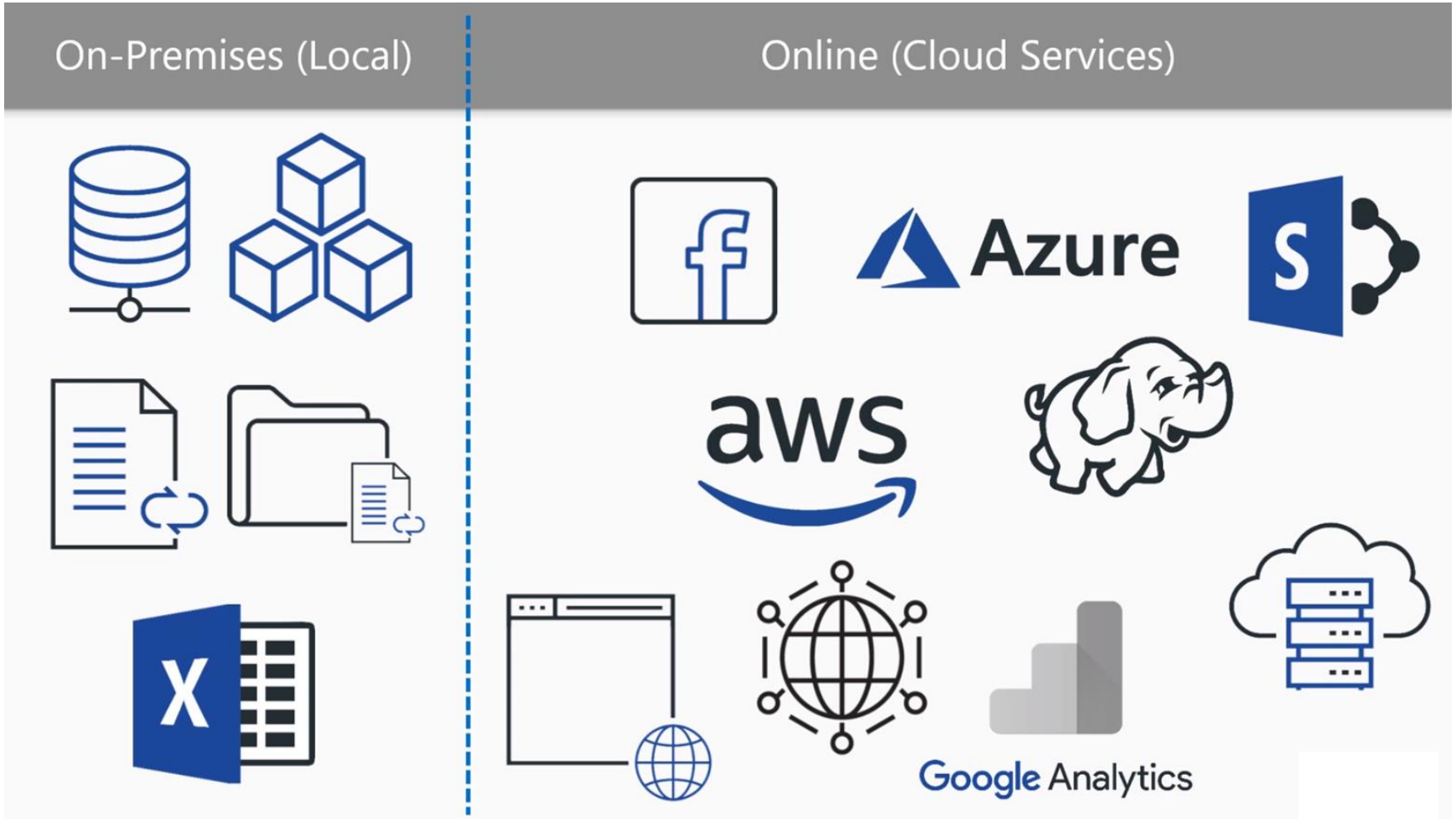
- **Extract** – Data can be extracted from a **variety of sources:** Databases, CSV files, Text files, Excel, Website and even PDF.

- **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.

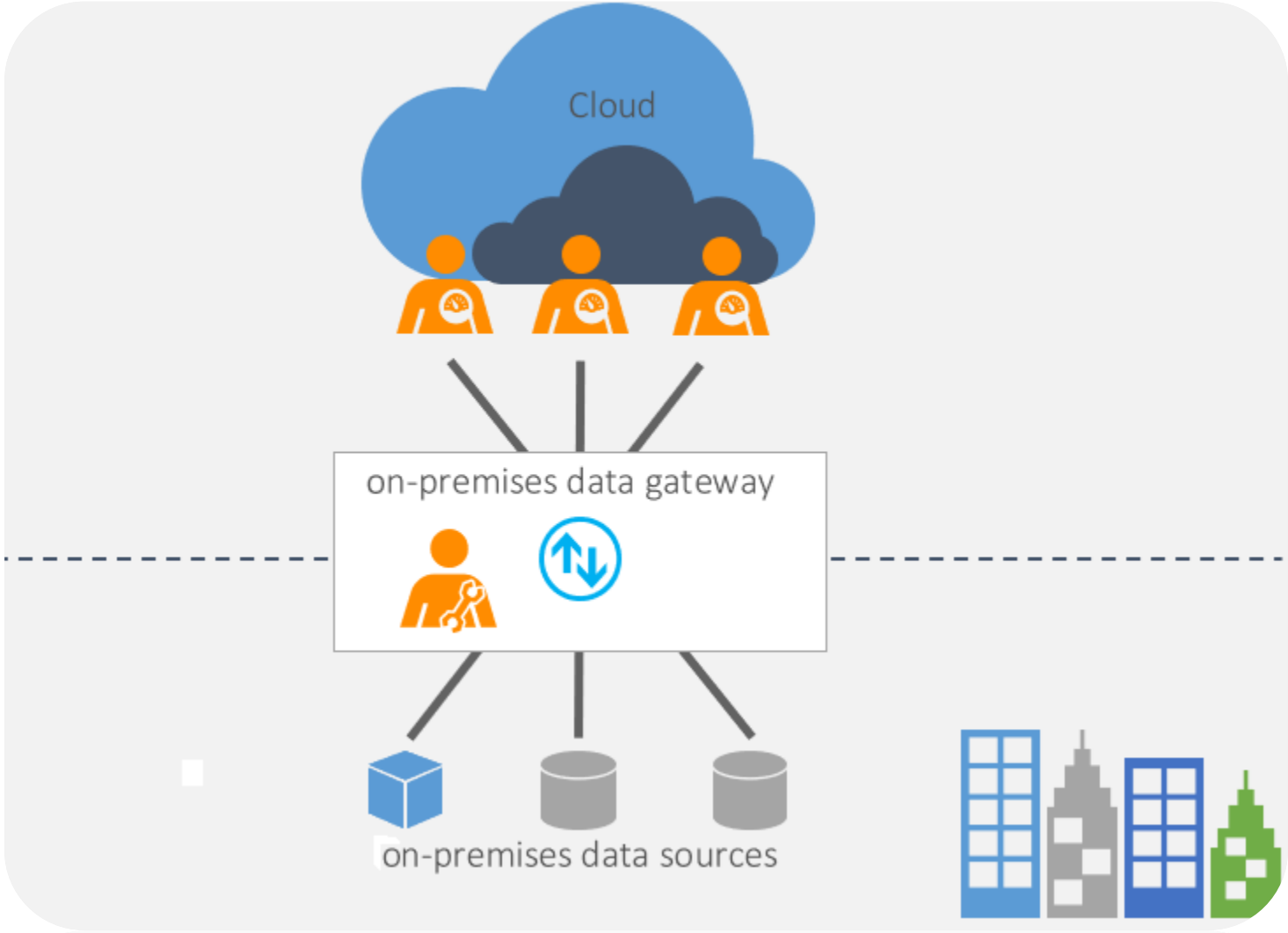
- **Load** – Once the data has been extracted and transformed, it needs to be put somewhere so that you can use it. From an Excel perspective, it can be **pushed into a worksheet, a data model, or another query.**

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.**

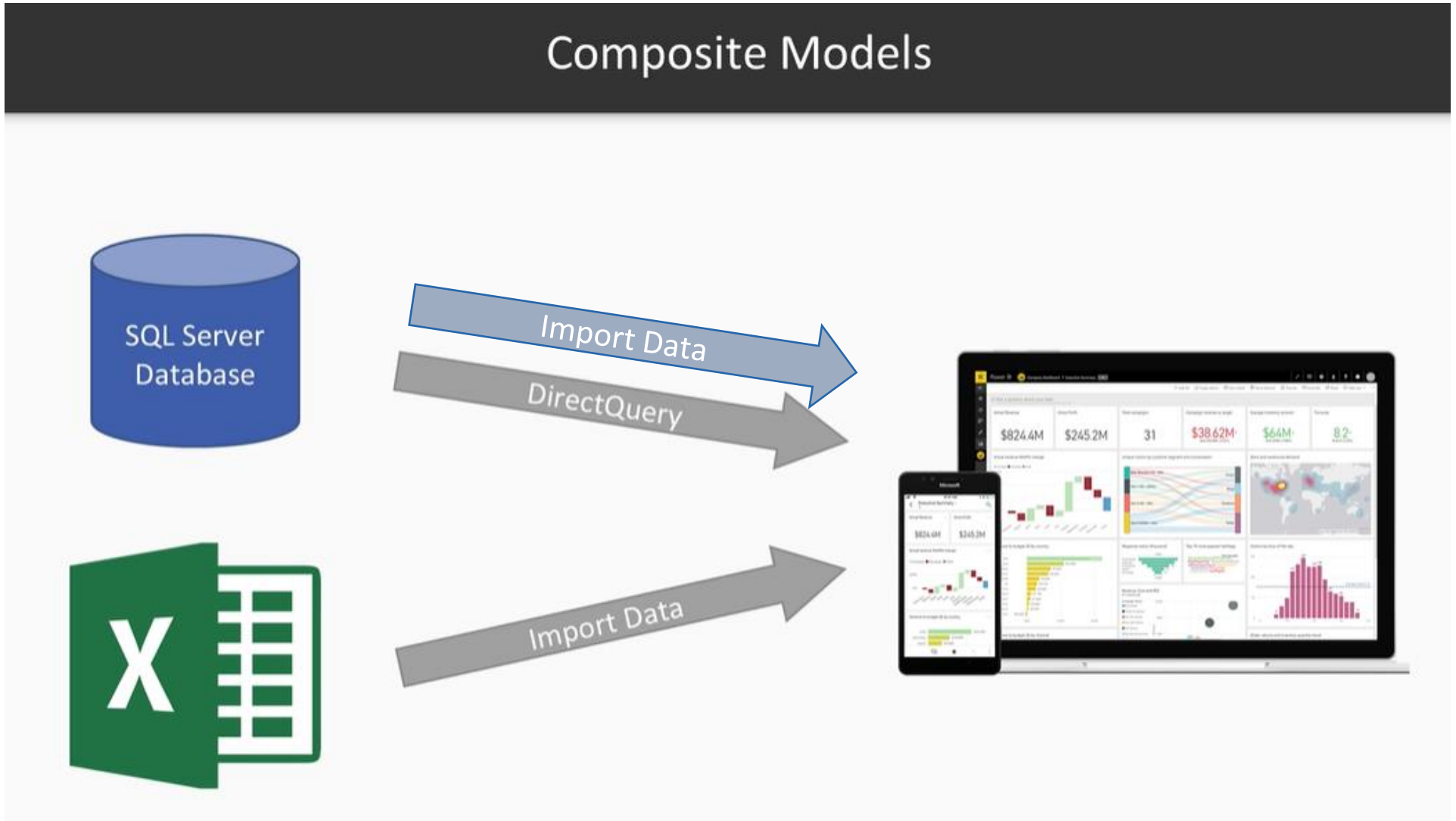
DATA PREPARATION
Get Data - Data Sources



DATA PREPARATION
Get Data - Data Gateway

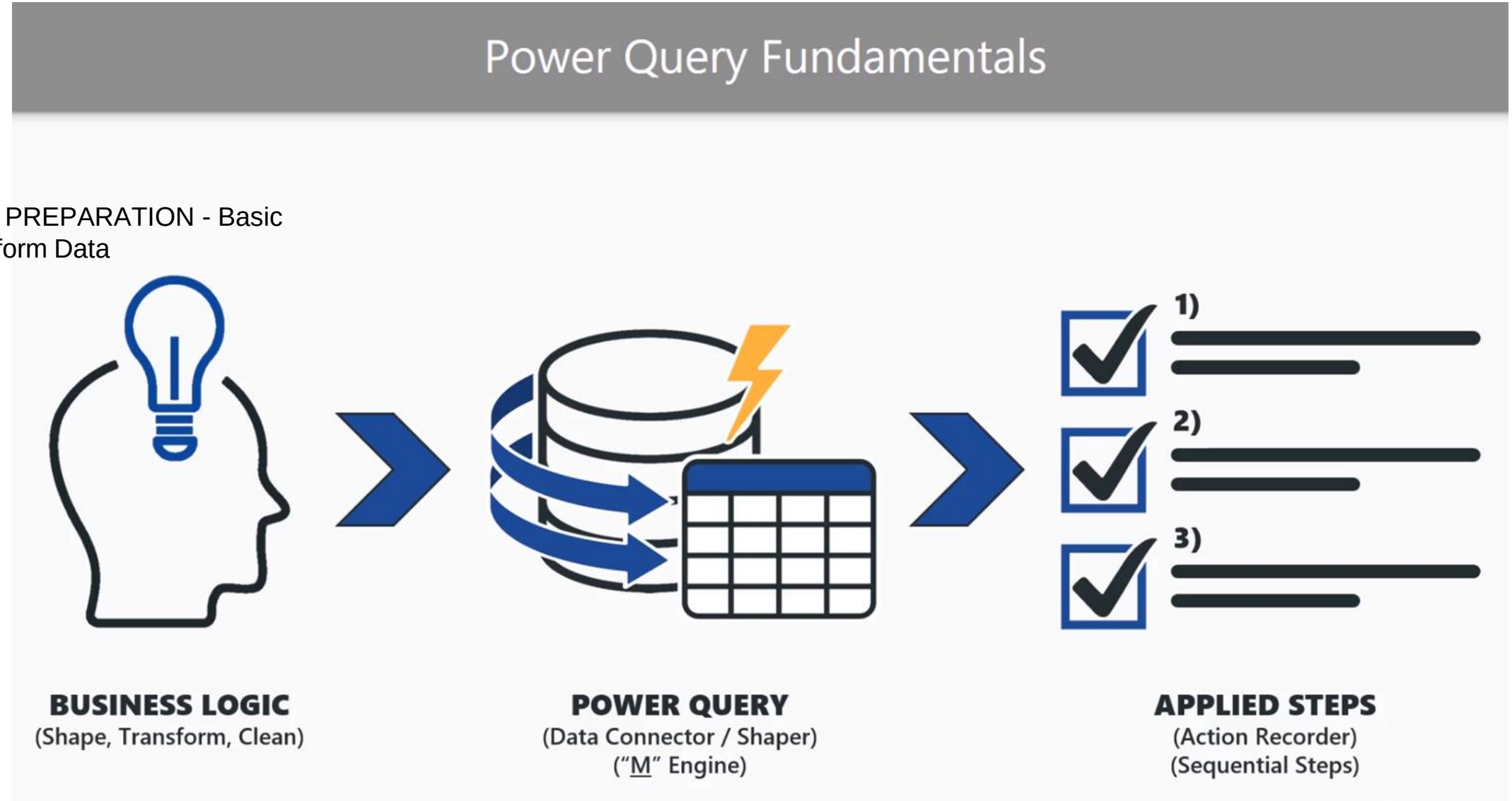


MASTERING DATA ANALYTICS

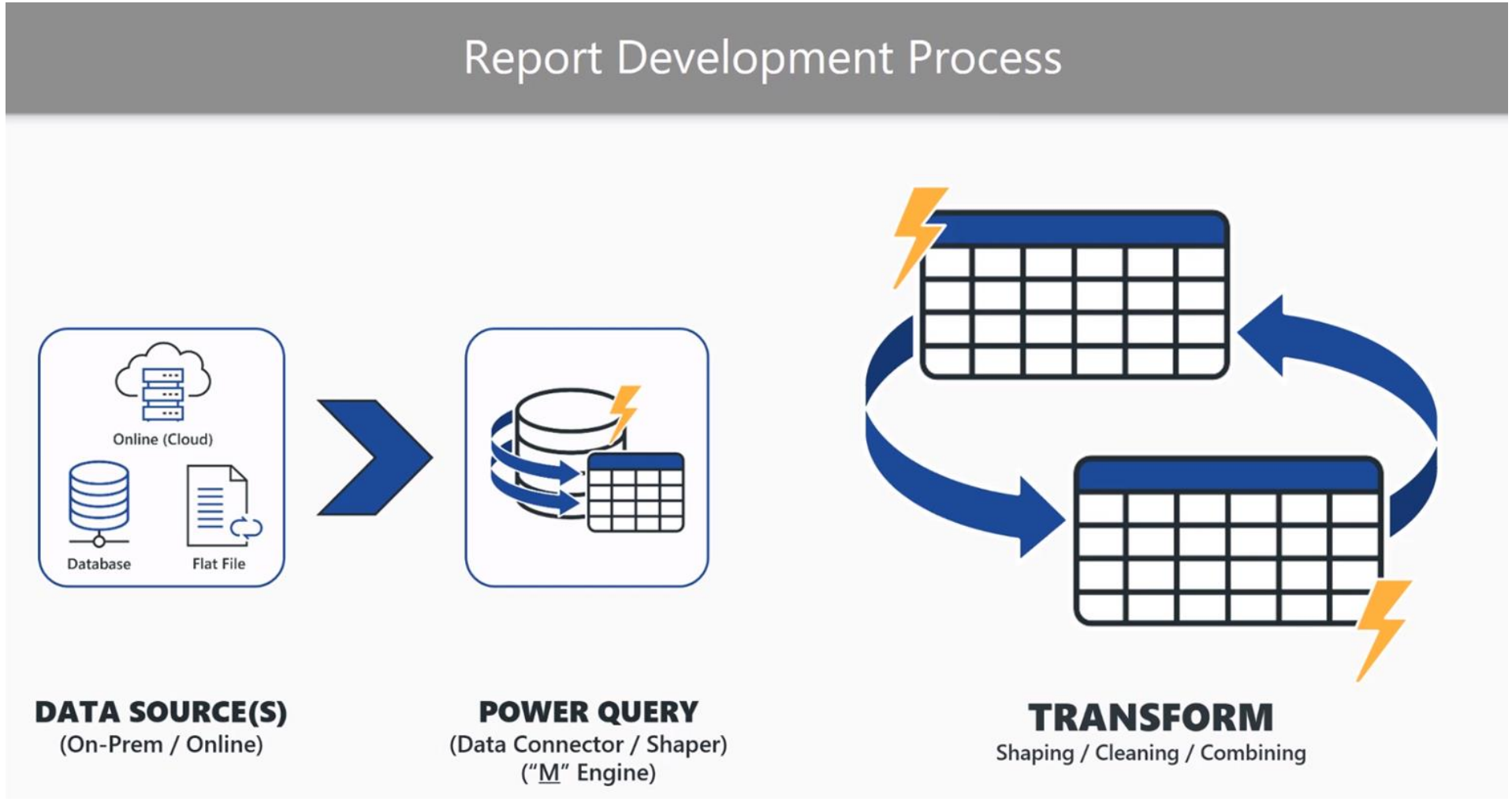


DATA PREPARATION - Basic Transform Data

DATA PREPARATION - Basic Transform Data



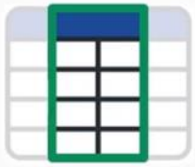
DATA PREPARATION - Basic Transform Data



DATA PREPARATION - Basic Transform Data

Essential Transformations For (Most) Tables

1) Choose Columns



2) Rename Columns

Prod_Name → Product Name

3) Change Data Type

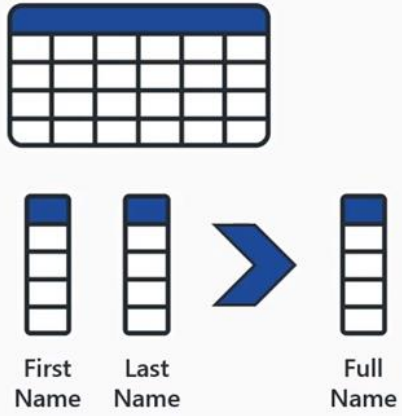
APPLIED STEPS
(Action Recorder)

DATA PREPARATION - Basic Transform Data
Add/ Modify Columns

MASTERING DATA ANALYTICS

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

- ☑ **1)** Add Columns
- ☑ **2)** Merge Columns
- ☑ **3)** Modify Columns

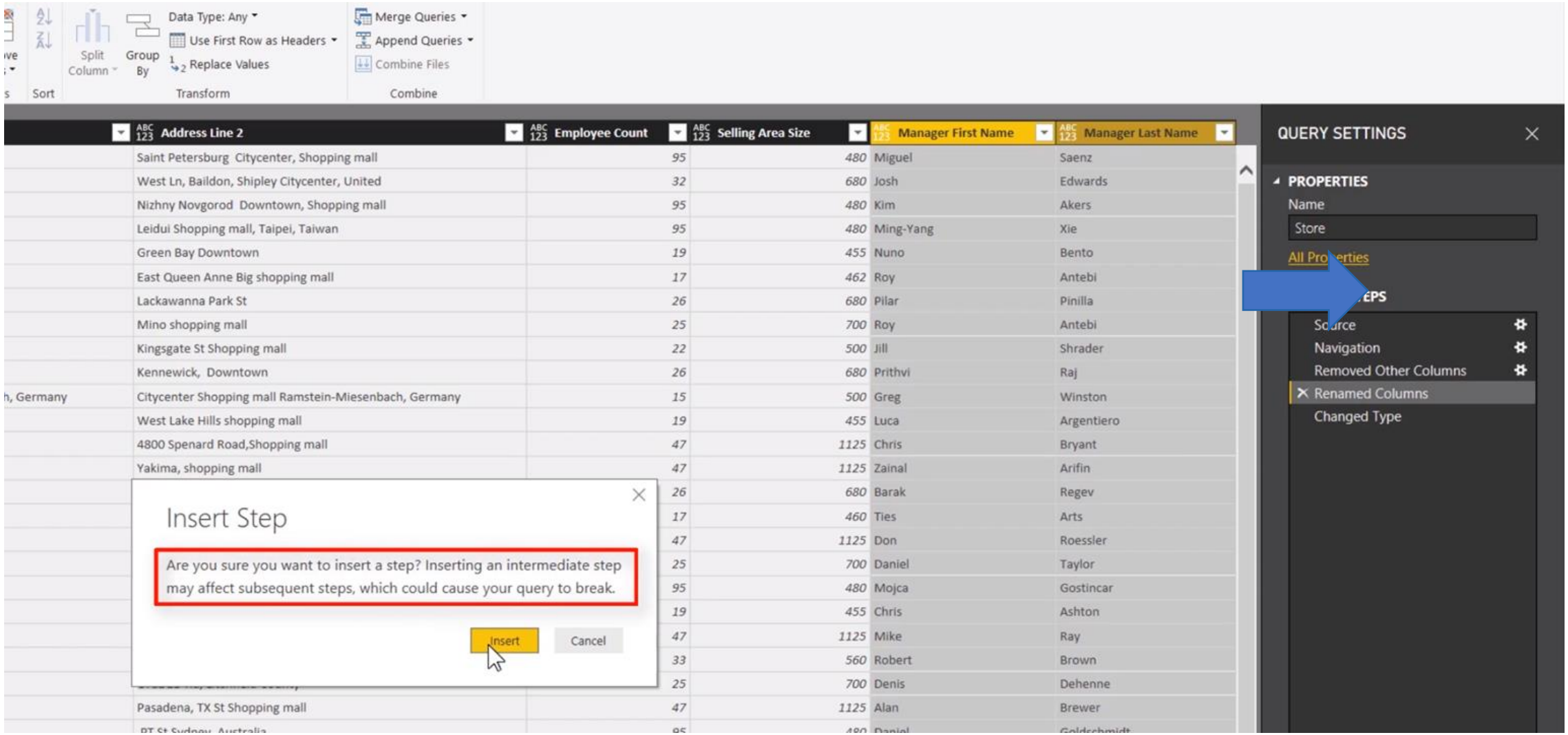


Apple | iPhone XXIII → iPhone XXIII

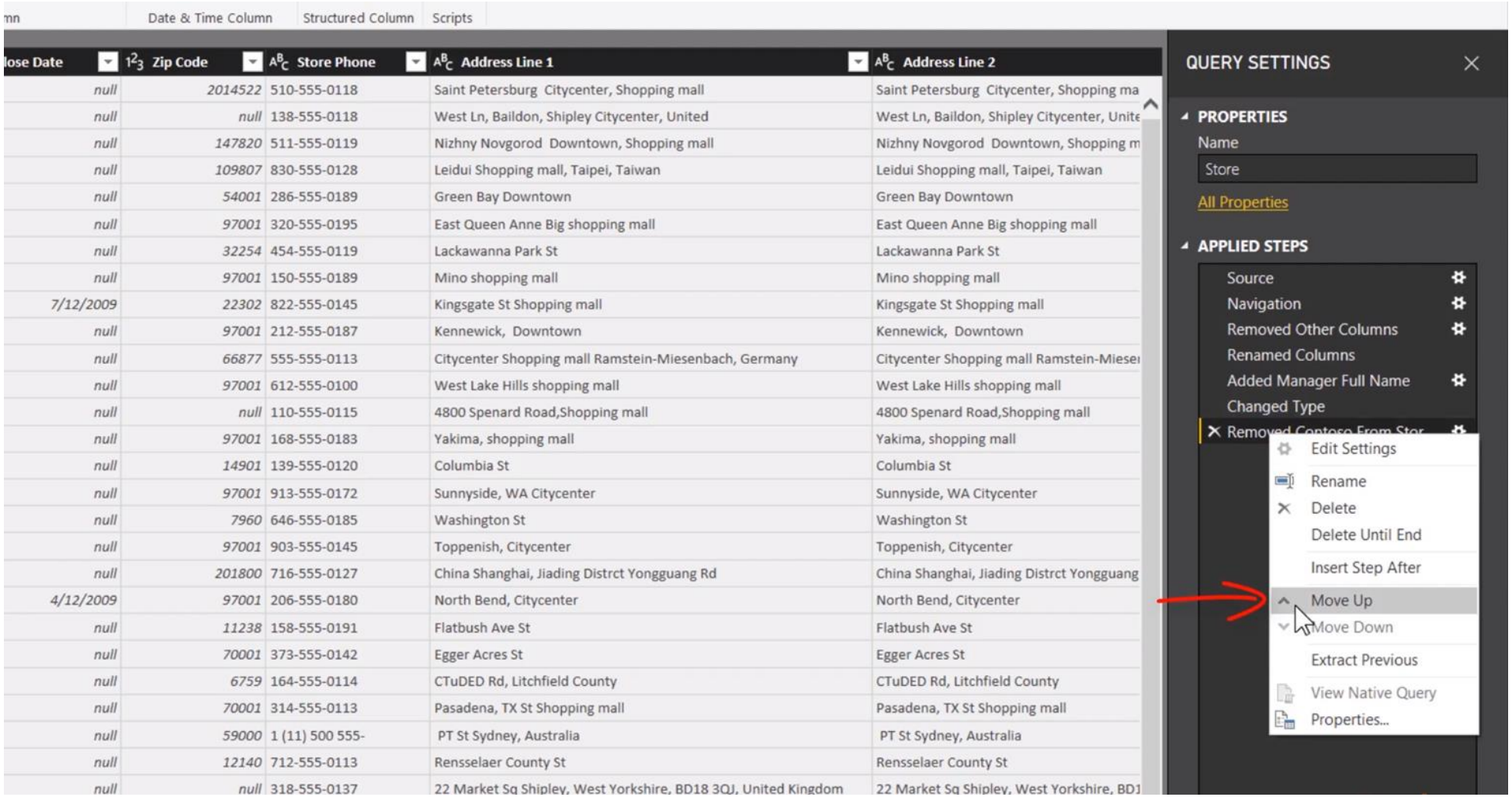
Sony | Playstation ∞ → Playstation ∞

APPLIED STEPS
(Action Recorder)

DATA PREPARATION - Basic Transform Data
Insert in Applied Steps



DATA PREPARATION - Basic Transform Data
Move in Applied Steps



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 ma
null	null	138-555-0118	West Ln, Baildon, Shipley Citycenter, United	West Ln, Baildon, Shipley Citycenter, Unite
null	147820	511-555-0119	Nizhny Novgorod Downtown, Shopping mall	Nizhny Novgorod Downtown, Shopping m
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-Miesen
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 Distrct Yongguang Rd	China Shanghai, Jiading Distrct Yongguang
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, BD1

MASTERING DATA ANALYTICS

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!!**

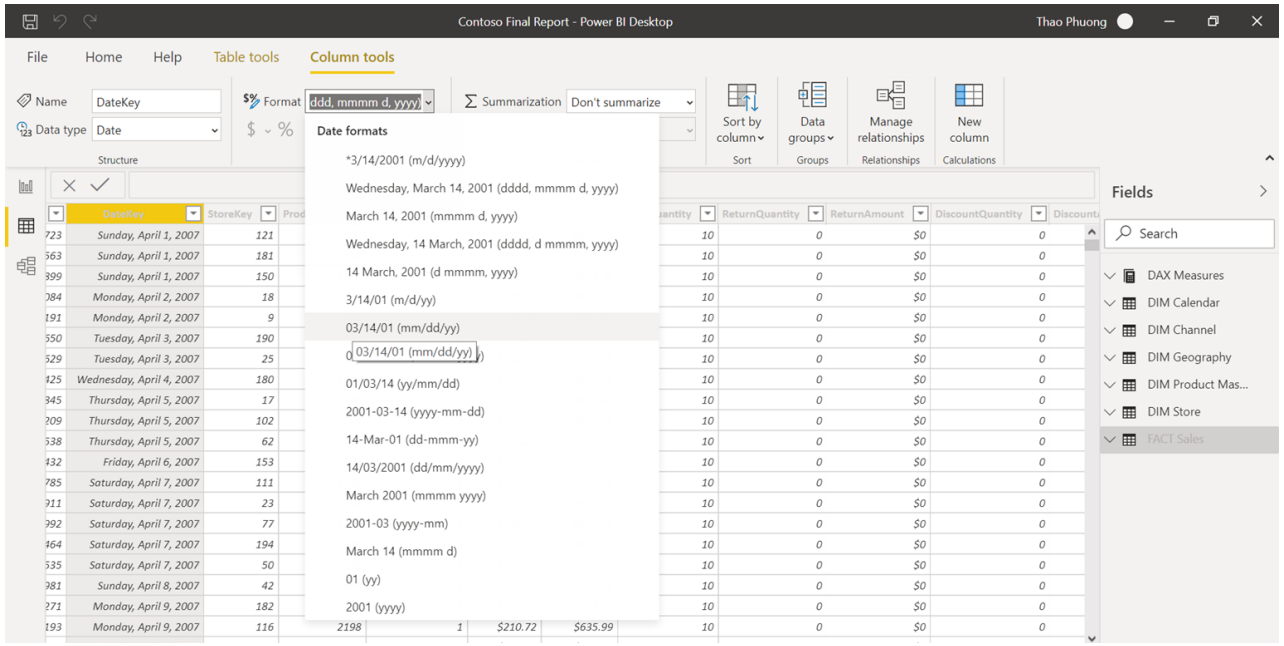
Set your ***Data Types*** *in the Query Editor*

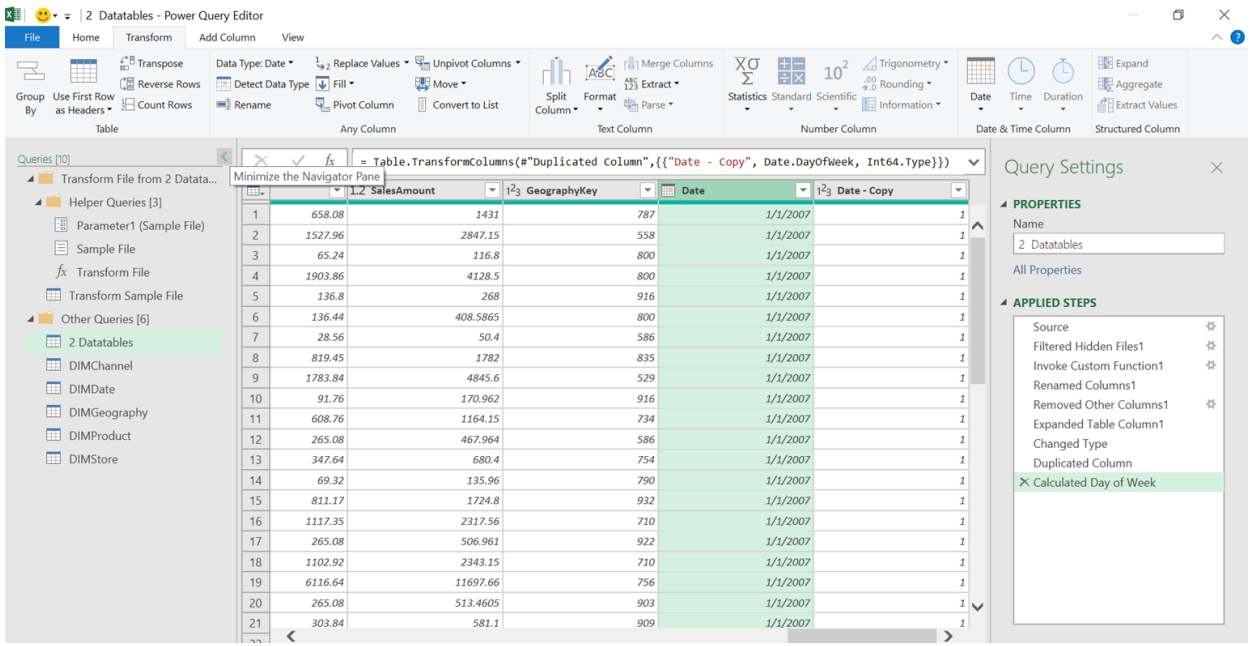
Set your ***Data Formats*** *($ %, etc) in the Data Model*

Pro Tip: Data type is different from data format

DATA PREPARATION - Basic Transform Data

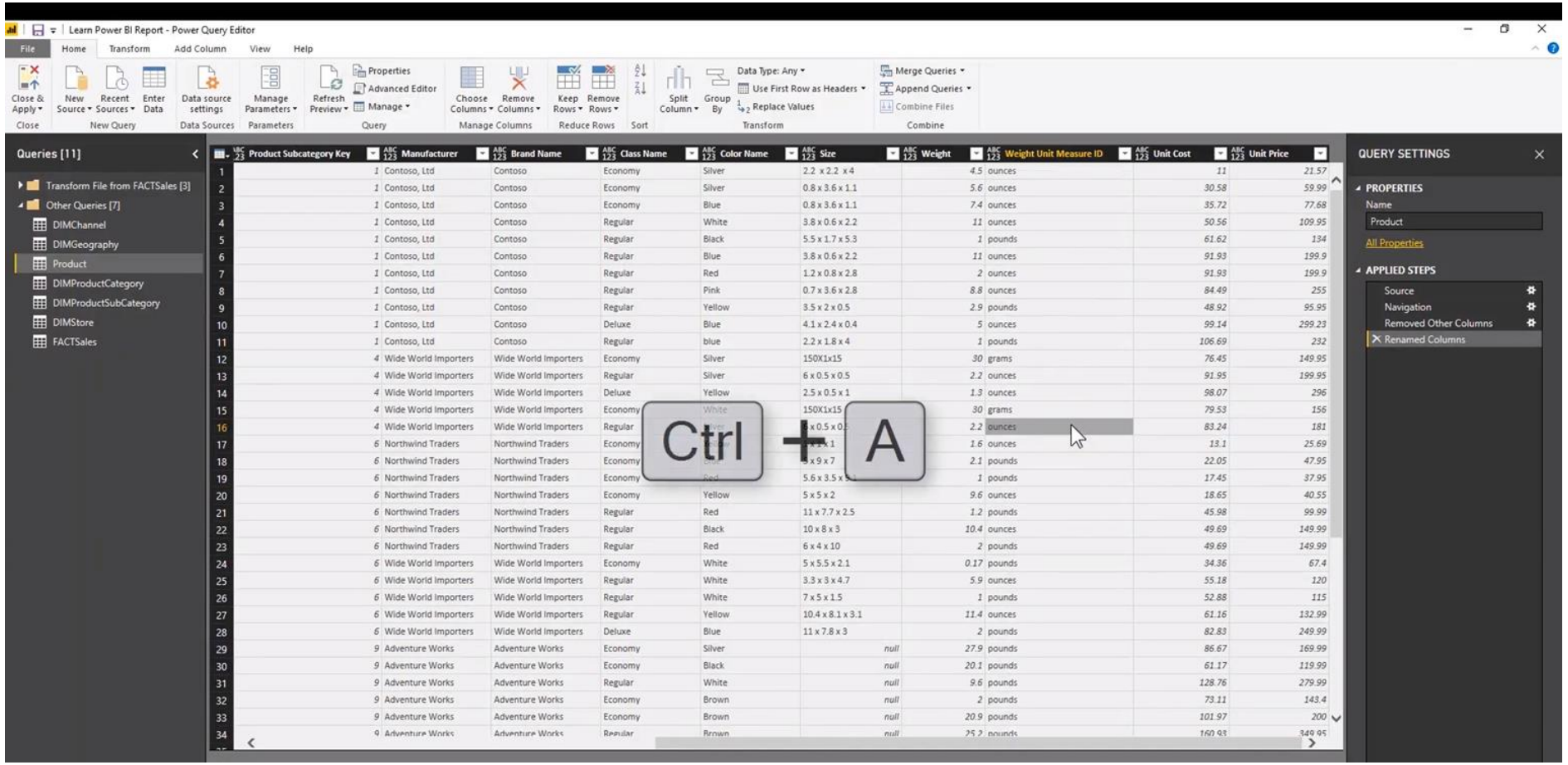
Data Type vs Data Format



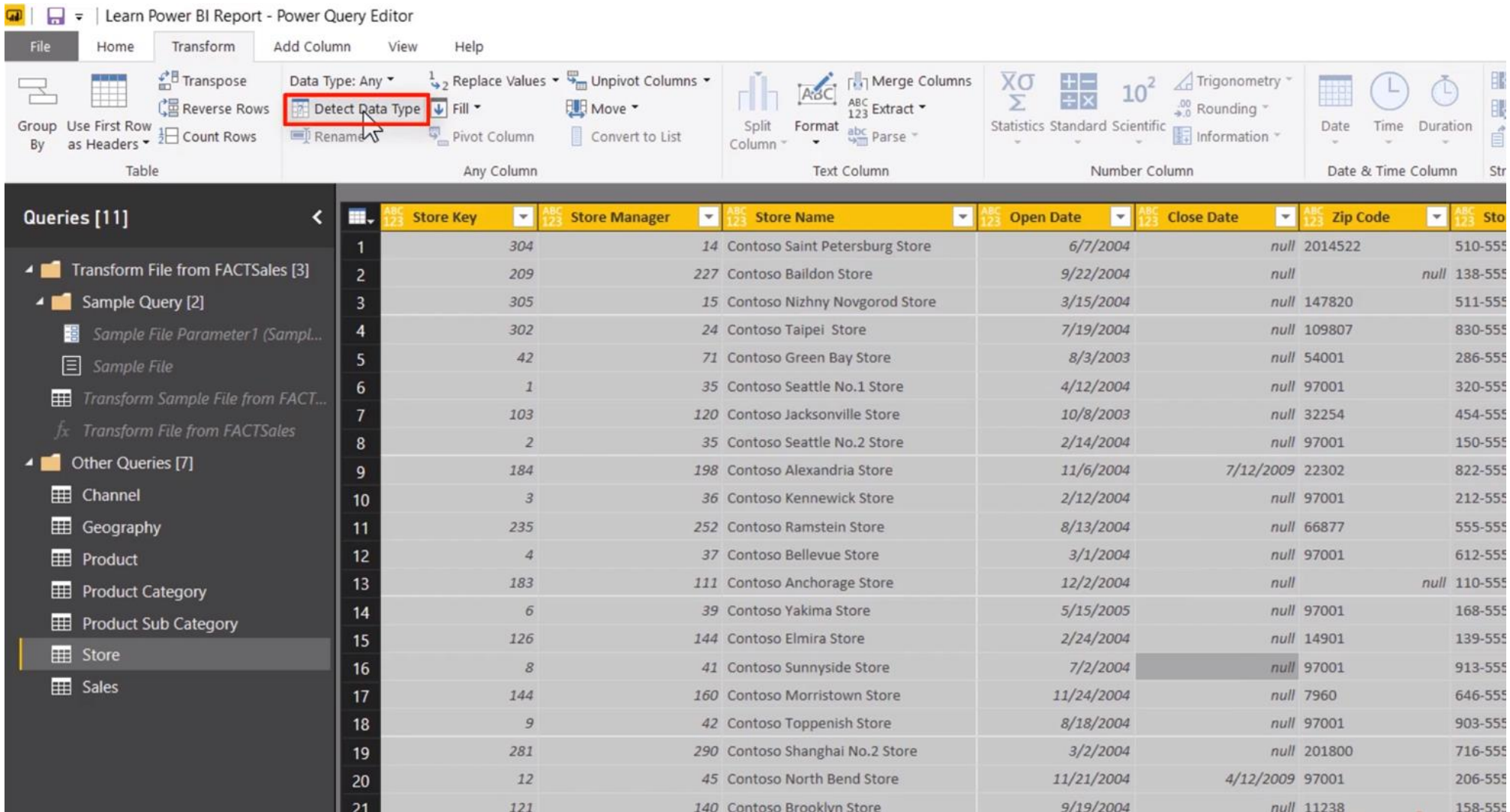


DATA PREPARATION - Basic Transform Data

Detect Data Type

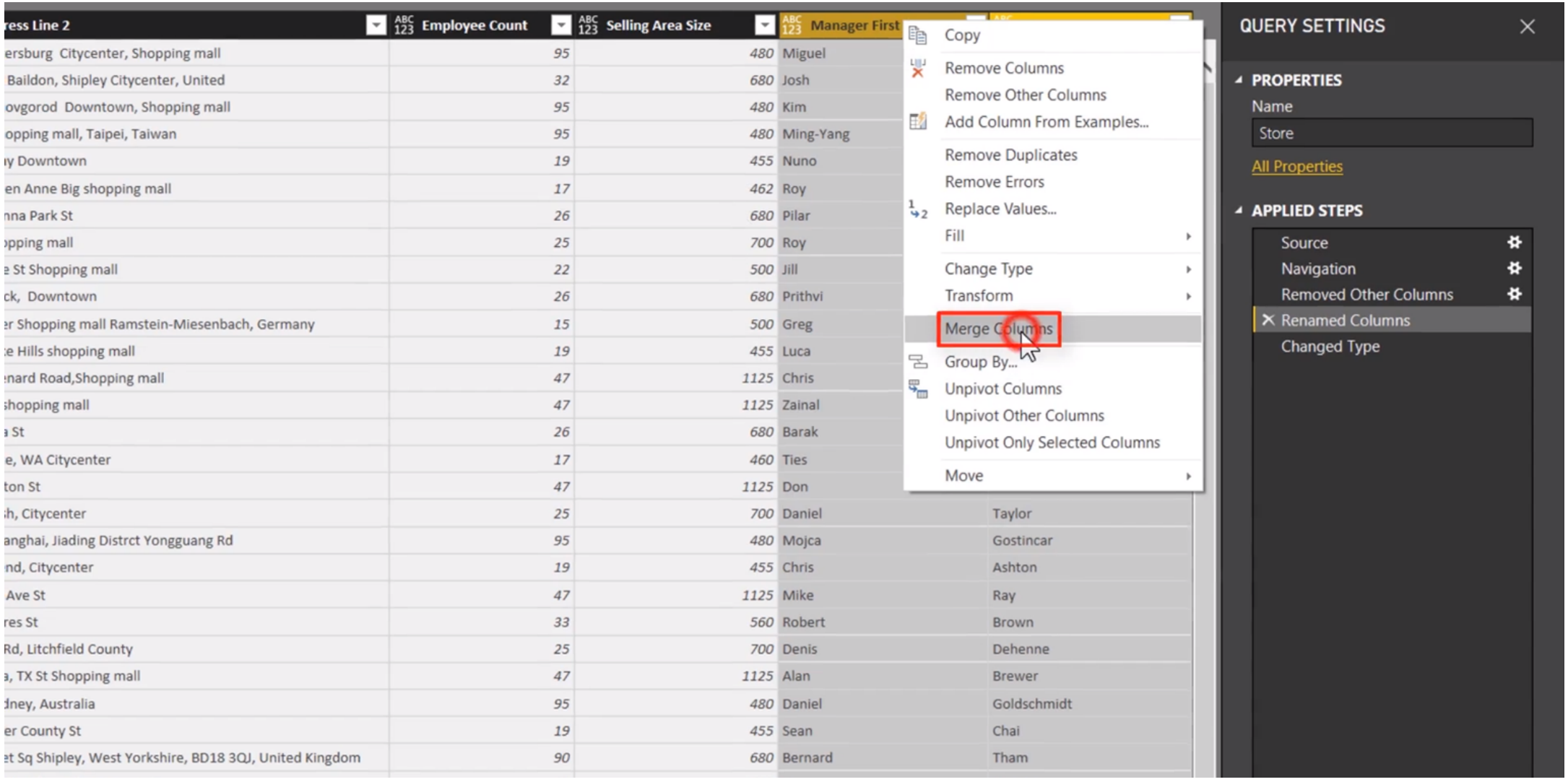


DATA PREPARATION - Basic Transform Data
Detect Data Type



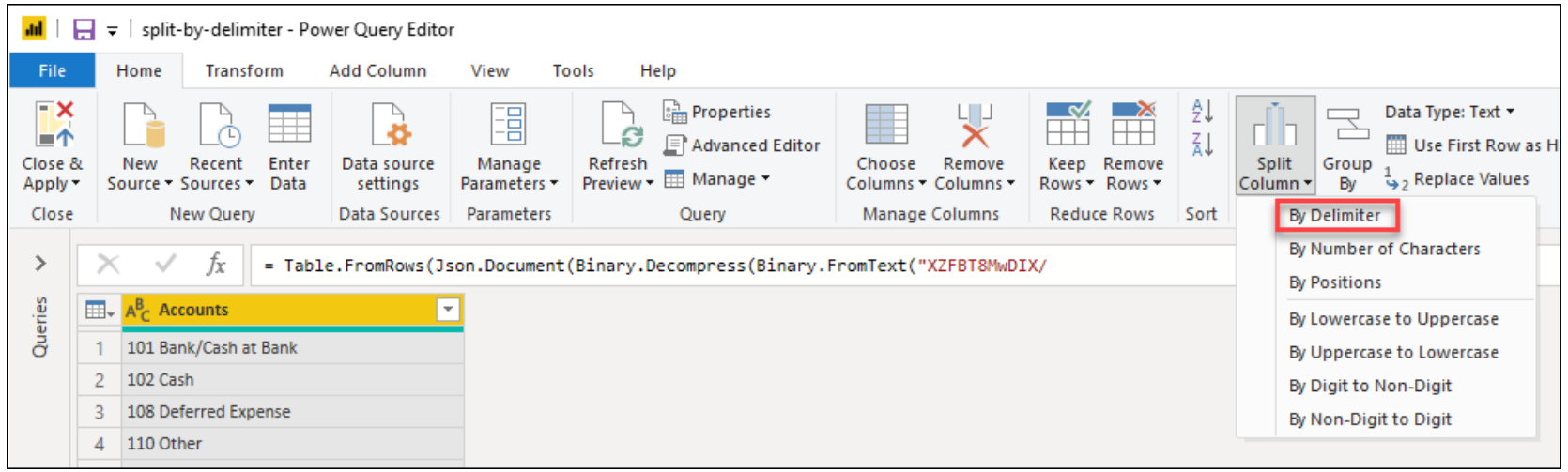
	Store Key	Store Manager	Store Name	Open Date	Close Date	Zip Code	Sto
1	304	14	Contoso Saint Petersburg Store	6/7/2004	null	2014522	510-555
2	209	227	Contoso Baildon Store	9/22/2004	null	null	138-555
3	305	15	Contoso Nizhny Novgorod Store	3/15/2004	null	147820	511-555
4	302	24	Contoso Taipei Store	7/19/2004	null	109807	830-555
5	42	71	Contoso Green Bay Store	8/3/2003	null	54001	286-555
6	1	35	Contoso Seattle No.1 Store	4/12/2004	null	97001	320-555
7	103	120	Contoso Jacksonville Store	10/8/2003	null	32254	454-555
8	2	35	Contoso Seattle No.2 Store	2/14/2004	null	97001	150-555
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	212-555
11	235	252	Contoso Ramstein Store	8/13/2004	null	66877	555-555
12	4	37	Contoso Bellevue Store	3/1/2004	null	97001	612-555
13	183	111	Contoso Anchorage Store	12/2/2004	null	null	110-555
14	6	39	Contoso Yakima Store	5/15/2005	null	97001	168-555
15	126	144	Contoso Elmira Store	2/24/2004	null	14901	139-555
16	8	41	Contoso Sunnyside Store	7/2/2004	null	97001	913-555
17	144	160	Contoso Morristown Store	11/24/2004	null	7960	646-555
18	9	42	Contoso Toppenish Store	8/18/2004	null	97001	903-555
19	281	290	Contoso Shanghai No.2 Store	3/2/2004	null	201800	716-555
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	158-555

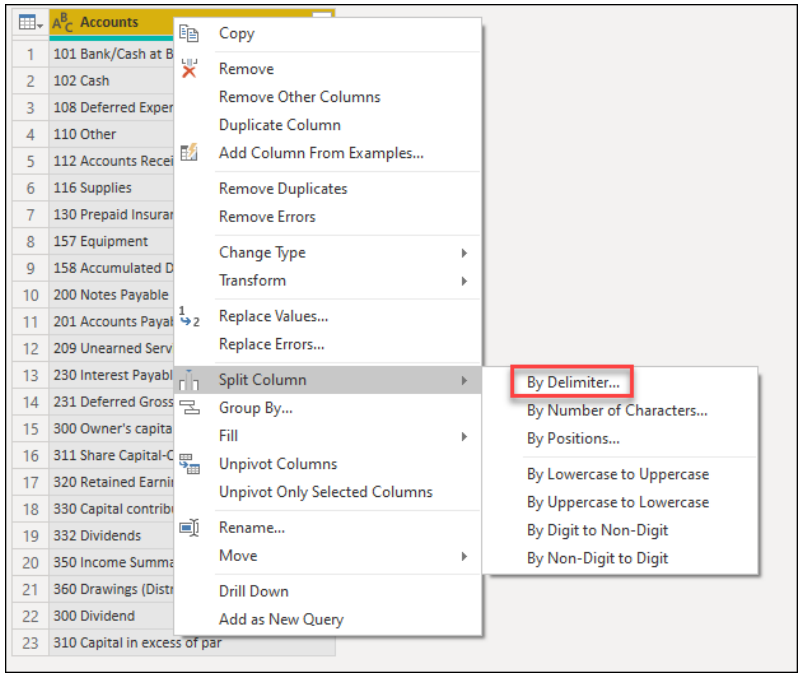
DATA PREPARATION - Basic Transform Data
Merge Column



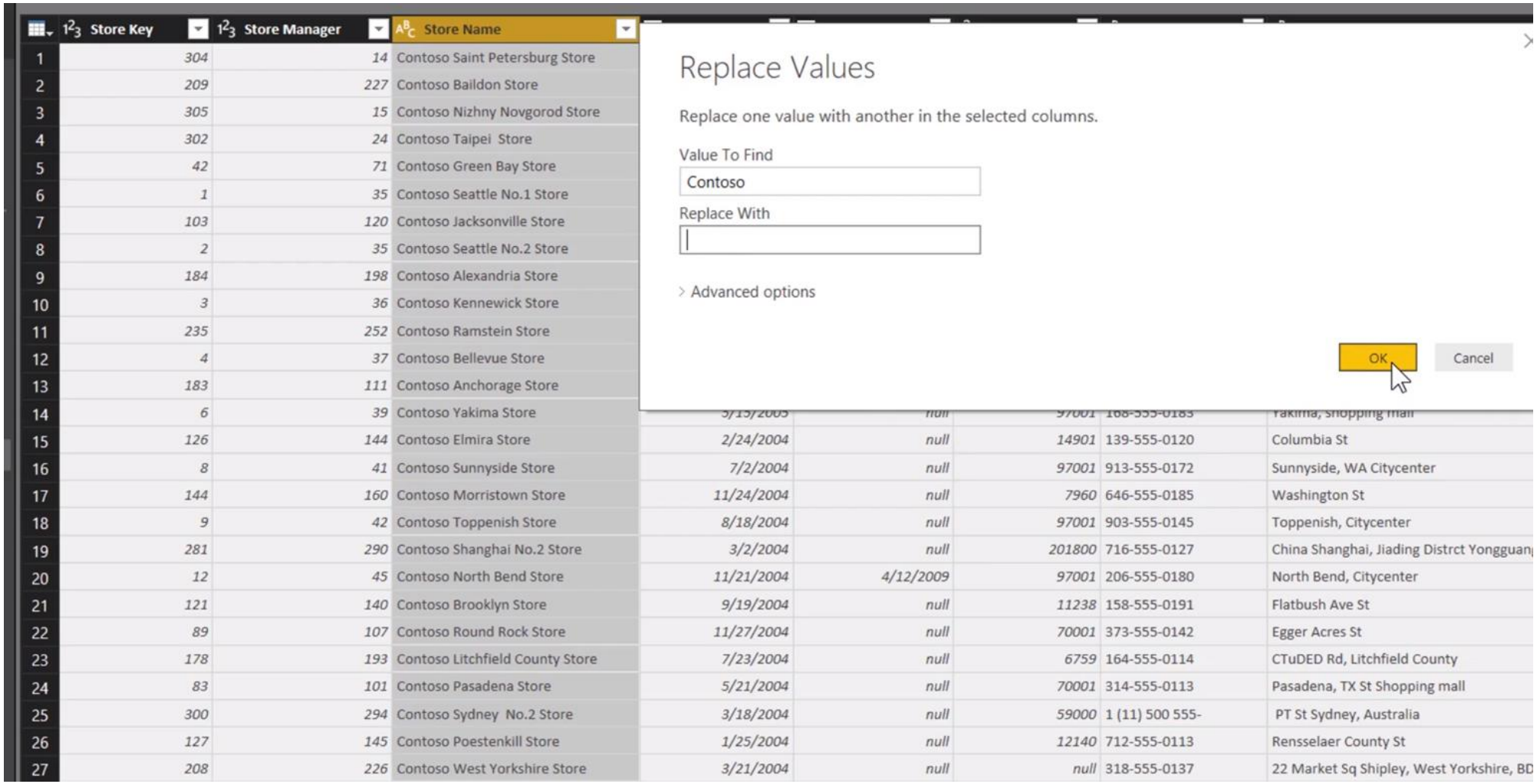
DATA PREPARATION - Basic Transform Data
Split Column

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





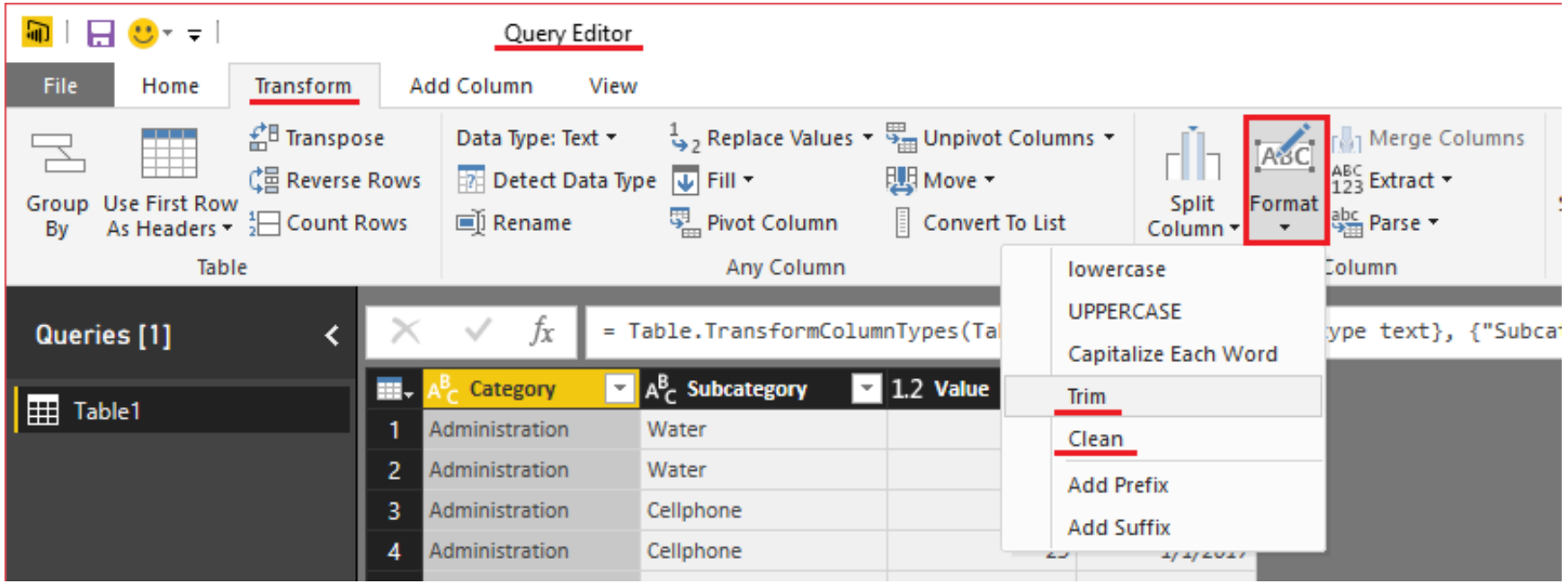
DATA PREPARATION - Basic Transform Data
Replace



	Store Key	Store Manager	Store Name					
1	304	14	Contoso Saint Petersburg Store					
2	209	227	Contoso Baildon Store					
3	305	15	Contoso Nizhny Novgorod Store					
4	302	24	Contoso Taipei Store					
5	42	71	Contoso Green Bay Store					
6	1	35	Contoso Seattle No.1 Store					
7	103	120	Contoso Jacksonville Store					
8	2	35	Contoso Seattle No.2 Store					
9	184	198	Contoso Alexandria Store					
10	3	36	Contoso Kennewick Store					
11	235	252	Contoso Ramstein Store					
12	4	37	Contoso Bellevue Store					
13	183	111	Contoso Anchorage Store					
14	6	39	Contoso Yakima Store	5/15/2005	null	97001	168-555-0183	Yakima, shopping mall
15	126	144	Contoso Elmira Store	2/24/2004	null	14901	139-555-0120	Columbia St
16	8	41	Contoso Sunnyside Store	7/2/2004	null	97001	913-555-0172	Sunnyside, WA Citycenter
17	144	160	Contoso Morristown Store	11/24/2004	null	7960	646-555-0185	Washington St
18	9	42	Contoso Toppenish Store	8/18/2004	null	97001	903-555-0145	Toppenish, Citycenter
19	281	290	Contoso Shanghai No.2 Store	3/2/2004	null	201800	716-555-0127	China Shanghai, Jiading Distrct Yongguan
20	12	45	Contoso North Bend Store	11/21/2004	4/12/2009	97001	206-555-0180	North Bend, Citycenter
21	121	140	Contoso Brooklyn Store	9/19/2004	null	11238	158-555-0191	Flatbush Ave St
22	89	107	Contoso Round Rock Store	11/27/2004	null	70001	373-555-0142	Egger Acres St
23	178	193	Contoso Litchfield County Store	7/23/2004	null	6759	164-555-0114	CTuDED Rd, Litchfield County
24	83	101	Contoso Pasadena Store	5/21/2004	null	70001	314-555-0113	Pasadena, TX St Shopping mall
25	300	294	Contoso Sydney No.2 Store	3/18/2004	null	59000	1 (11) 500 555-	PT St Sydney, Australia
26	127	145	Contoso Poestenkill Store	1/25/2004	null	12140	712-555-0113	Rensselaer County St
27	208	226	Contoso West Yorkshire Store	3/21/2004	null	null	318-555-0137	22 Market Sq Shipley, West Yorkshire, BD

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!



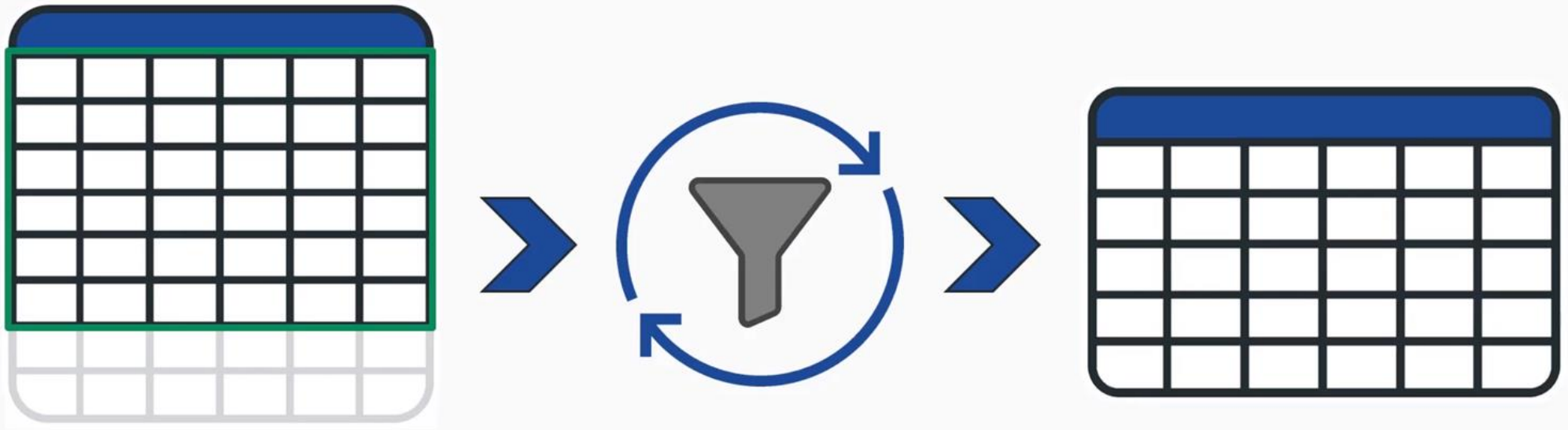
Untrimmed Data	Trimmed Data
Marry Angela	Marry Angela
Joey Tribiani	Joey Tribiani
Mathew Pery	Mathew Pery
Jenier Aniston	Jenier Aniston

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

MASTERING DATA ANALYTICS

Why Should We Filter Data?

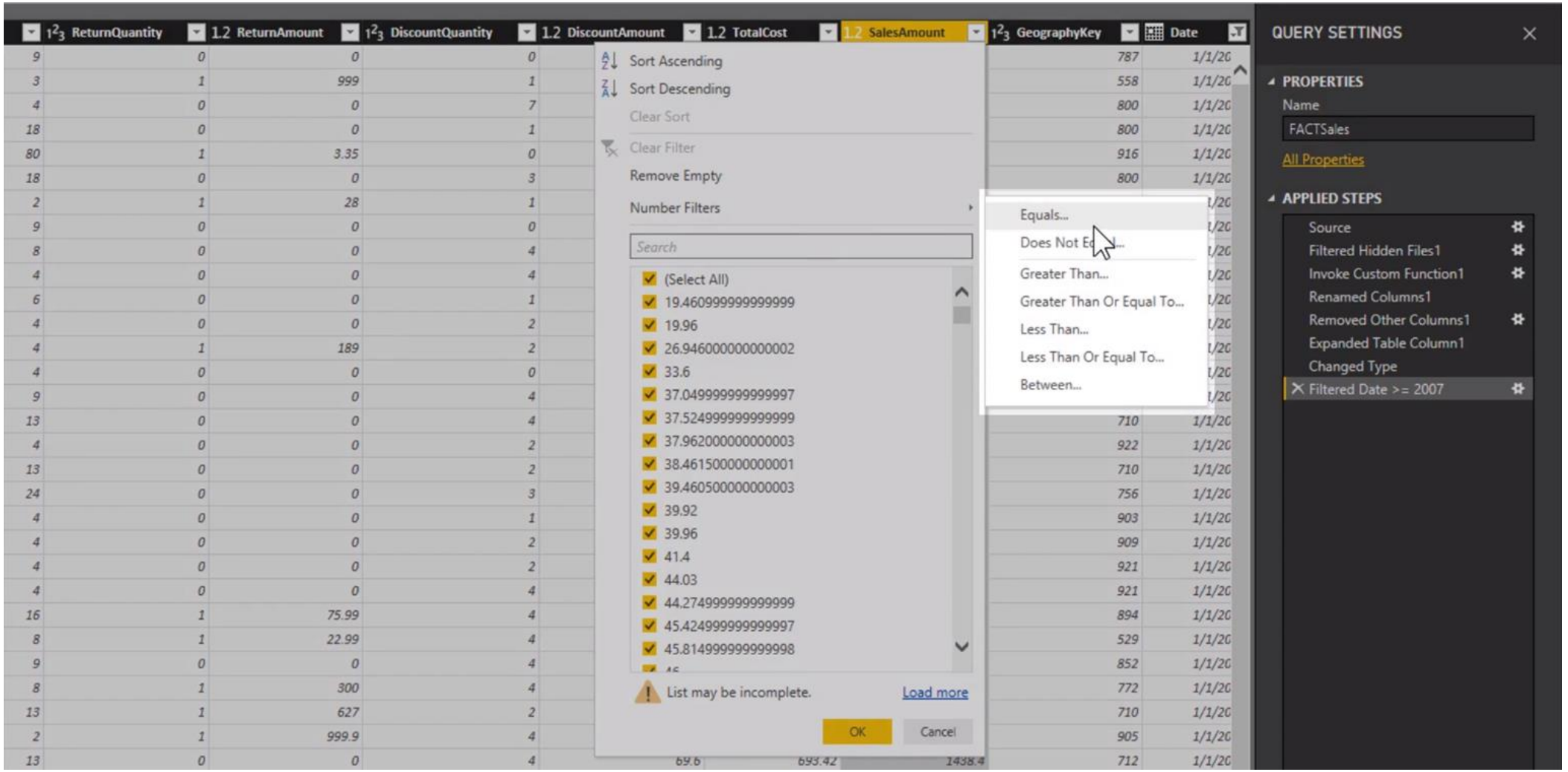


ORIGINAL TABLE
(Rows To Keep)

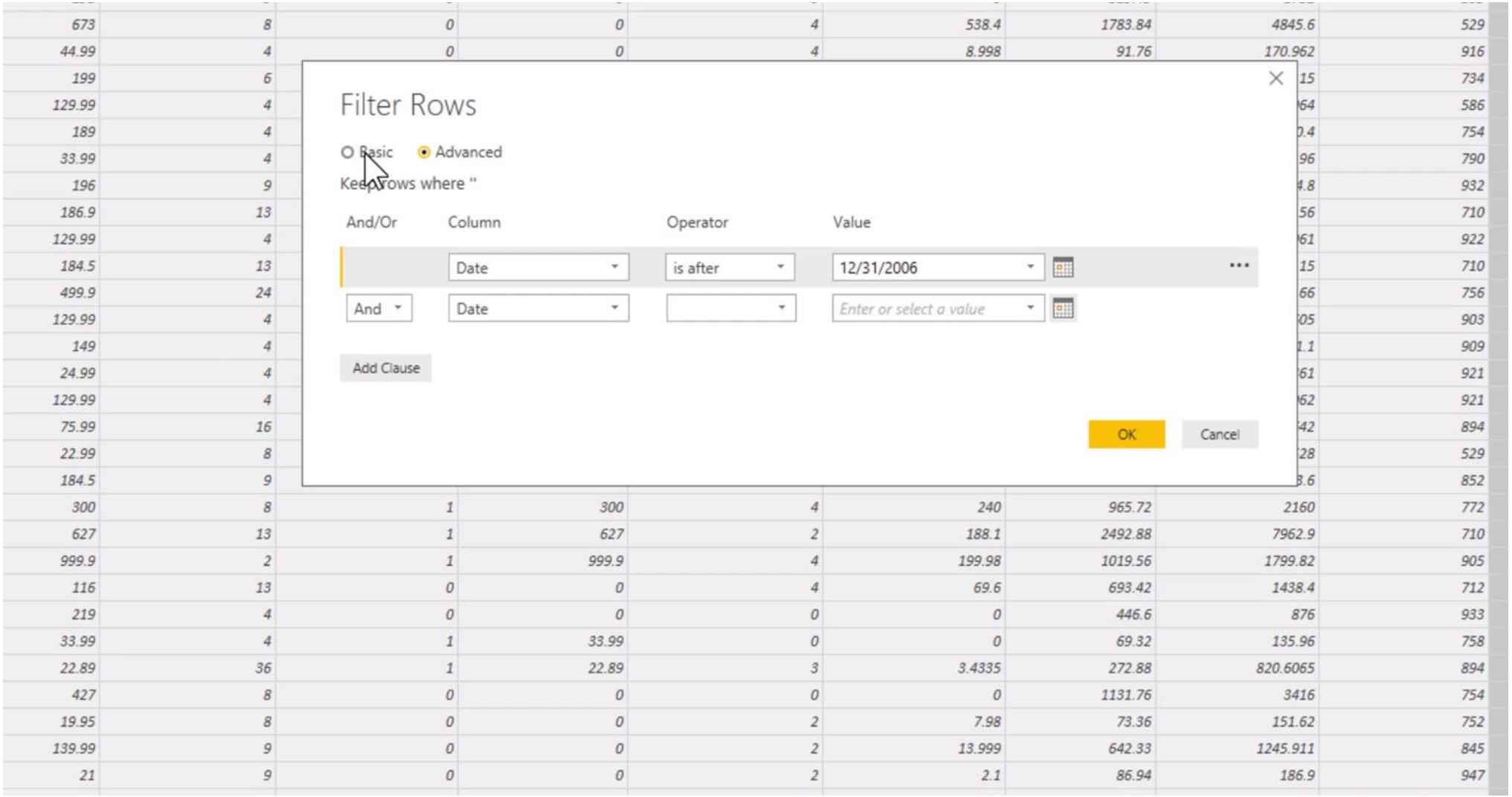
APPLY FILTERS
(Filtering = Reducing Rows)

TRANSFORMED TABLE
(Only necessary rows)

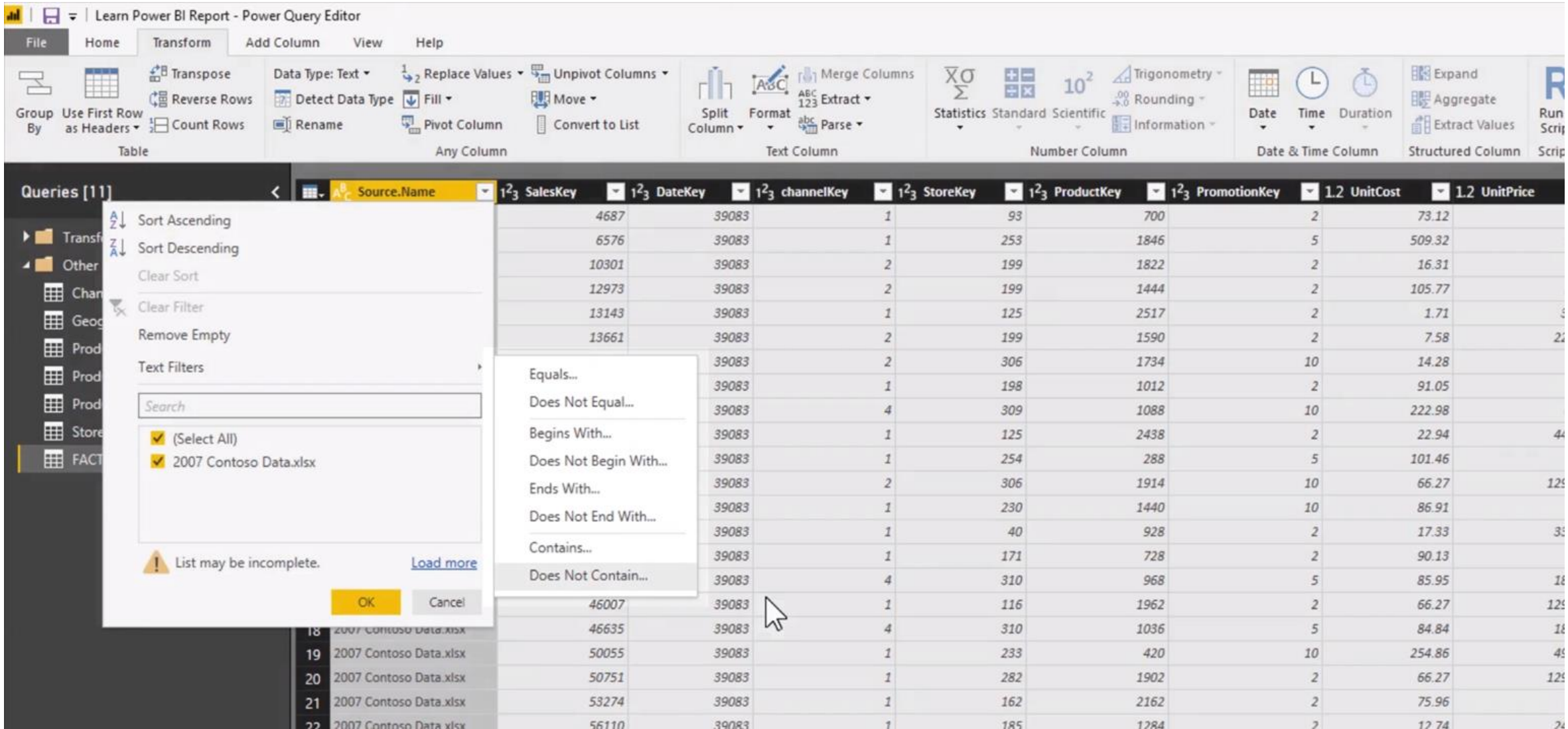
DATA PREPARATION - Basic Transform Data
Filter



DATA PREPARATION - Basic Transform Data
Advanced Filter

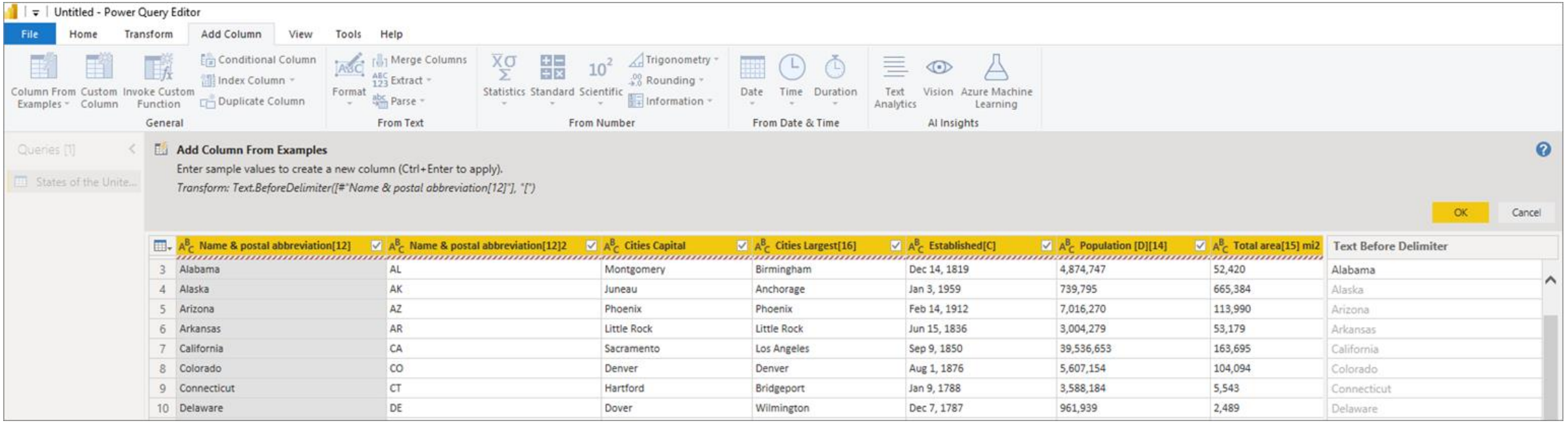


DATA PREPARATION - Basic Transform Data
Advanced Filter



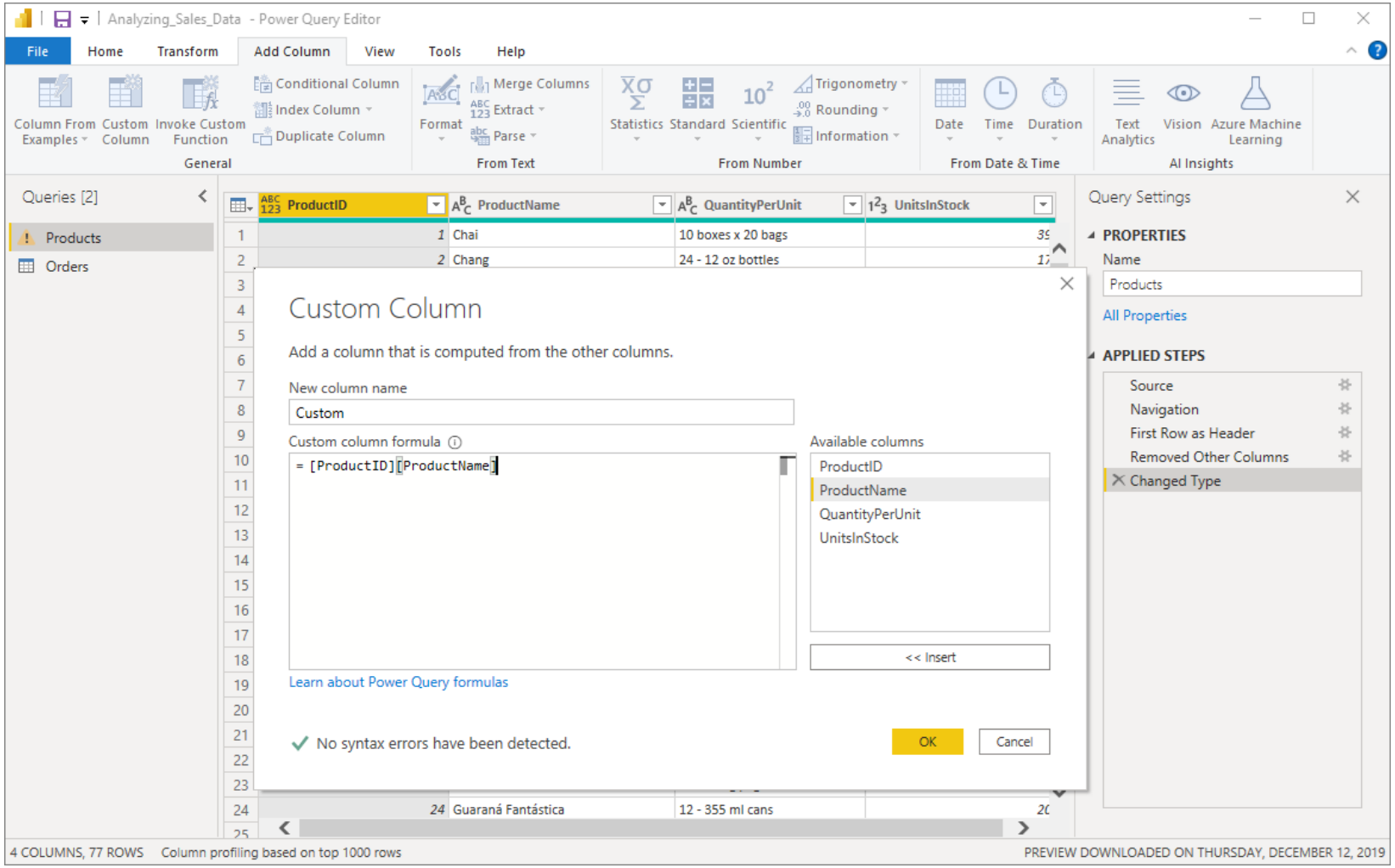
DATA PREPARATION - Basic Transform Data

Add Column from Example



	Name & postal abbreviation[12]	Name & postal abbreviation[12]2	Cities Capital	Cities Largest[16]	Established[C]	Population [D][14]	Total area[15] mi2	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

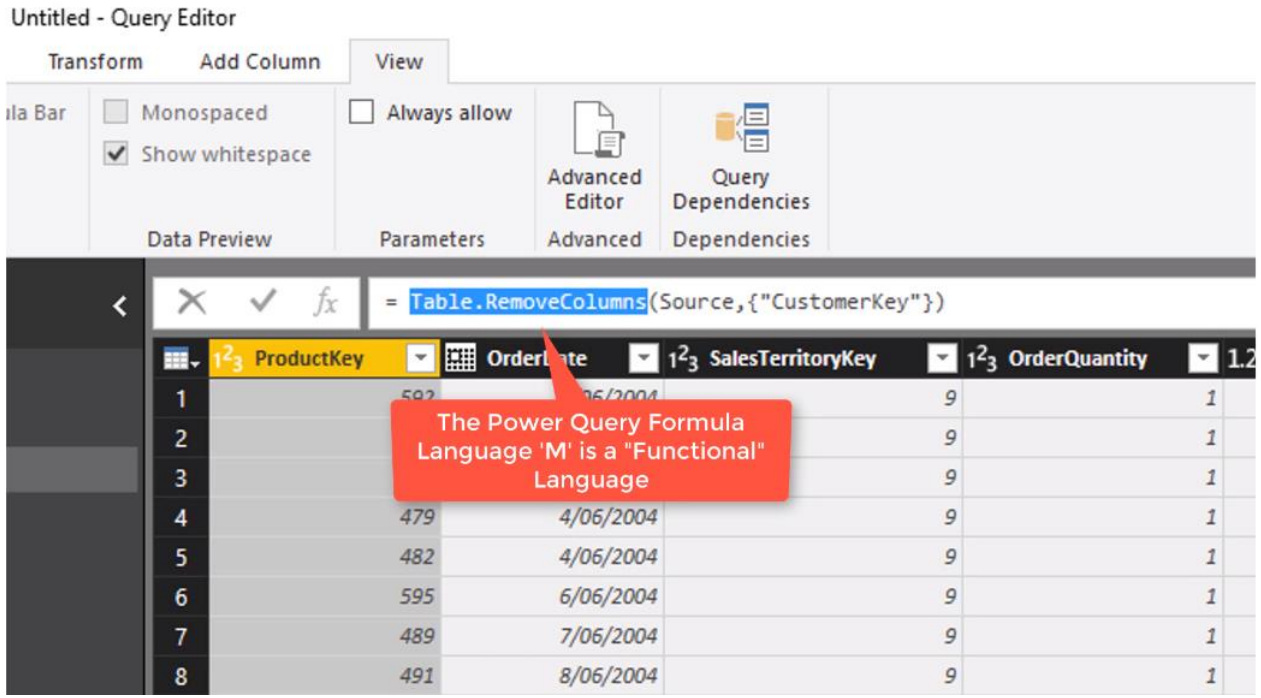


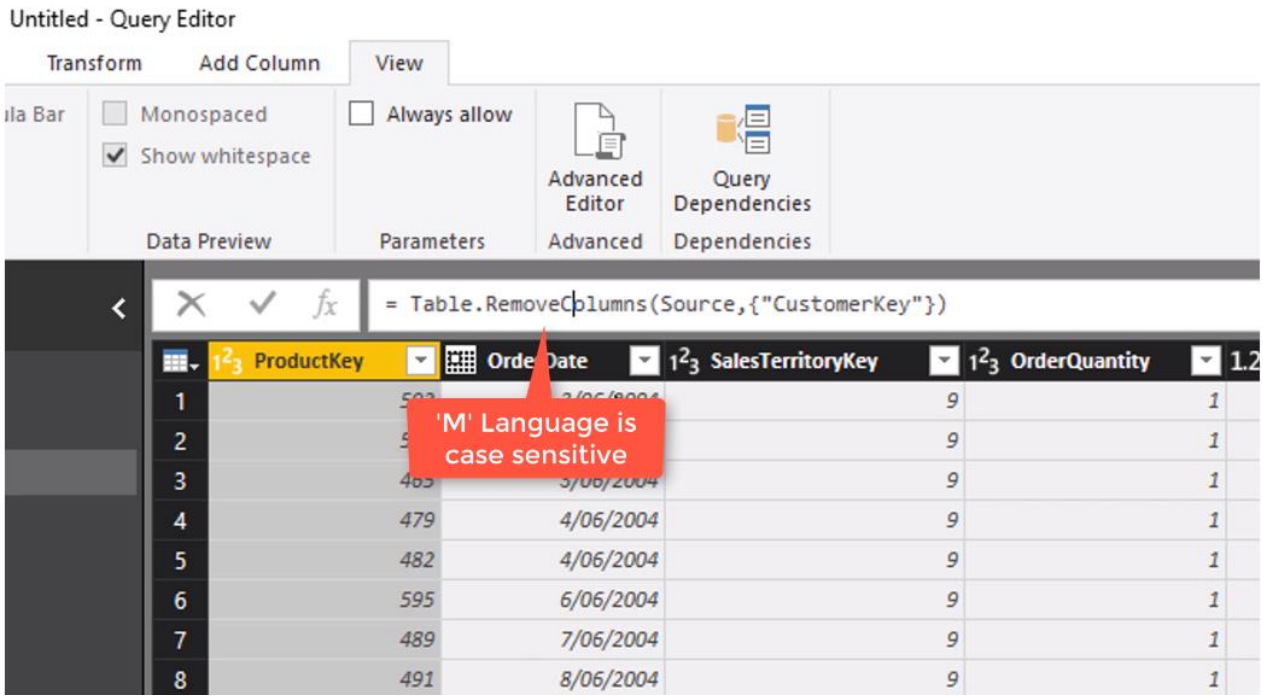
MASTERING DATA ANALYTICS

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

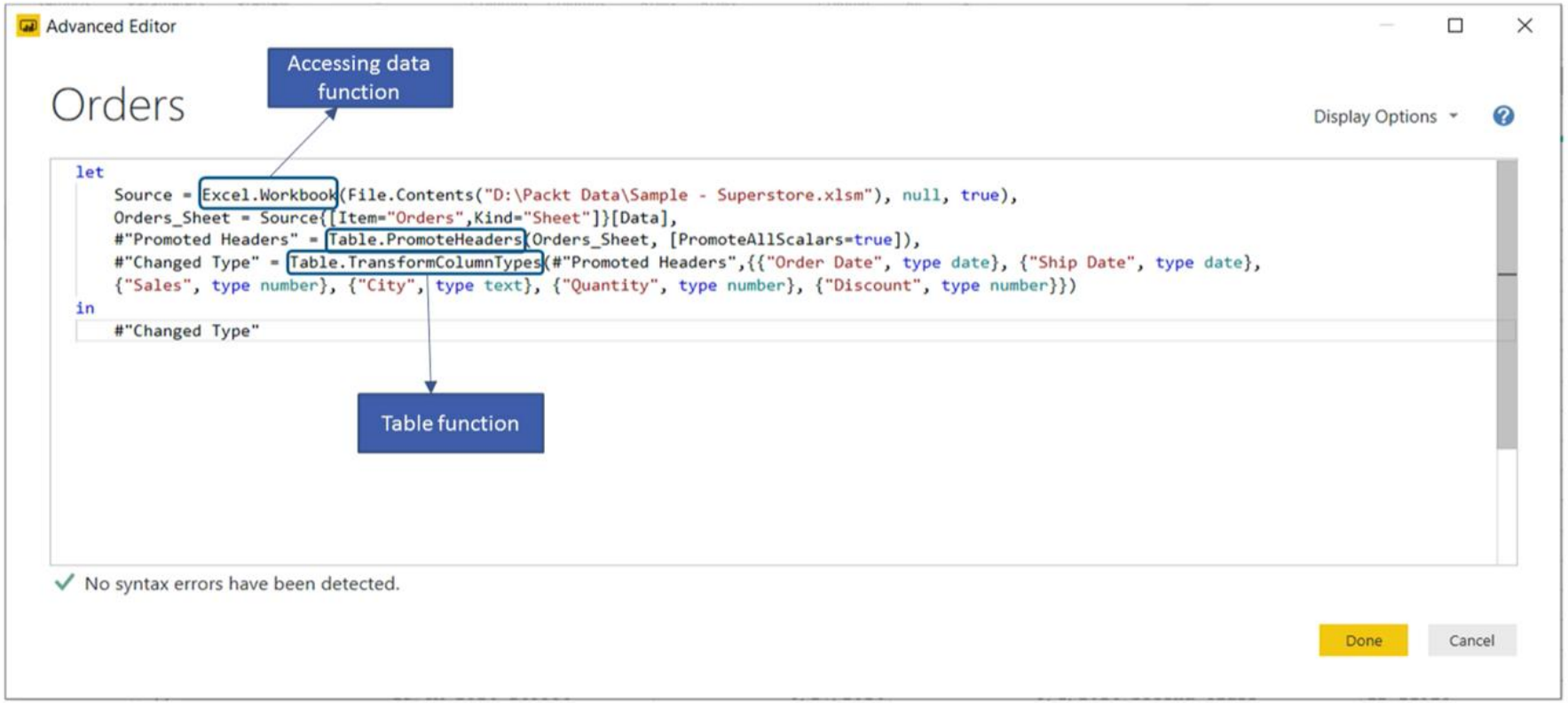
DATA PREPARATION - Basic Transform Data
M-code



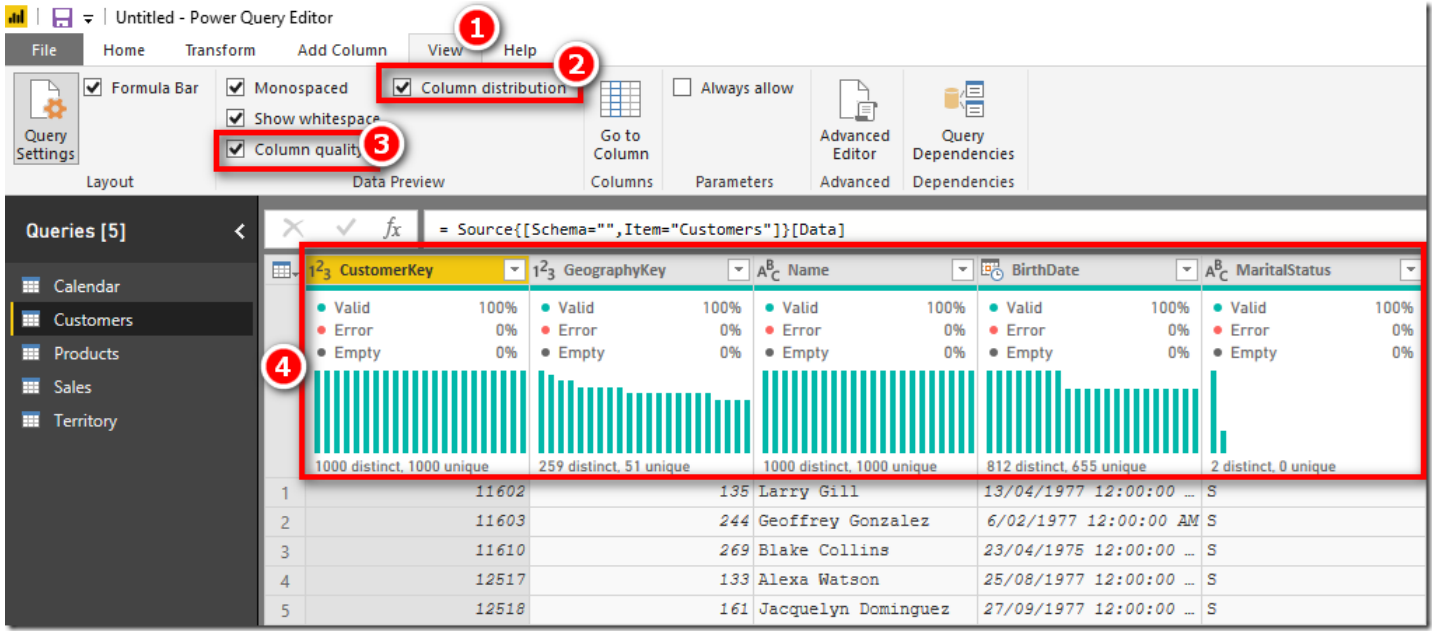


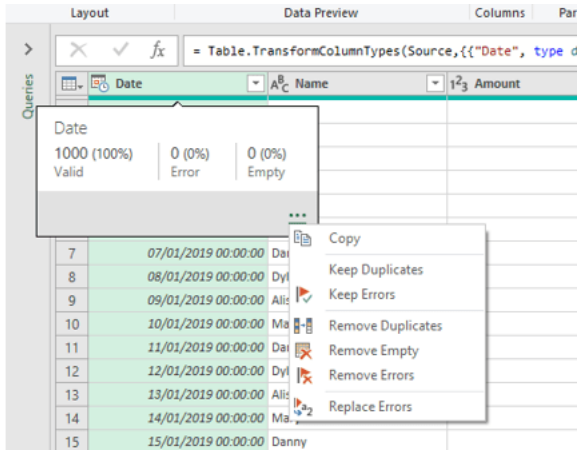
MASTERING DATA ANALYTICS

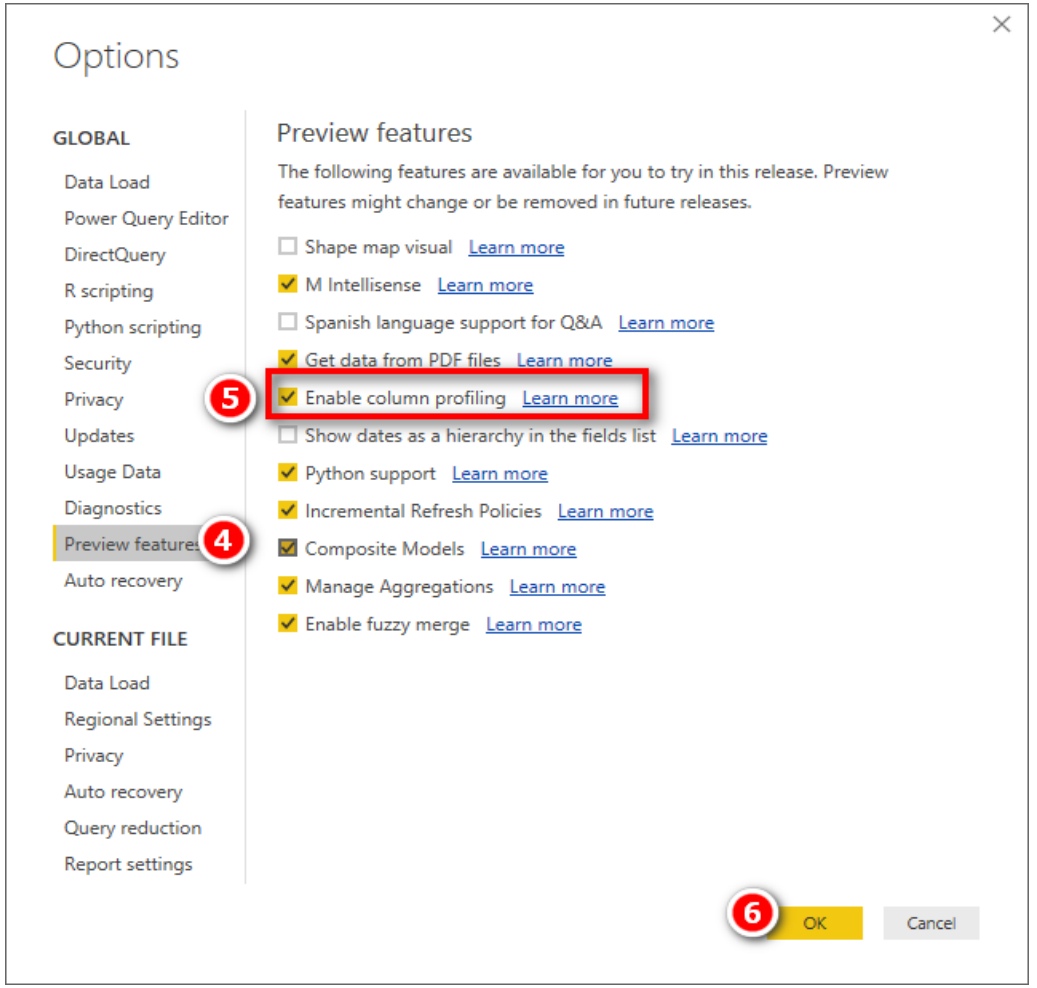
Power Query M Functions (Continued)



DATA PREPARATION - Basic Transform Data
Data profiling

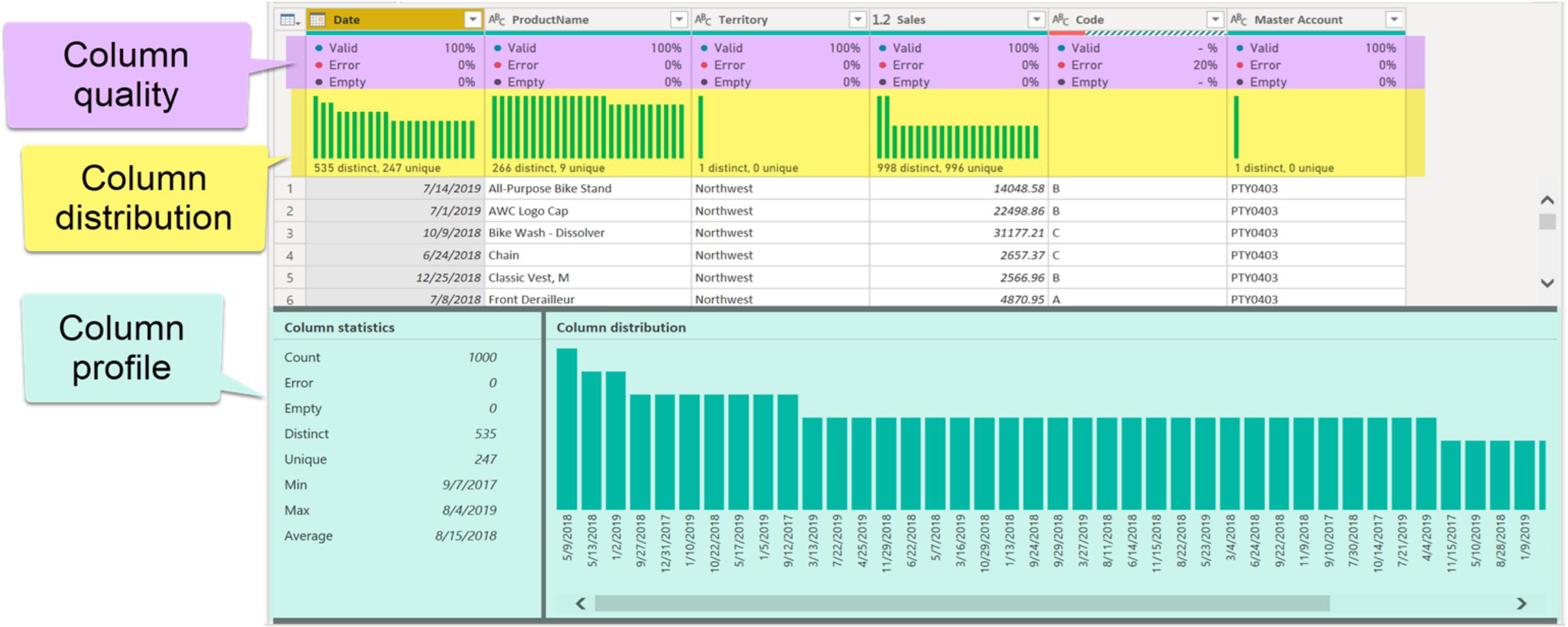




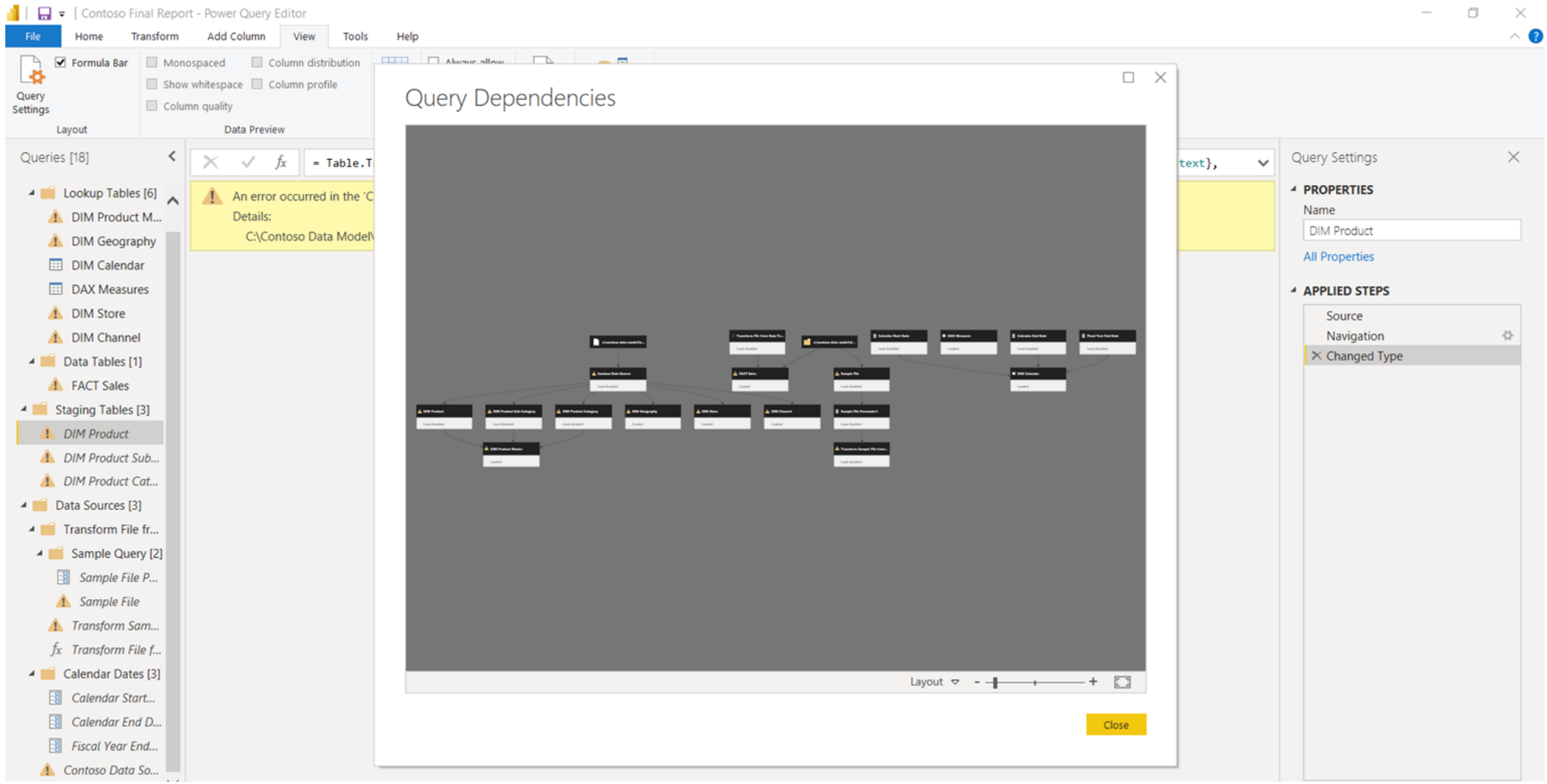


DATA PREPARATION - Basic Transform Data

Data profiling

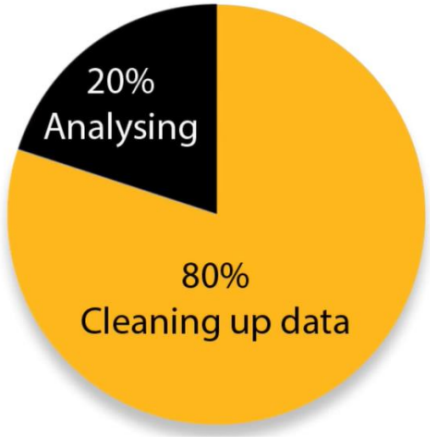


DATA PREPARATION - Basic Transform Data
Query Dependencies



DATA PREPARATION - Basic Transform Data
Cleaning Data

HOW DATA SCIENTISTS SPEND TIME



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

DATA PREPARATION – Data Issue

DATA PREPARATIO
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**

Cross Table /Contingency Table / Pivot Table

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

Regular 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.

DATA PREPARATIO
Understand Structure of Data

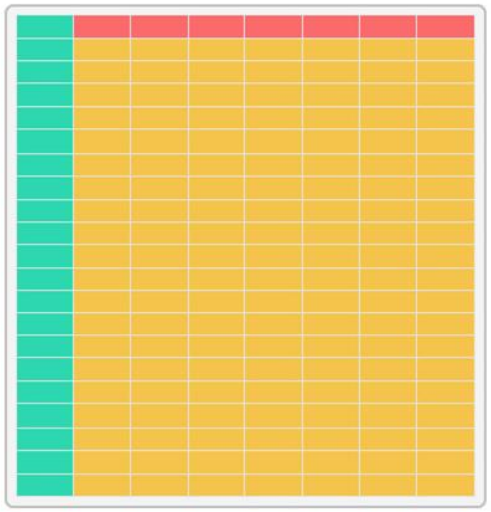
Figure 3.5
Stacked Data

	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	61600	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 PREPARATIO
Understand Structure of Data



Wide format

→

Narrow format

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 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

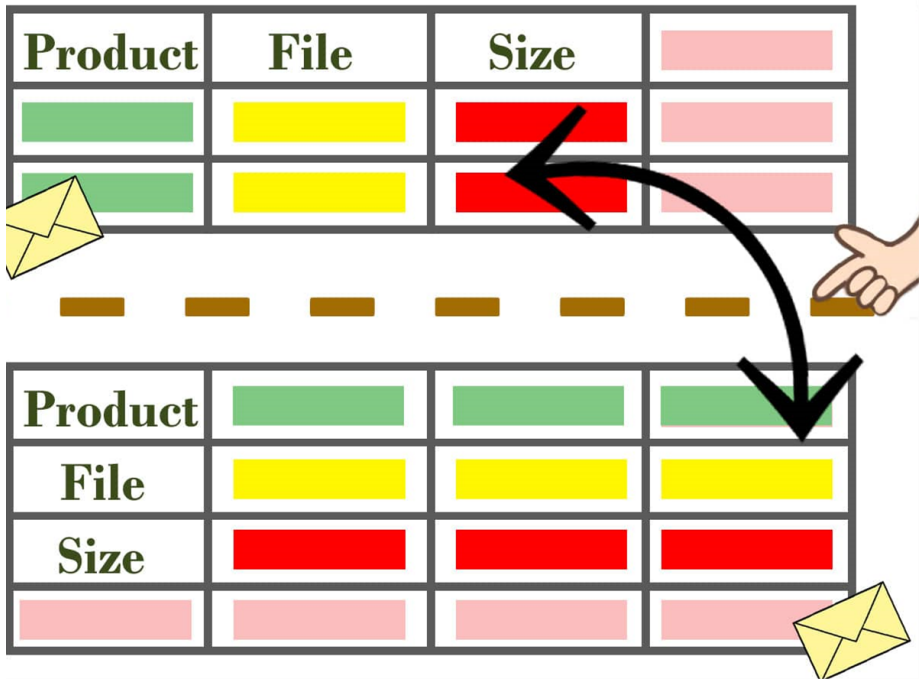
1. **Transpose Table**
2. **Cross Tabulation**
   Pivot + Unpivot
3. **Aggregation (Group by)**

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	

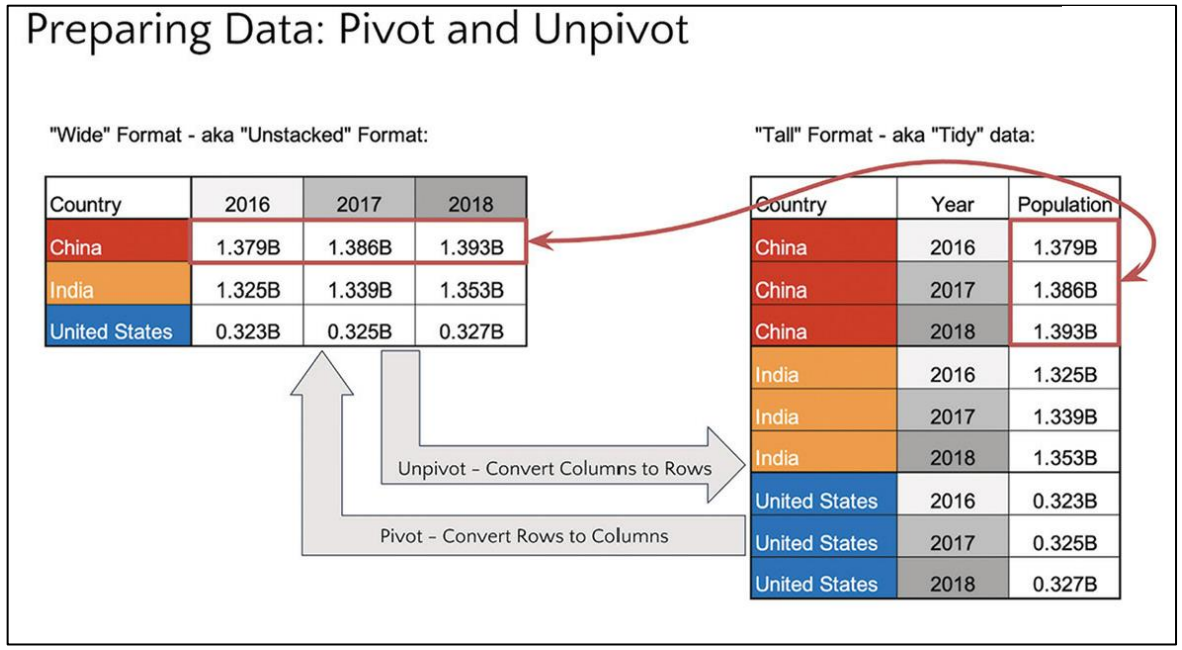
Region	Area	Aspen	Beaut	Bellen	Carlota	Doublers	FlatTop	Quad	Sunbell	Sunset	Sunshine	V-Rang	Yanaki
East	East 1	10									76		
East	East 2		16	77		7						8	42
East	East 3				39			87					
East	East 4						4		44	23			
MidWest	MidWest 1		30	47	57			112			60		
MidWest	MidWest 2	74							47				
MidWest	MidWest 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

MASTERING DATA ANALYTICS

UNPIVOT & PIVOT



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

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

MASTERING DATA ANALYTICS

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	133095

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

MASTERING DATA ANALYTICS

Not Parsed Correctly

Name
Smith, John

→

LastName	FirstName
Smith	John

Extra Characters

Name
"John Smith"

→

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 Street → 125 Main Street

Duplicate Data Records

ID	Name
1	John Smith
1	John Smith
2	Jane Doe

→

ID	Name
1	John Smith
2	Jane Doe

Incorrect Data

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

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

DATA PREPARATION – Data Issue
Dirty Data

MASTERING DATA ANALYTICS

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

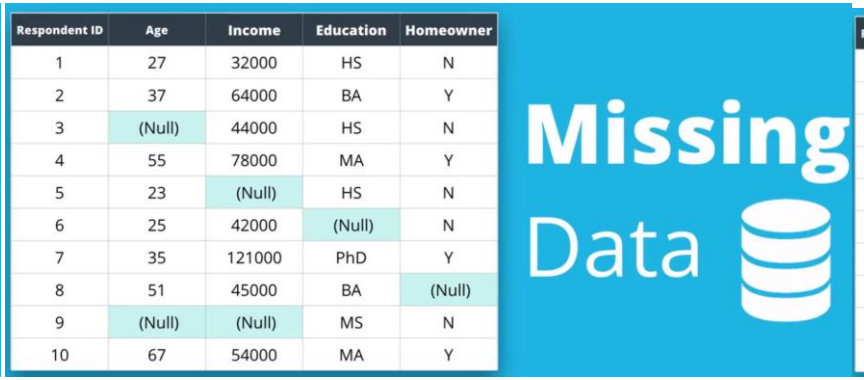
MASTERING DATA ANALYTICS

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

Blank/ Empty cells (CSV)

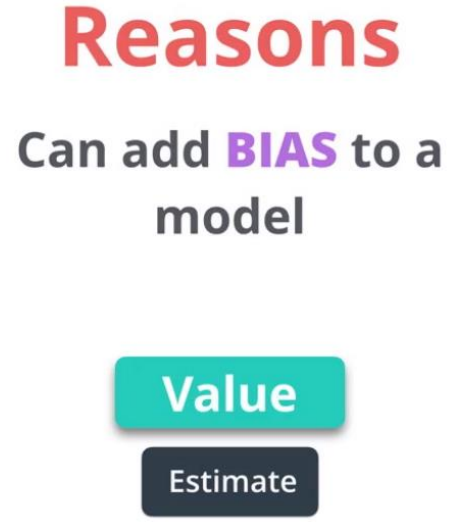
Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3	(Null)	44000	HS	N
4	55	78000	MA	Y
5	23	(Null)	HS	N
6	25	42000	(Null)	N
7	35	121000	PhD	Y
8	51	45000	BA	(Null)
9	(Null)	(Null)	MS	N
10	67	54000	MA	Y



Null value (Database)

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

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*

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

Real 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

Average age without missing values:
44.5

Average age with missing values:
40

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.

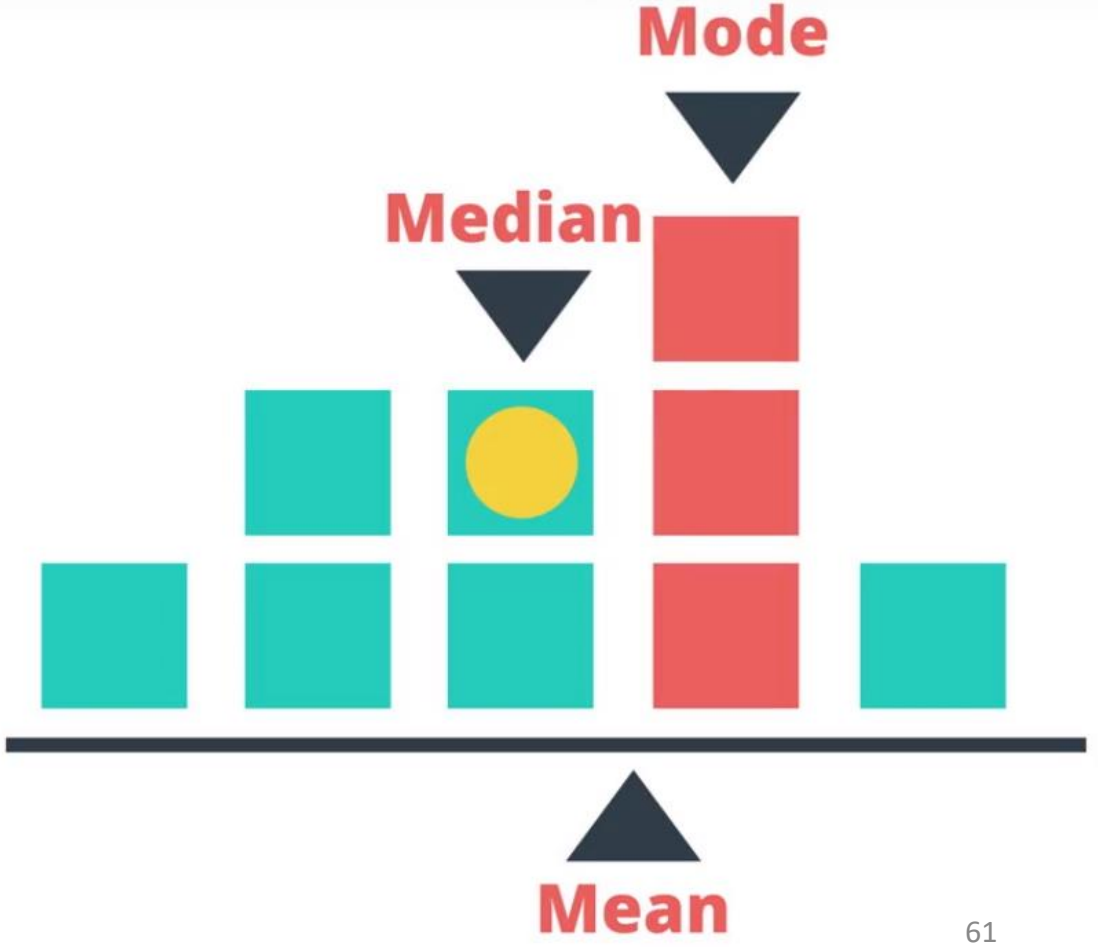
However, you should **make sure** that deleting missing data **doesn't have adverse effects on your analysis.** *For example, if a particular demographic tended to leave a response blank in a survey, then removing records with blank entries will mean that a part of the population is underrepresented.*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

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



DATA PREPARATION – Data Issue
Missing Data – Select the method

What methodology might be the best approach

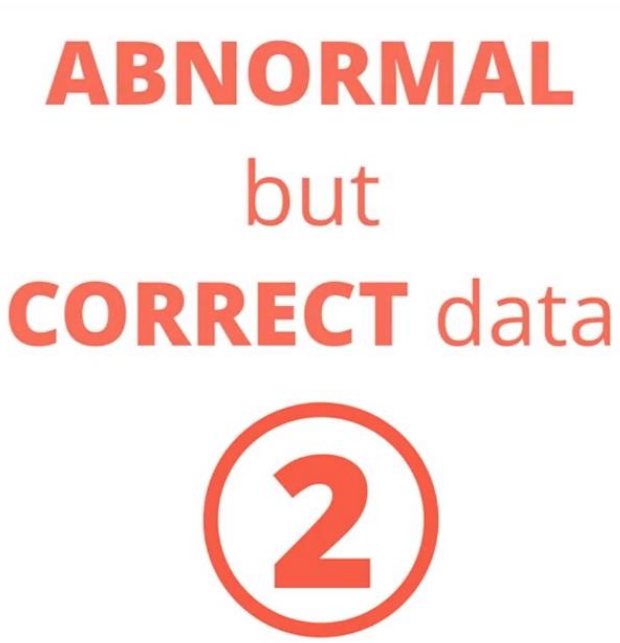
1. **How much data is really missing? (>=20%)**
2. **How the missing data is distributed across the dataset? (2/10 predictor variables missed)**
3. **Whether those specific variables are actually significant to our analysis and model making process**
4. **The missing data is numeric or categorical**

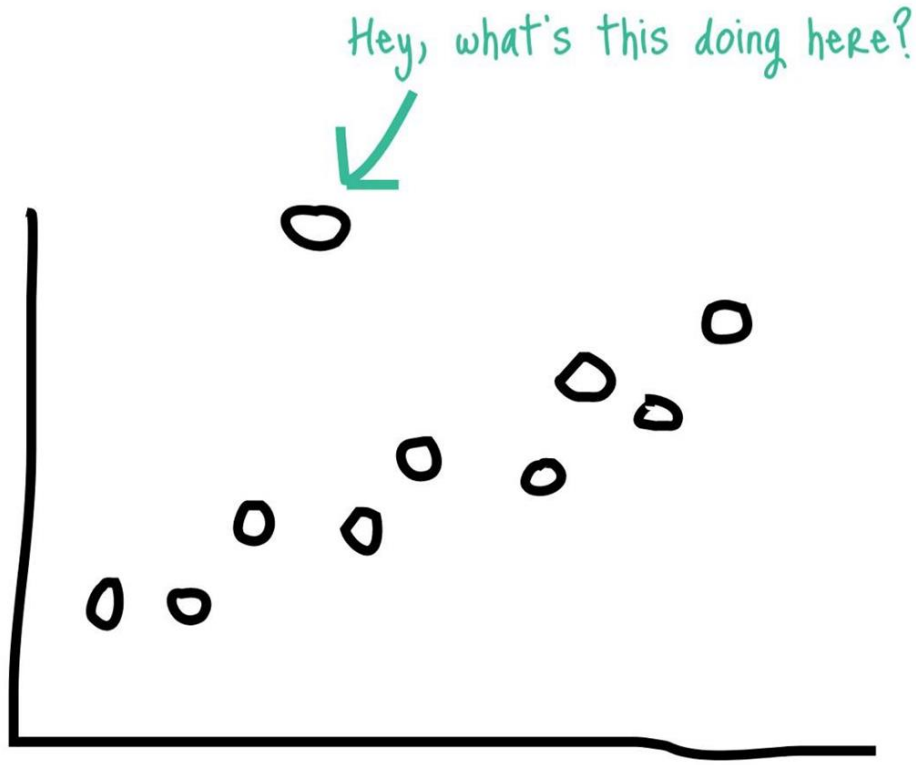
Numeric or **Categorical** ?

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

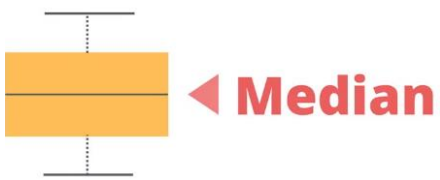
Identifying outliers in the data helps us understand **how vulnerable** our model would be to a **small set of observations.**

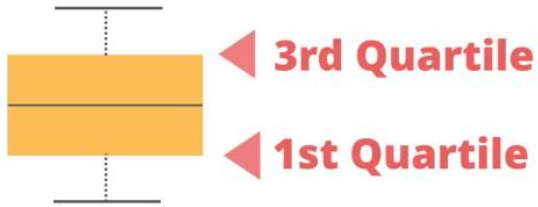




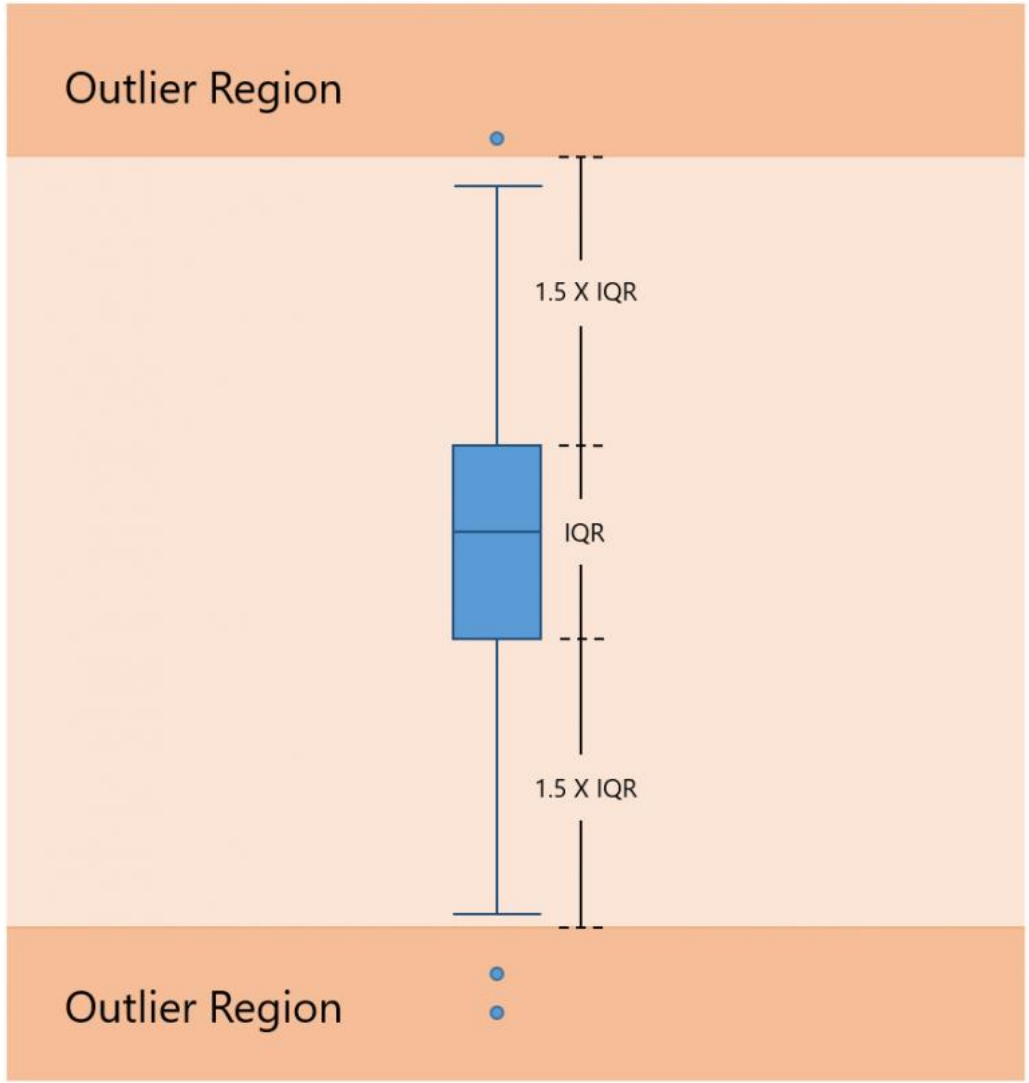
DATA PREPARATION – Data Issue
Outliers

Identifying outliers more **methodically** rather than simply *eyeballing* them
Violin Plot: shows the volume of the distribution
Others: *z-scores or standard deviations*





DATA PREPARATION – Data Issue
Outliers



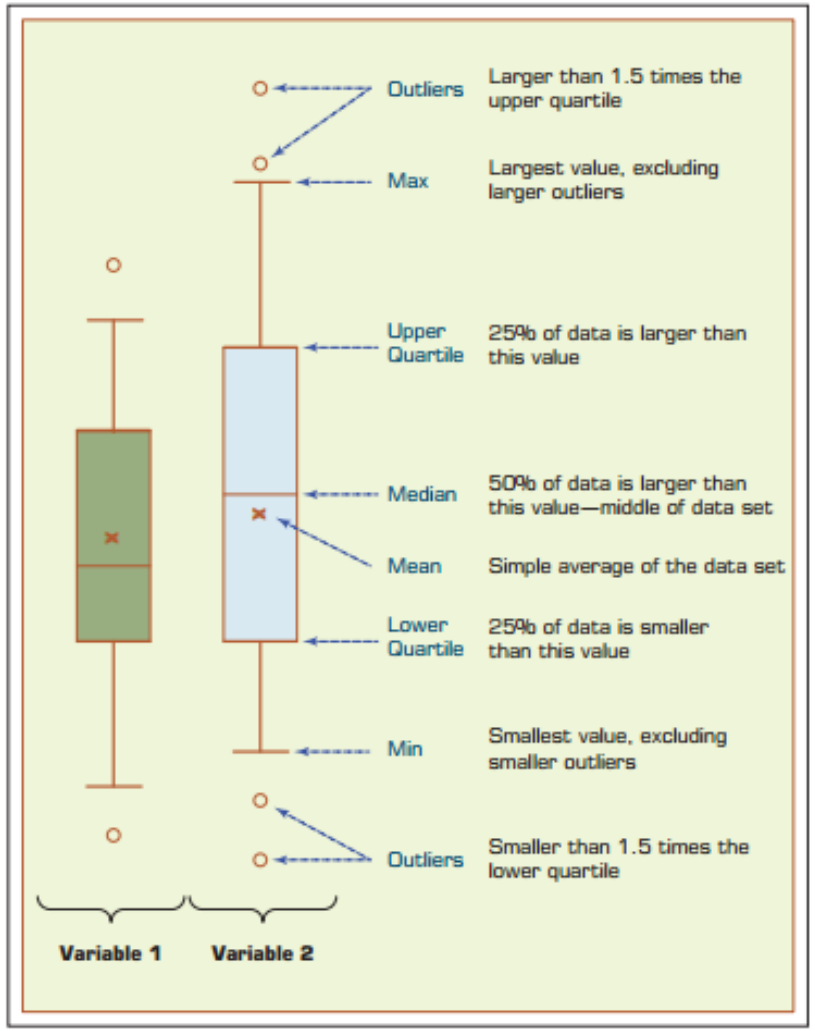


FIGURE 2.8 Understanding the Specifics about Box-and-Whiskers Plots.

DATA PREPARATION – Data Issue
Outliers

Steps

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

- 1 . Calculate 1st quartile `Q1` and 3rd quartile `Q3` of the dataset. You can use the Excel function `QUARTILE.INC` or `QUARTILE.EXC`
- 2 . Calculate the Interquartile Range: `IQR = Q3 - Q1`
- 3 . Add `1.5 * IQR` to Q3 to get the upper fence: `Upper Fence = Q3 + 1.5 * IQR`
- 4 . Subtract `1.5 * IQR` to Q1 to get the lower fence: `Lower Fence = Q1 - 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

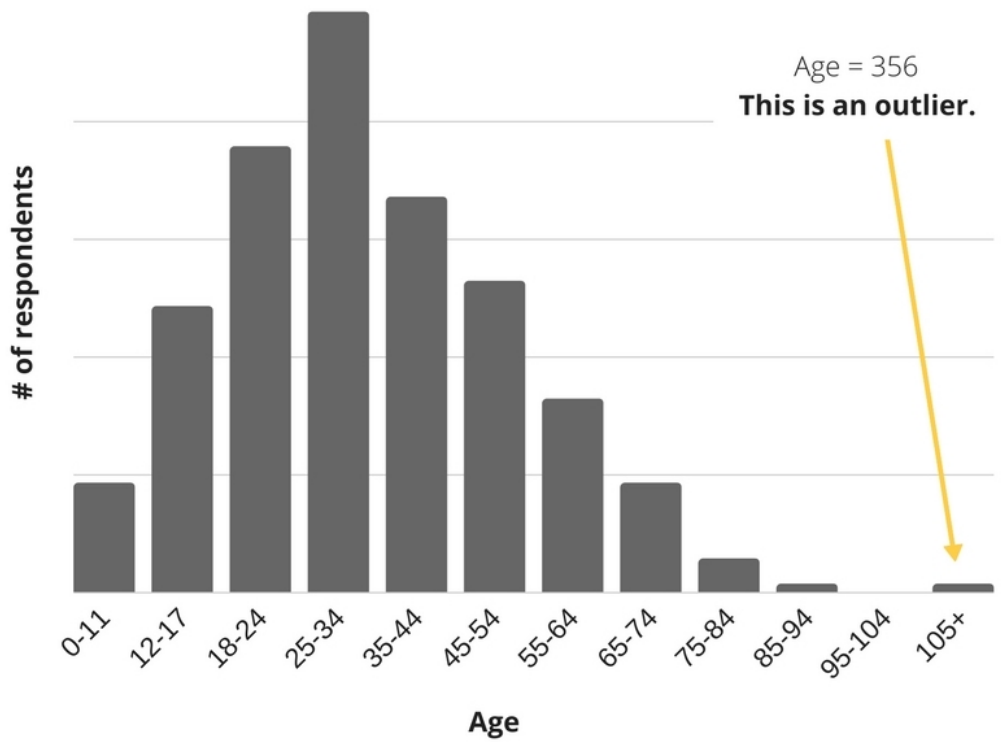
Ex: Age: 299

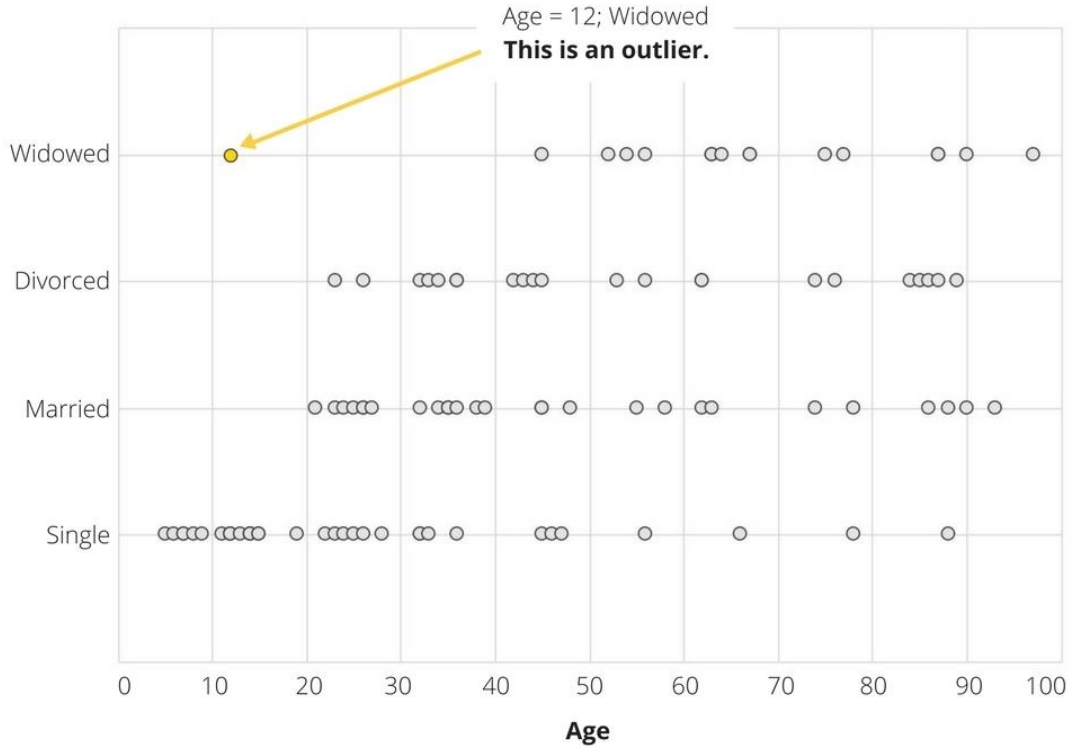
2. Delete the record from the dataset

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

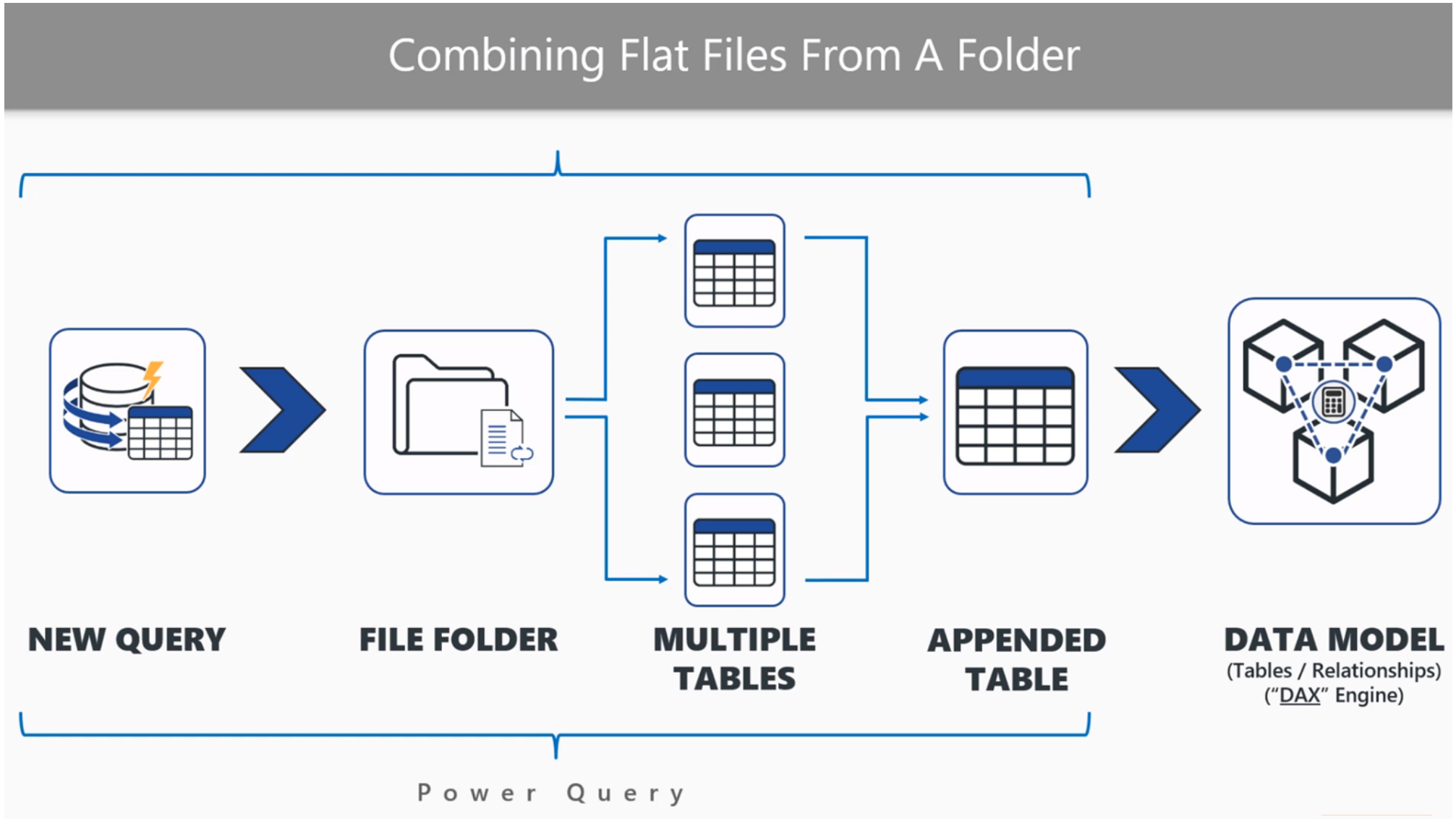
REMOVE Insignificant Outliers

DATA PREPARATION – Data Issue
Outliers



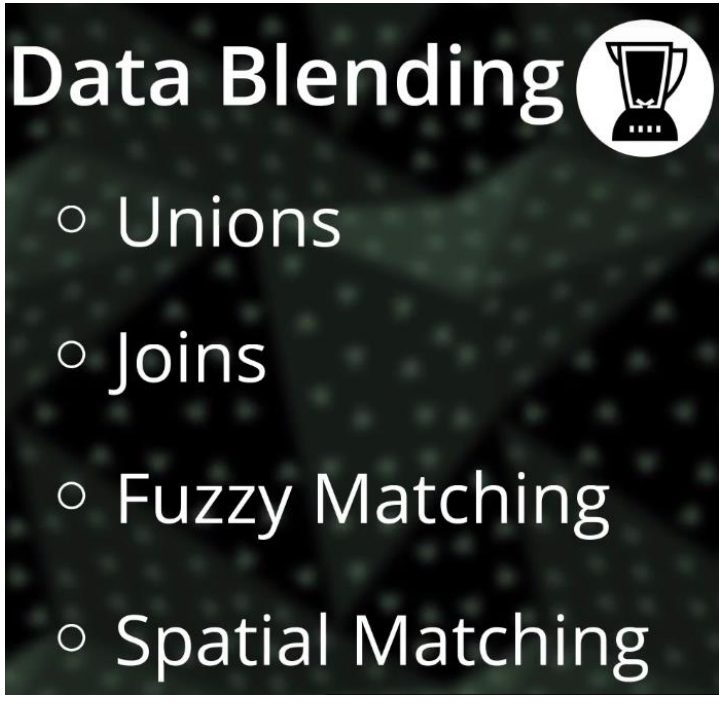


DATA PREPARATION – Data Issue
Combine Data from Folder



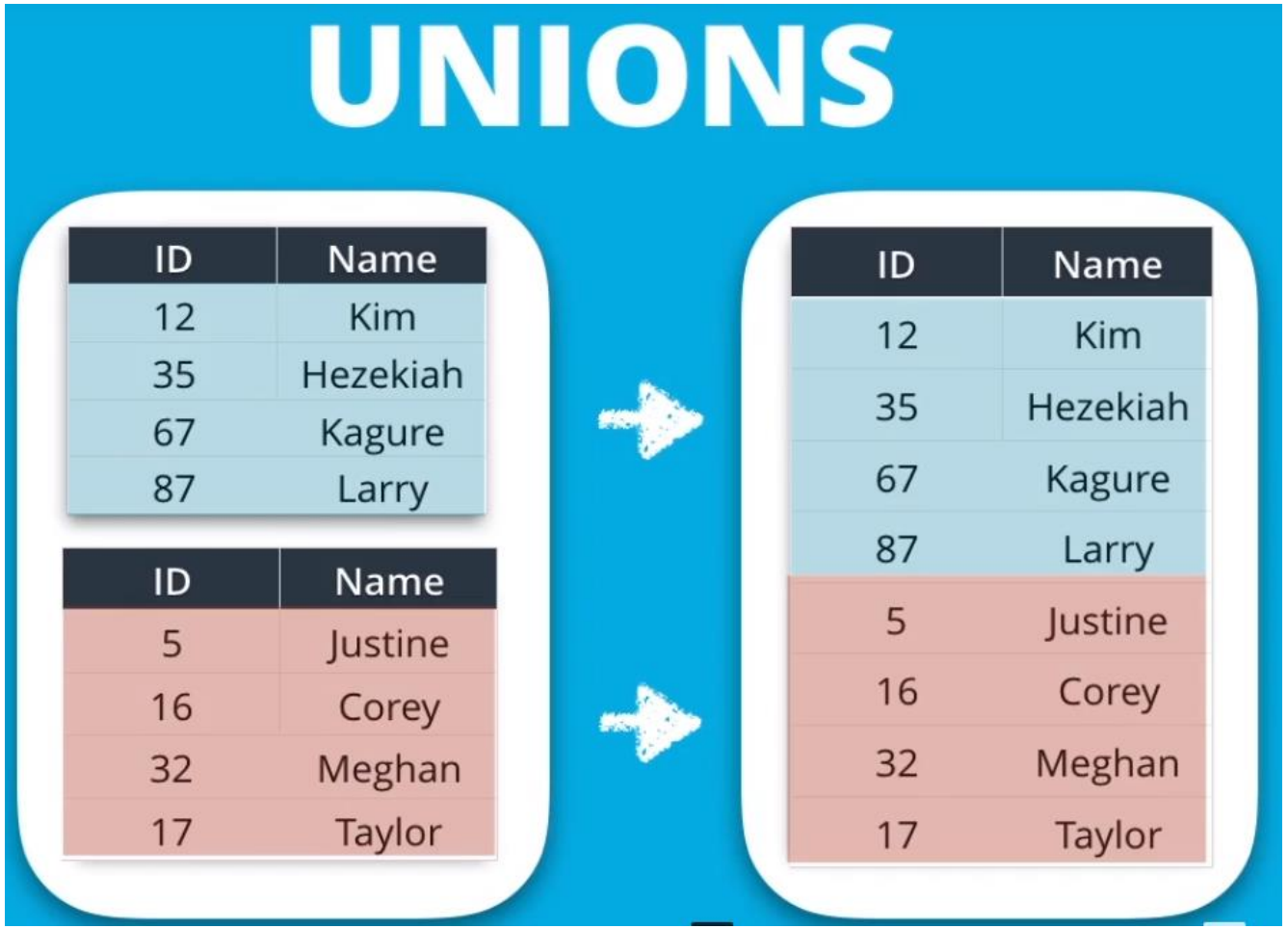
DATA PREPARATION - Data Blending

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



DATA 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

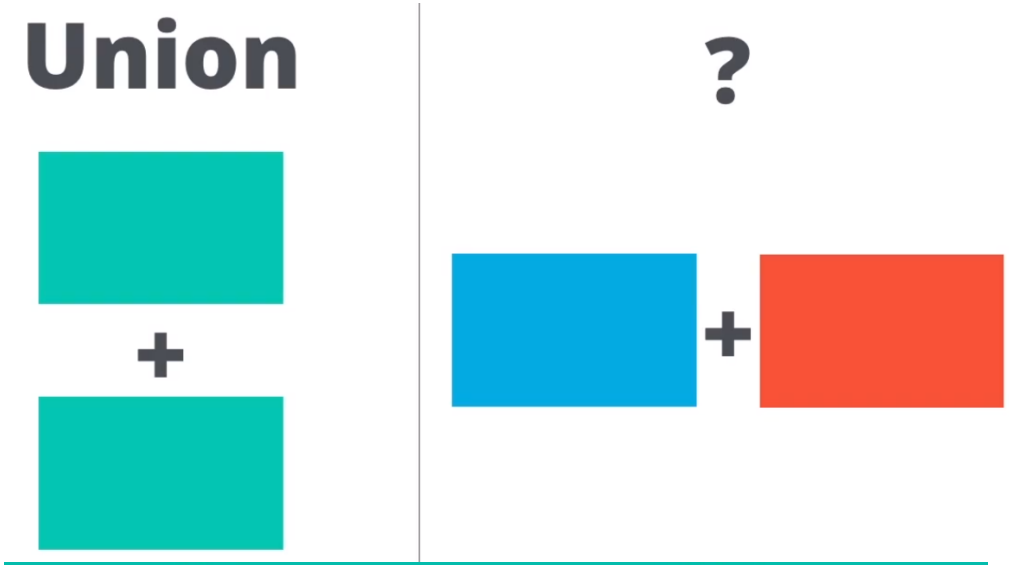


Table 1

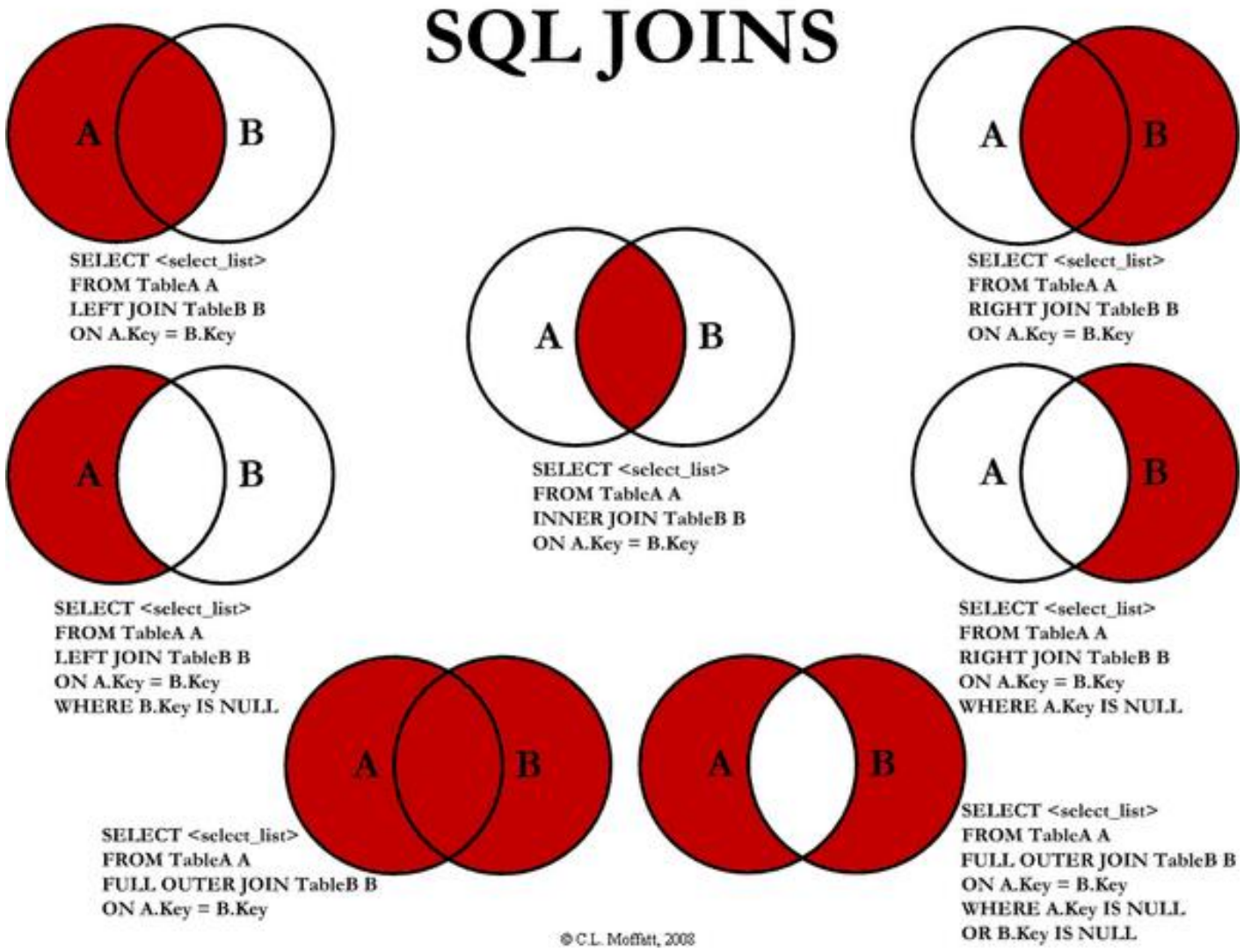
ID	Transaction	Spend
5	5	500
16	10	750
32	1	200
17	2	50

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



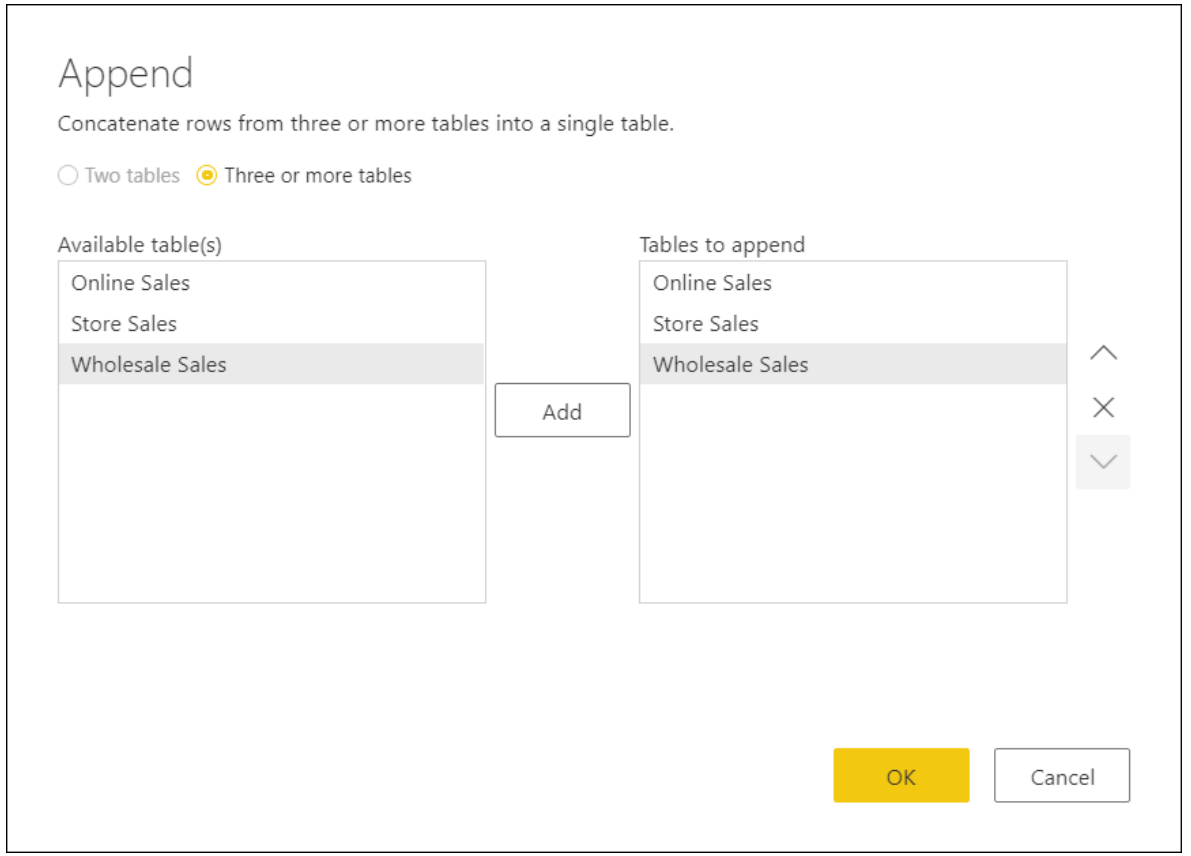
DATA PREPARATION - Data Blending
Union Query

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

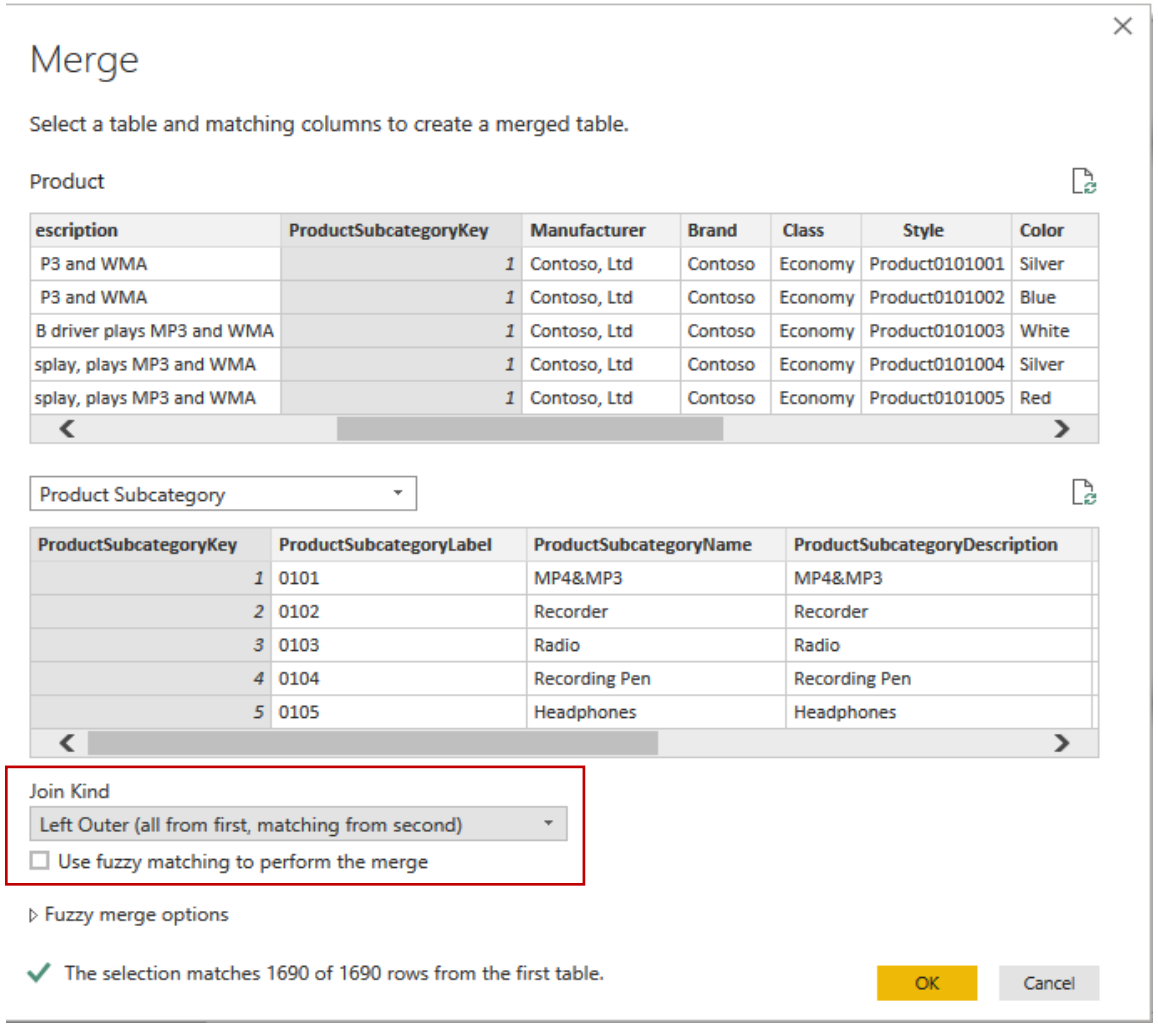
	Channel Name	Date	CustomerID	Units
1	Online	6/2/2020	US-187982	20
2	Online	6/2/2020	US-235789	30
3	Online	6/5/2020	US-187982	15
4	Online	6/5/2020	US-235789	40

- **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



DATA PREPARATION - Data Blending
Merge Query



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.

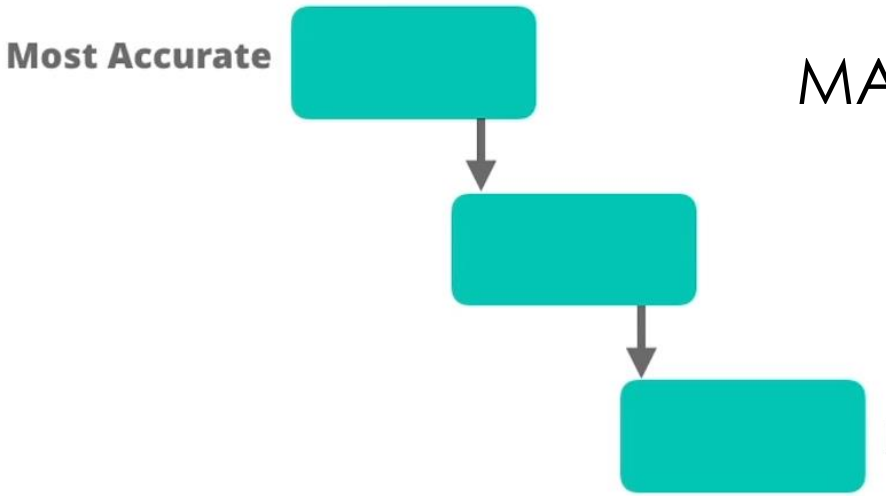
Fuzzy Matching Algorithms

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.



MATCHES

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

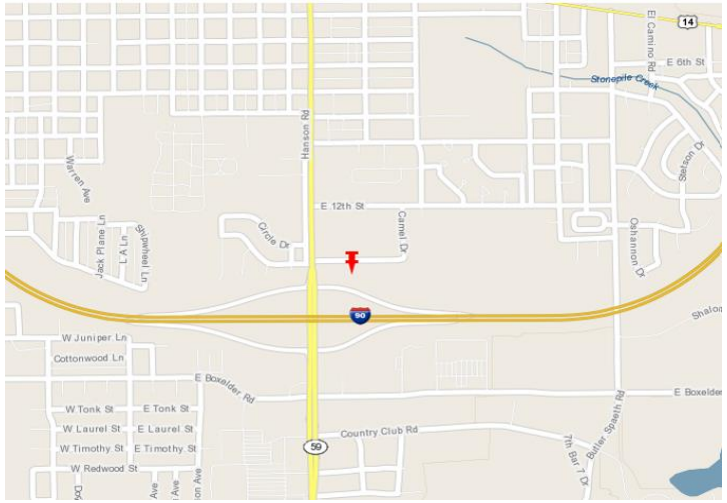
DATA PREPARATION - Data Blending
Spatial Matching

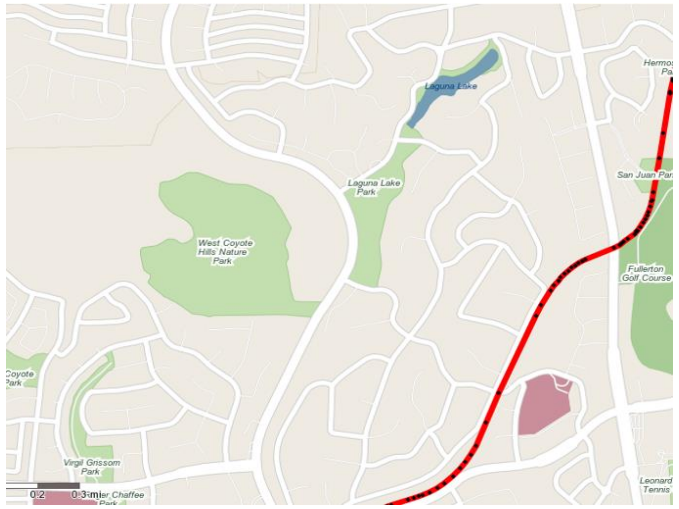
Types of Spatial Data
All of these location data examples are represented by **points, lines, or polygons**

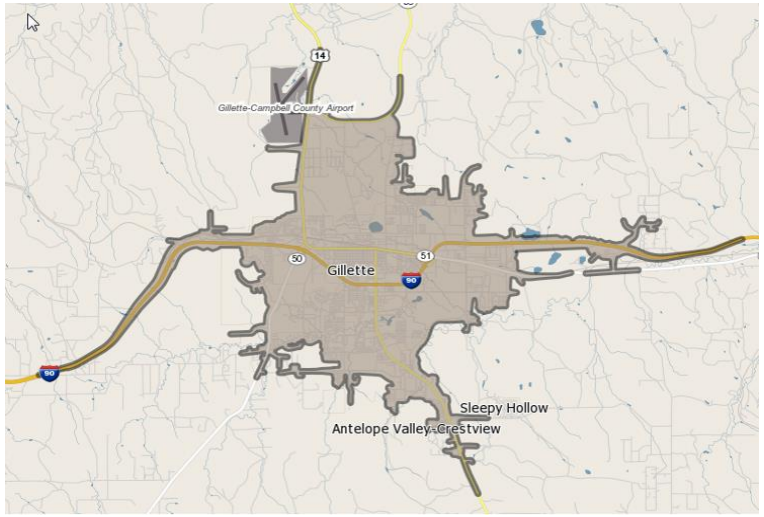
Points
A point, also referred to as a centroid, is in the form of a **latitude and longitude** which we use to **pinpoint** its exact location.

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

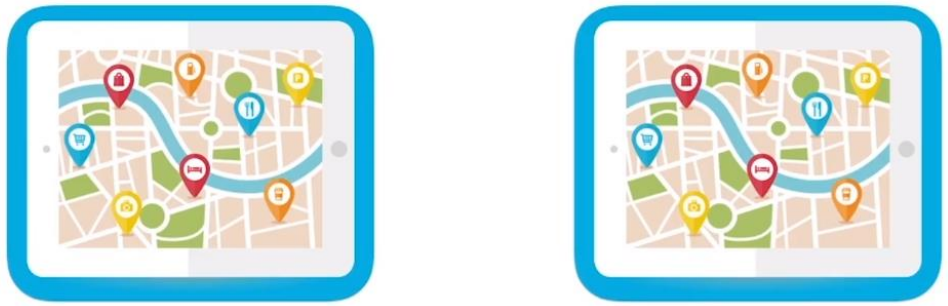
Polygons
Polygons are made up of a **series of longitude and latitude coordinates** defining all of the vertices of a region.



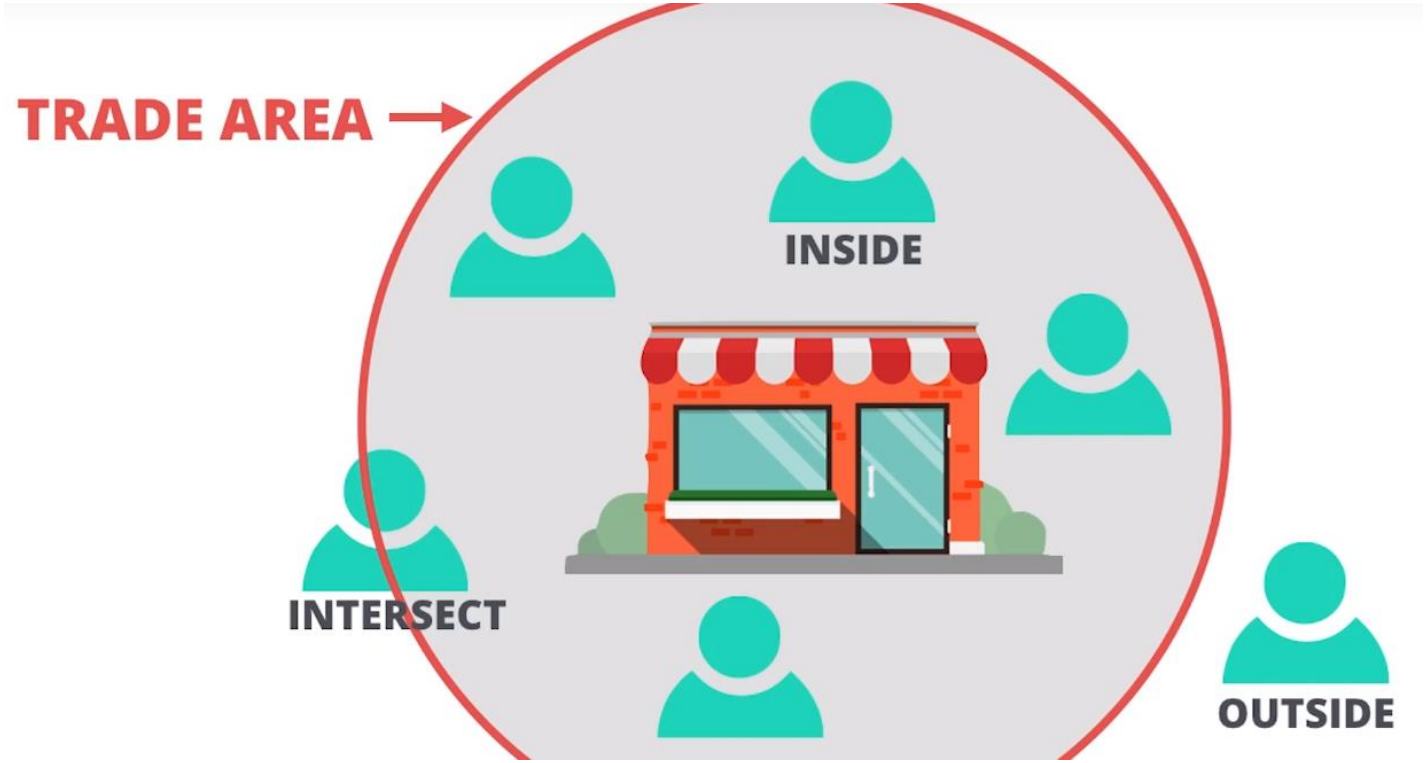




DATA PREPARATION - Data Blending
Spatial Matching



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 characers			
		Unexpected Pattern			
		Incorrect Data			
		Duplicate Data Records			
		Misspelled Entries			
	Missing Data	Deleting Missing Data			
		Imputation			
		Advanced methods			
	Outliers	Errors: Cross-check & fix			
		Errors: Delect			
		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

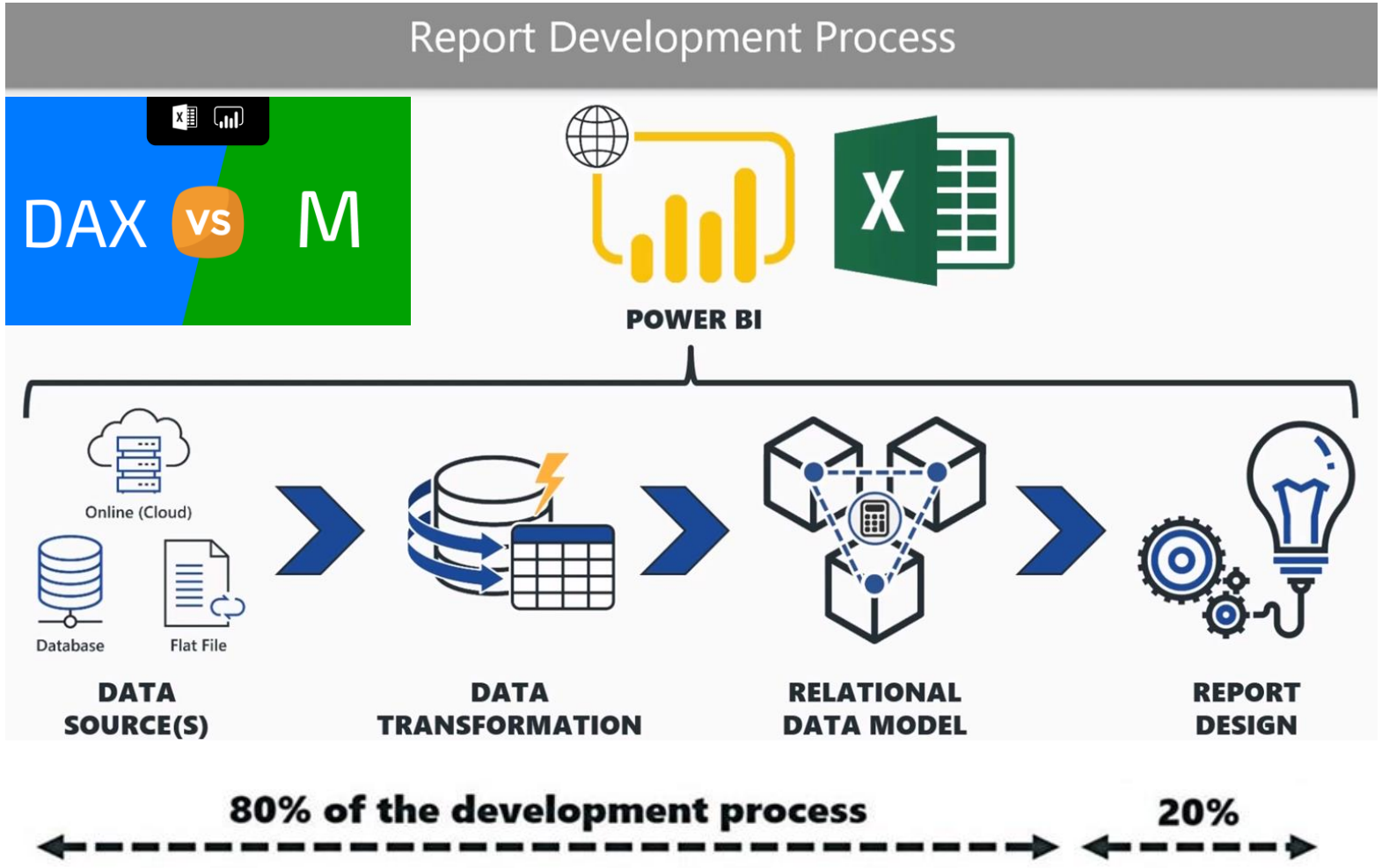
1. Data Model Overview
2. Fact & Dimension
3. Schema
4. Cardinality
5. Cross Filter Direction
6. Hierarchies

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



MASTERING DATA ANALYTICS

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

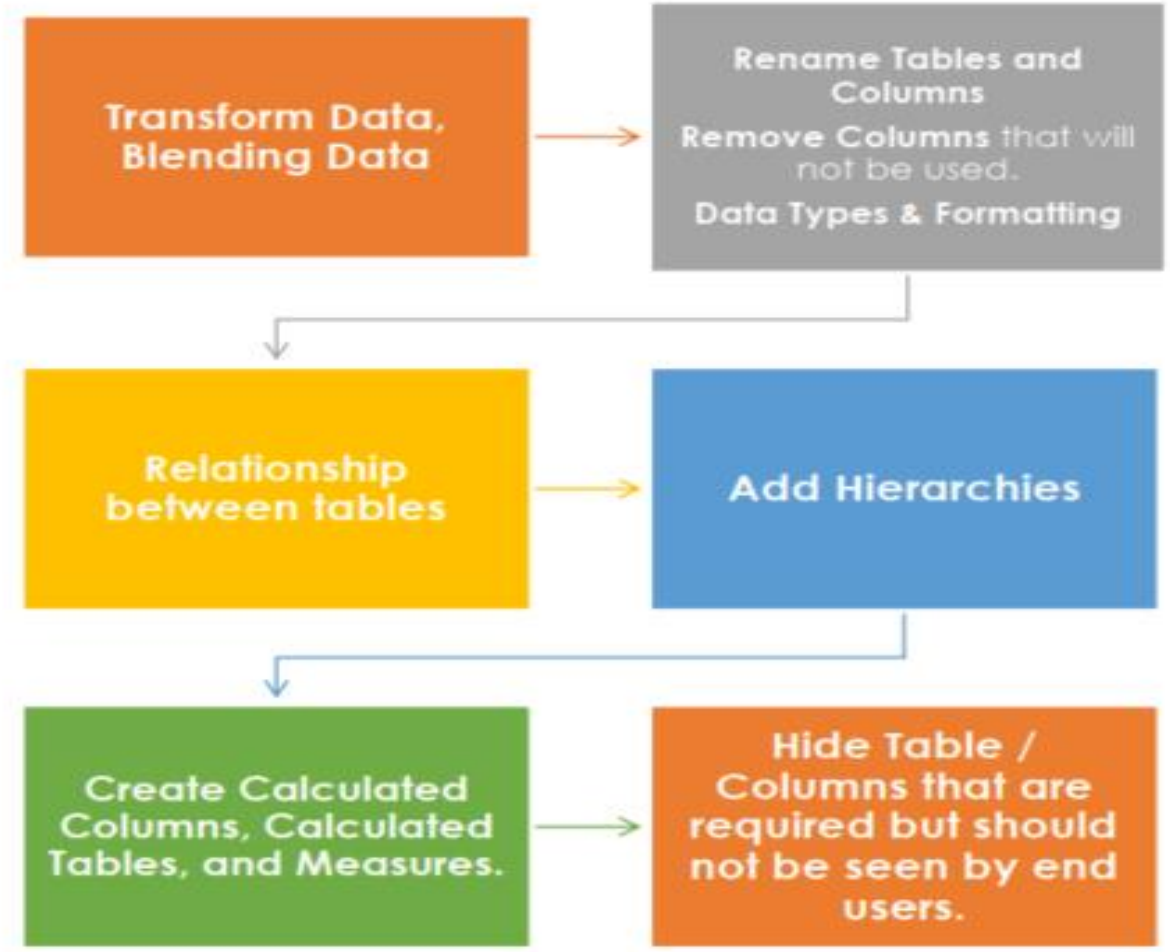
MASTERING DATA ANALYTICS

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



DATA MODEL
Components

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

MASTERING DATA ANALYTICS

Components of a data model – Fact Table

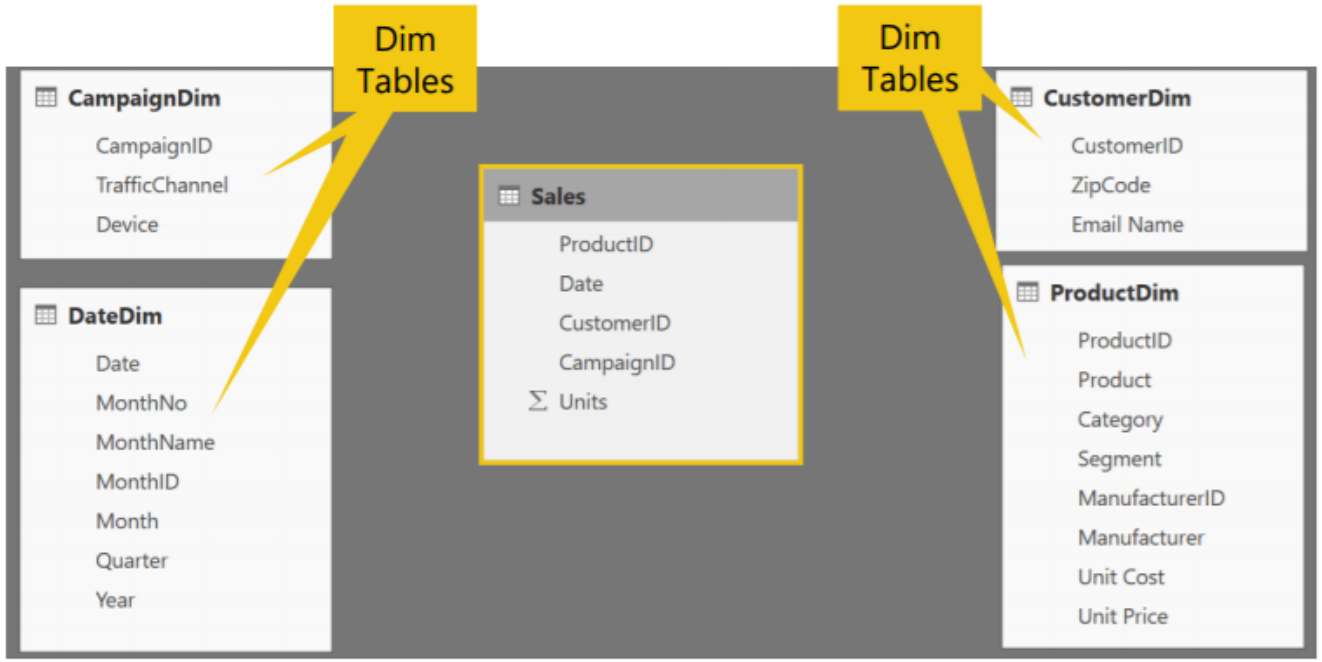
Fact Table

- 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

MASTERING DATA ANALYTICS

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

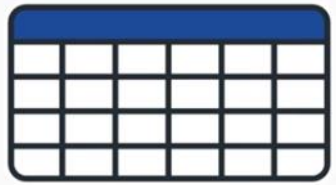
DATA MODEL
Fact Tables and Dimension Tables

MASTERING DATA ANALYTICS

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

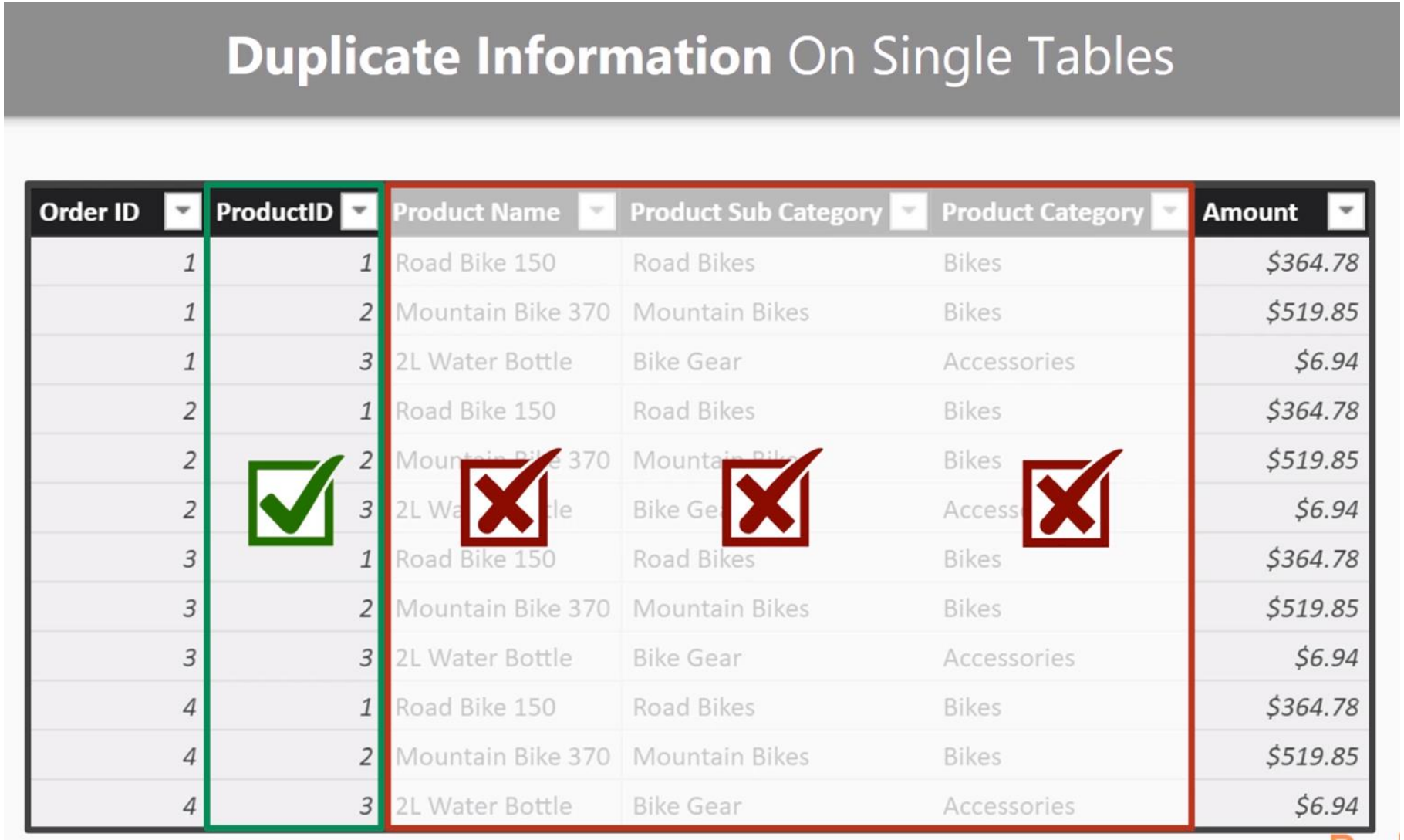
DATA MODEL
Fact Tables and Dimension Tables

MASTERING DATA ANALYTICS

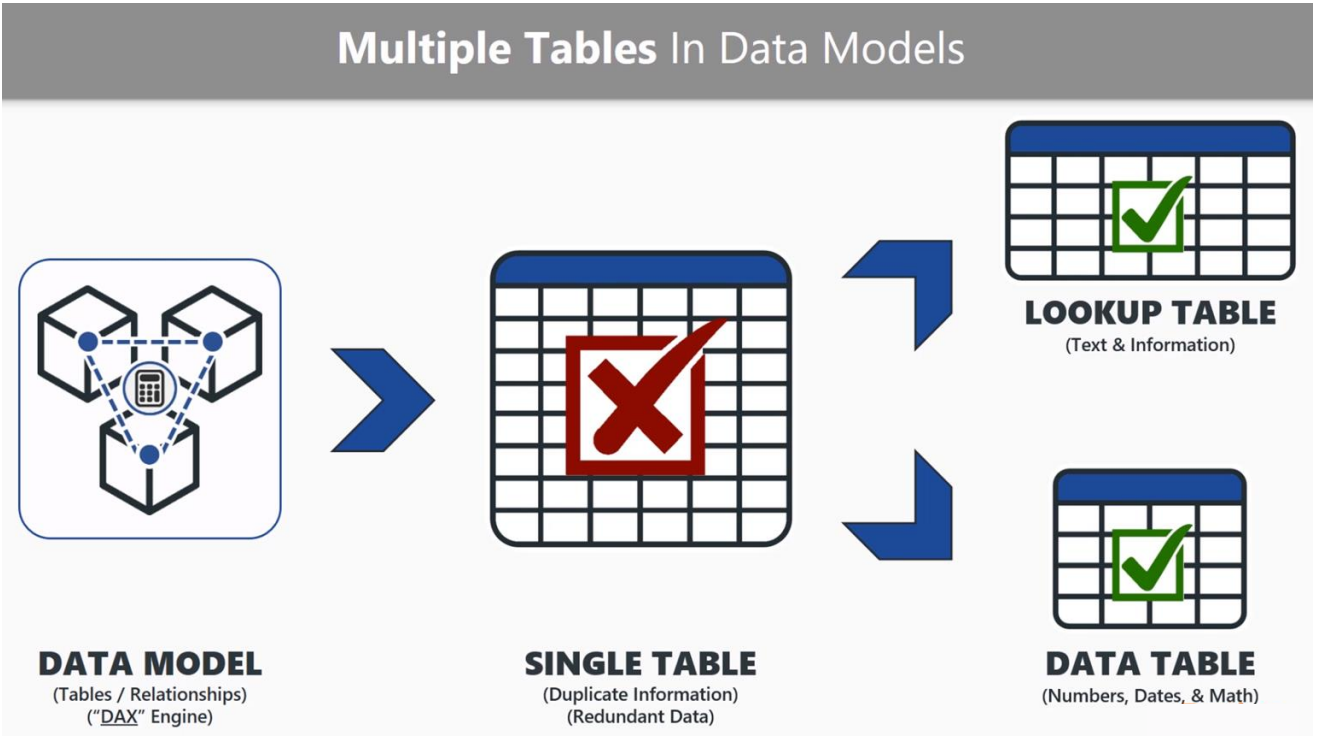
Duplicate Information On Single Tables

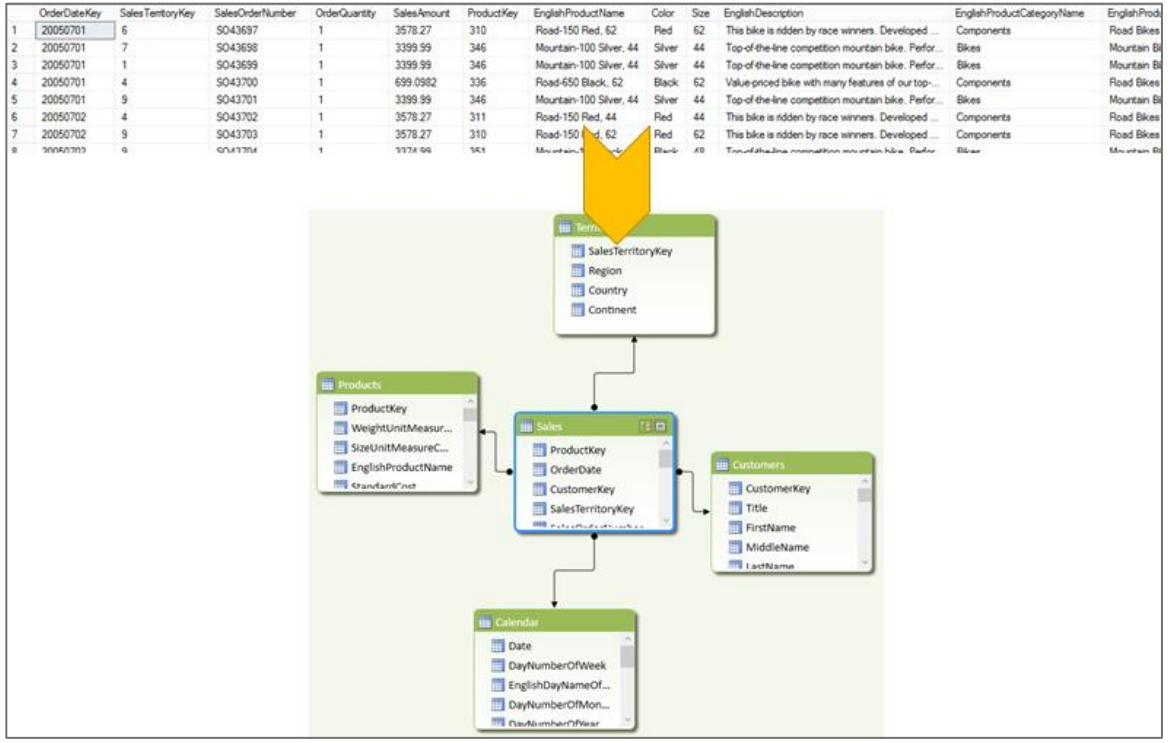
Order ID	ProductID	Product Name	Product Sub Category	Product Category	Amount
1	1	Road Bike 150	Road Bikes	Bikes	$364.78
1	2	Mountain Bike 370	Mountain Bikes	Bikes	$519.85
1	3	2L Water Bottle	Bike Gear	Accessories	$6.94
2	1	Road Bike 150	Road Bikes	Bikes	$364.78
2	2	Mountain Bike 370	Mountain Bikes	Bikes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



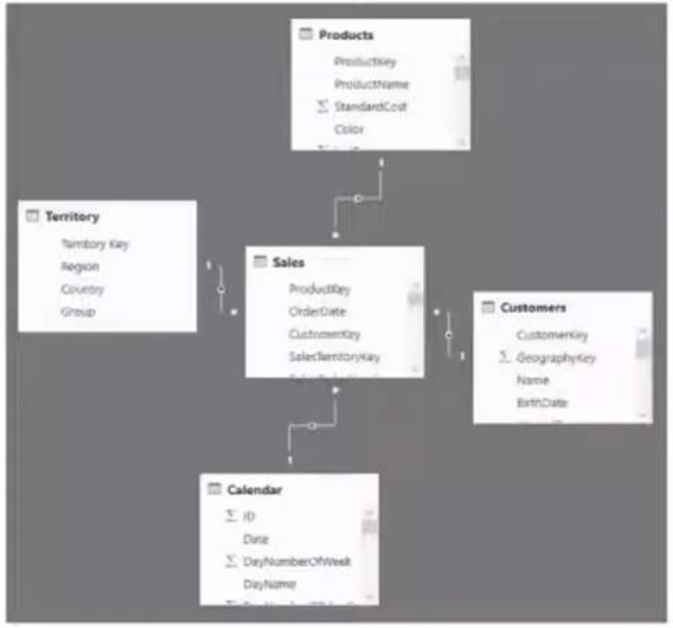
DATA MODEL
Fact Tables and Dimension Tables



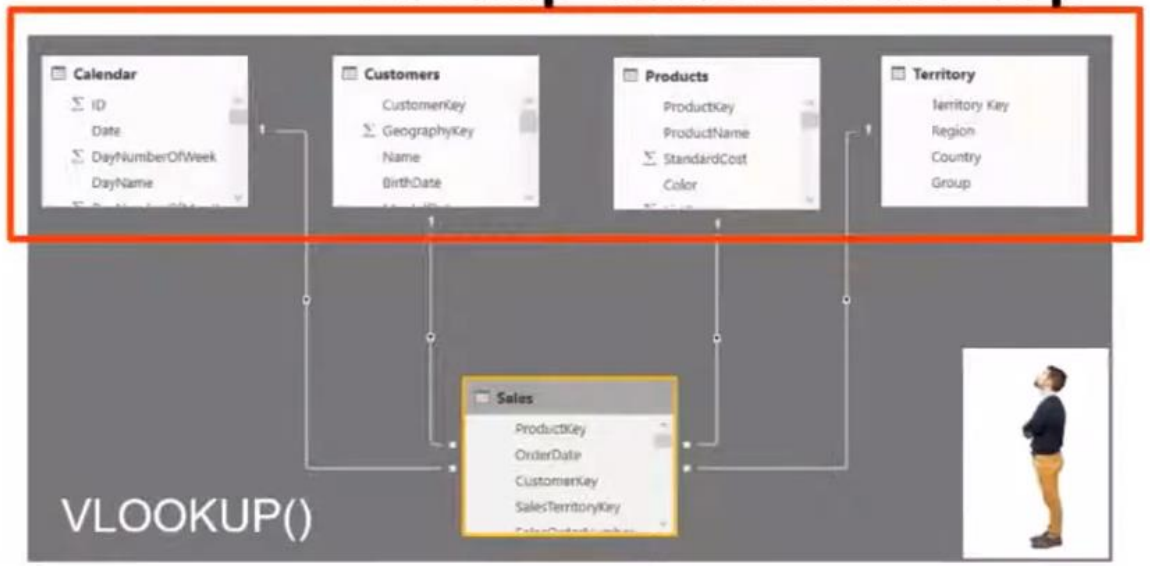


Star Schema is the Optimal Structure

One Entity, One Table

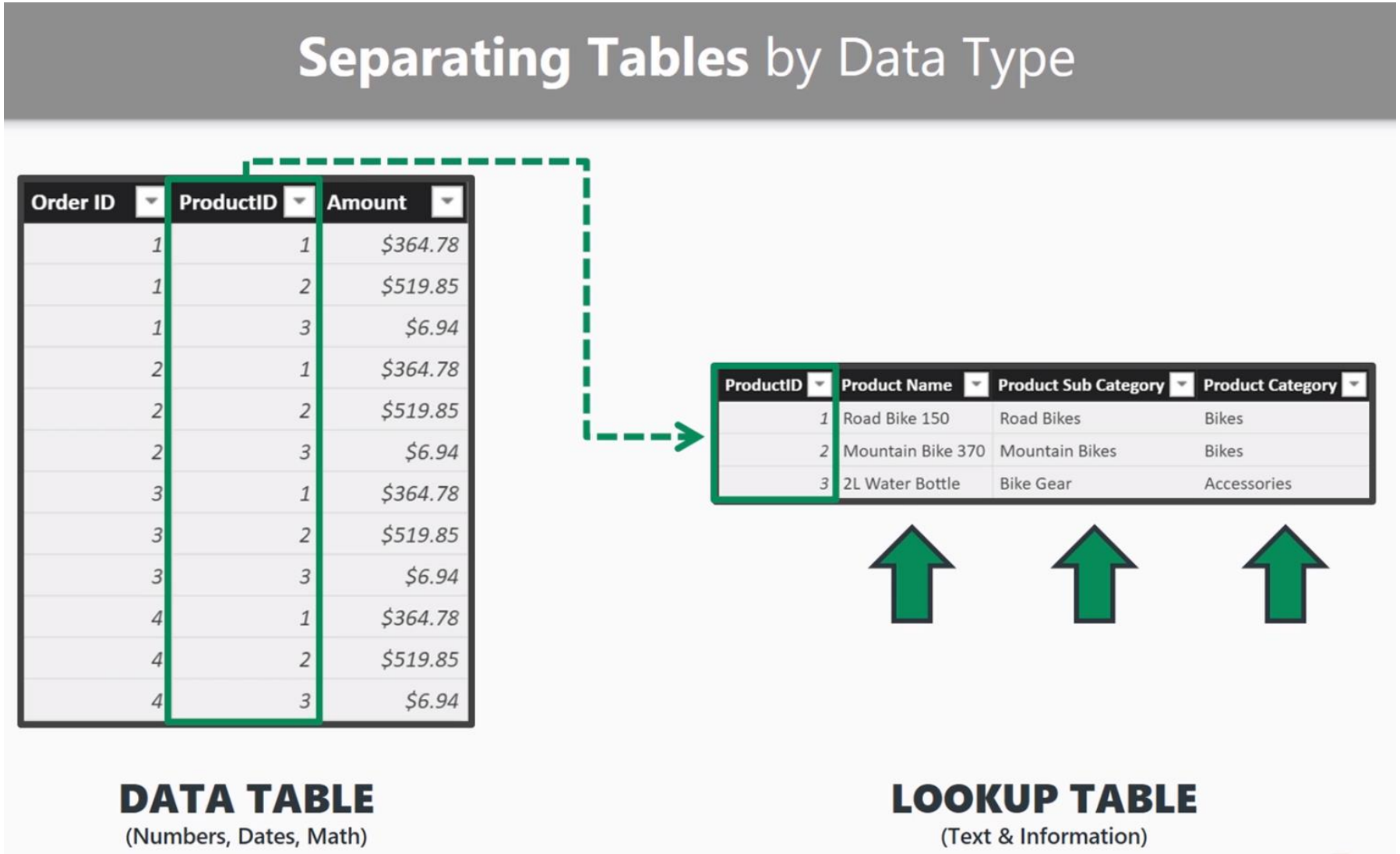


Lookup Tables at the Top



- 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



DATA MODEL
Fact Tables and Dimension Tables

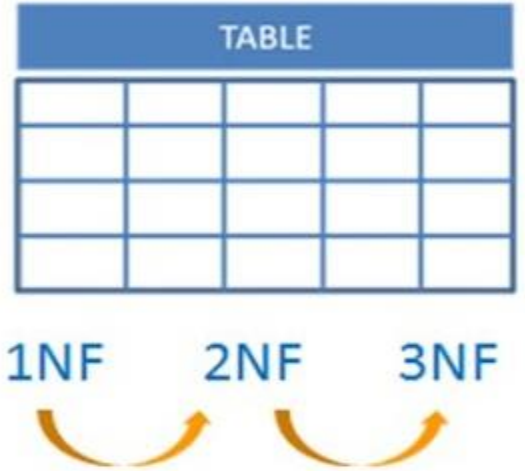
OLTP

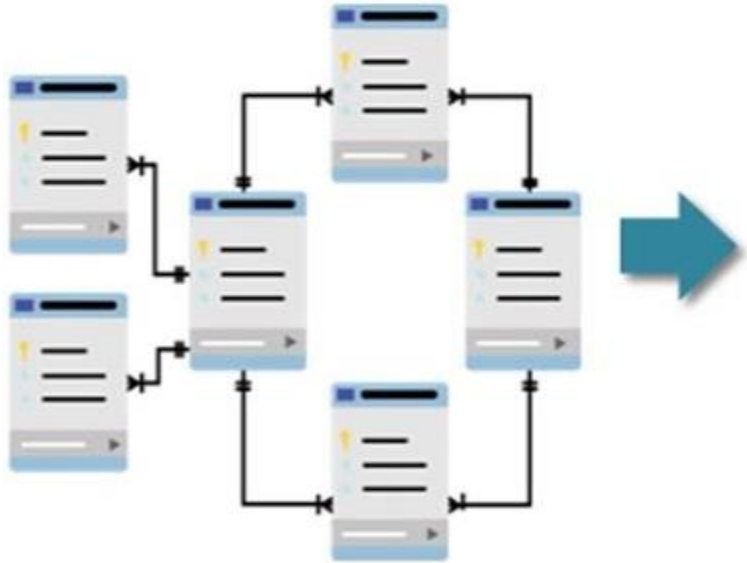
Data Warehouse

ID No.	Name	D.o.B.	Phone	Class	Tutor	Room
356	Jess	3 Mar 1995	7564356	5B	Mr Noggin	56
412	Hamad	12 Nov 1994	7465846	5B	Mr Noggin	56
459	Sita	9 Jan 1994	8565634	6Y	Ms Take	18
502	Hamad	3 Mar 1995	6554546	5B	Mr Noggin	56

Flat Table

Database Normalization





MASTERING DATA ANALYTICS

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

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

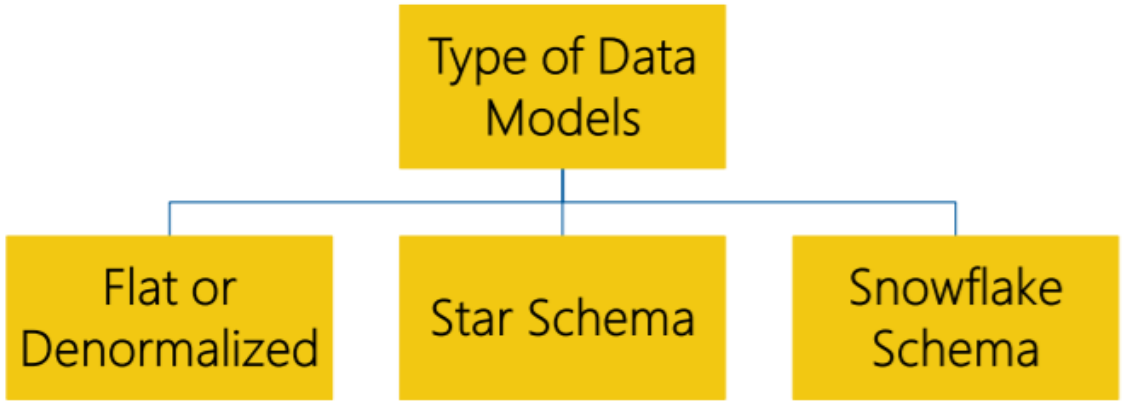
If it won't be aggregated, it's a Dimension

DATA MODEL

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

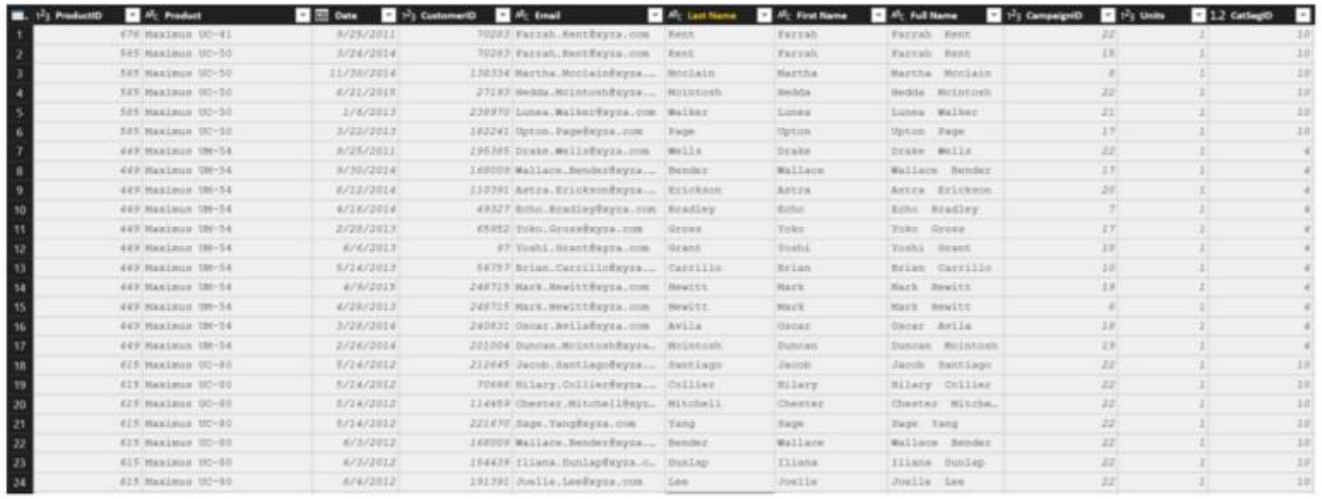
MASTERING DATA ANALYTICS

Data Model Brings Facts and Dimensions Together



MASTERING DATA ANALYTICS

Flat or Denormalized Schema

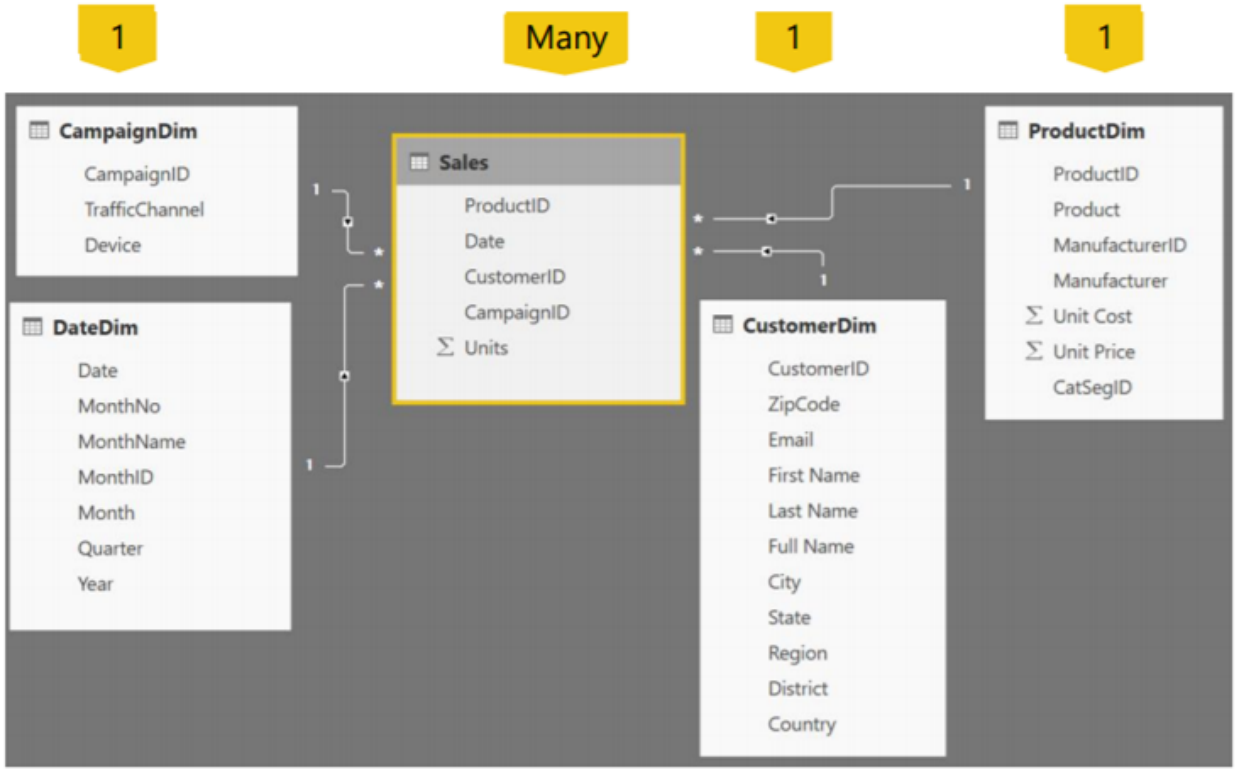


	ProductID	Product	Date	CustomerID	Email	Last Name	First Name	Full Name	CampaignID	Units	CatSegID
1	676	Maximus UC-41	9/25/2011	70283	Farrah.Kent@xyza.com	Kent	Farrah	Farrah Kent	22	1	10
2	585	Maximus UC-50	3/26/2014	70283	Farrah.Kent@xyza.com	Kent	Farrah	Farrah Kent	19	1	10
3	585	Maximus UC-50	11/30/2014	138334	Martha.Mcclain@xyza...	Mcclain	Martha	Martha Mcclain	6	1	10
4	585	Maximus UC-50	6/21/2013	27183	Hedda.Mcintosh@xyza...	Mcintosh	Hedda	Hedda Mcintosh	22	1	10
5	585	Maximus UC-50	1/6/2013	238970	Lunea.Walker@xyza.com	Walker	Lunea	Lunea Walker	21	1	10
6	585	Maximus UC-50	3/22/2013	162241	Upton.Page@xyza.com	Page	Upton	Upton Page	17	1	10
7	449	Maximus UM-54	9/25/2011	196385	Drake.Wells@xyza.com	Wells	Drake	Drake Wells	22	1	4
8	449	Maximus UM-54	9/30/2014	168009	Wallace.Bender@xyza...	Bender	Wallace	Wallace Bender	17	1	4
9	449	Maximus UM-54	8/12/2014	110391	Astra.Erickson@xyza...	Erickson	Astra	Astra Erickson	20	1	4
10	449	Maximus UM-54	6/16/2014	49327	Echo.Bradley@xyza.com	Bradley	Echo	Echo Bradley	7	1	4
11	449	Maximus UM-54	2/28/2013	65952	Yoko.Gross@xyza.com	Gross	Yoko	Yoko Gross	17	1	4
12	449	Maximus UM-54	6/6/2013	87	Yoshi.Grant@xyza.com	Grant	Yoshi	Yoshi Grant	13	1	4
13	449	Maximus UM-54	5/14/2013	56757	Brian.Carrillo@xyza...	Carrillo	Brian	Brian Carrillo	10	1	4
14	449	Maximus UM-54	6/6/2013	248725	Mark.Hewitt@xyza.com	Hewitt	Mark	Mark Hewitt	19	1	4
15	449	Maximus UM-54	6/26/2013	248725	Mark.Hewitt@xyza.com	Hewitt	Mark	Mark Hewitt	6	1	4
16	449	Maximus UM-54	3/26/2014	240831	Oscar.Avila@xyza.com	Avila	Oscar	Oscar Avila	18	1	4
17	449	Maximus UM-54	2/26/2014	201004	Duncan.Mcintosh@xyza...	Mcintosh	Duncan	Duncan Mcintosh	19	1	4
18	615	Maximus UC-80	5/14/2012	212645	Jacob.Santiago@xyza...	Santiago	Jacob	Jacob Santiago	22	1	10
19	615	Maximus UC-80	5/14/2012	70686	Hilary.Collier@xyza...	Collier	Hilary	Hilary Collier	22	1	10
20	615	Maximus UC-80	5/14/2012	114459	Chester.Mitchell@xyz...	Mitchell	Chester	Chester Mitche...	22	1	10
21	615	Maximus UC-80	5/14/2012	221470	Sage.Yang@xyza.com	Yang	Sage	Sage Yang	22	1	10
22	615	Maximus UC-80	6/3/2012	168009	Wallace.Bender@xyza...	Bender	Wallace	Wallace Bender	22	1	10
23	615	Maximus UC-80	6/3/2012	194479	Iliana.Dunlap@xyza.c...	Dunlap	Iliana	Iliana Dunlap	22	1	10
24	615	Maximus UC-80	6/6/2012	191391	Joelle.Lee@xyza.com	Lee	Joelle	Joelle 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

MASTERING DATA ANALYTICS

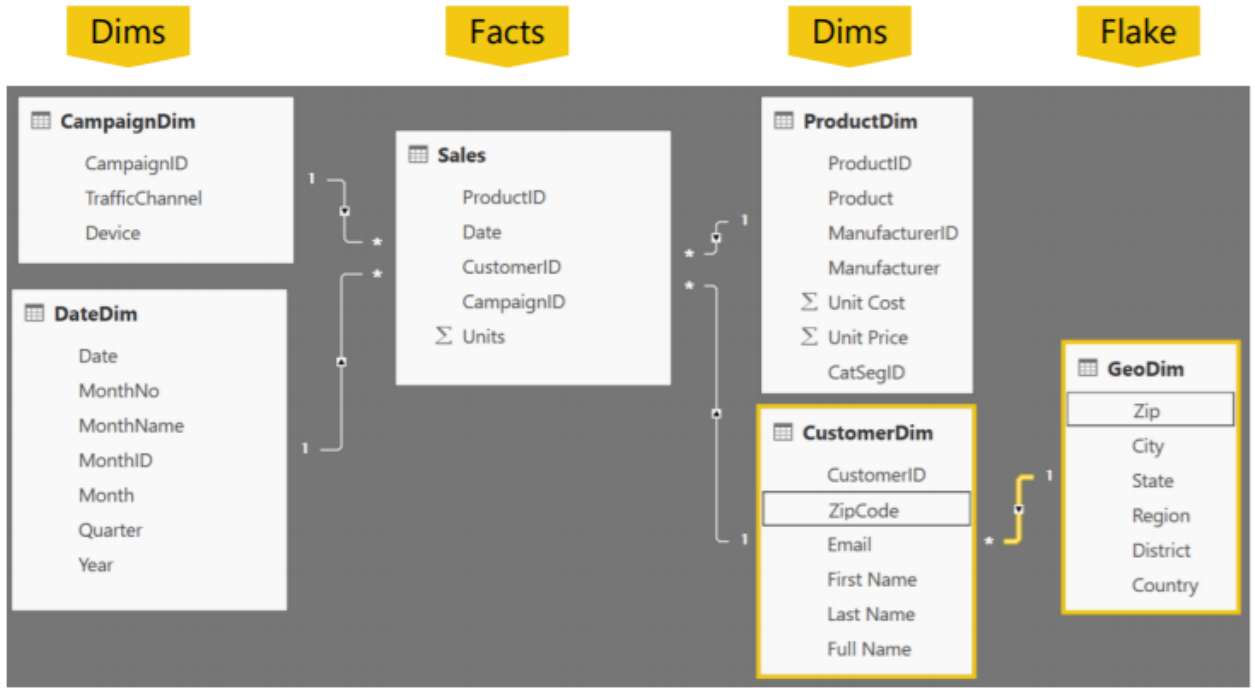
Star Schema



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

MASTERING DATA ANALYTICS

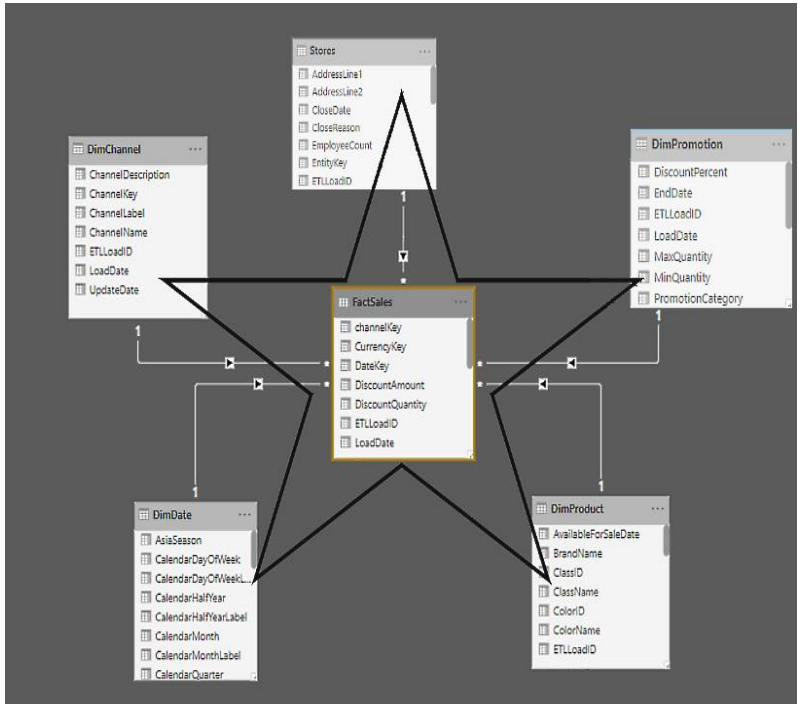
Snowflake Schema



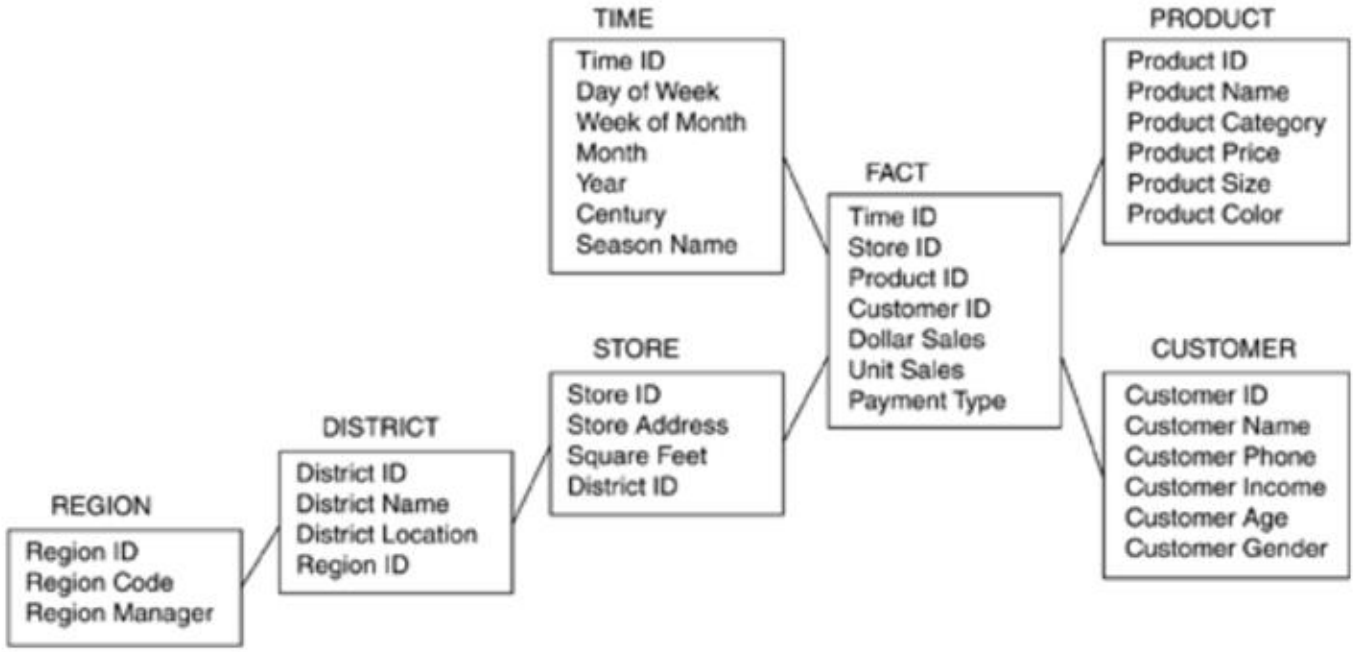
- Center is a Star schema
- **Fact** table in middle
- Surrounded by **Dims**
- Dims "snowflake" off of other Dims
- If you have many, it looks like a 'Snowflake'
- Dim or Fact tables can be the "Many" side of the relationship

DATA MODEL
Schema

Star Schema

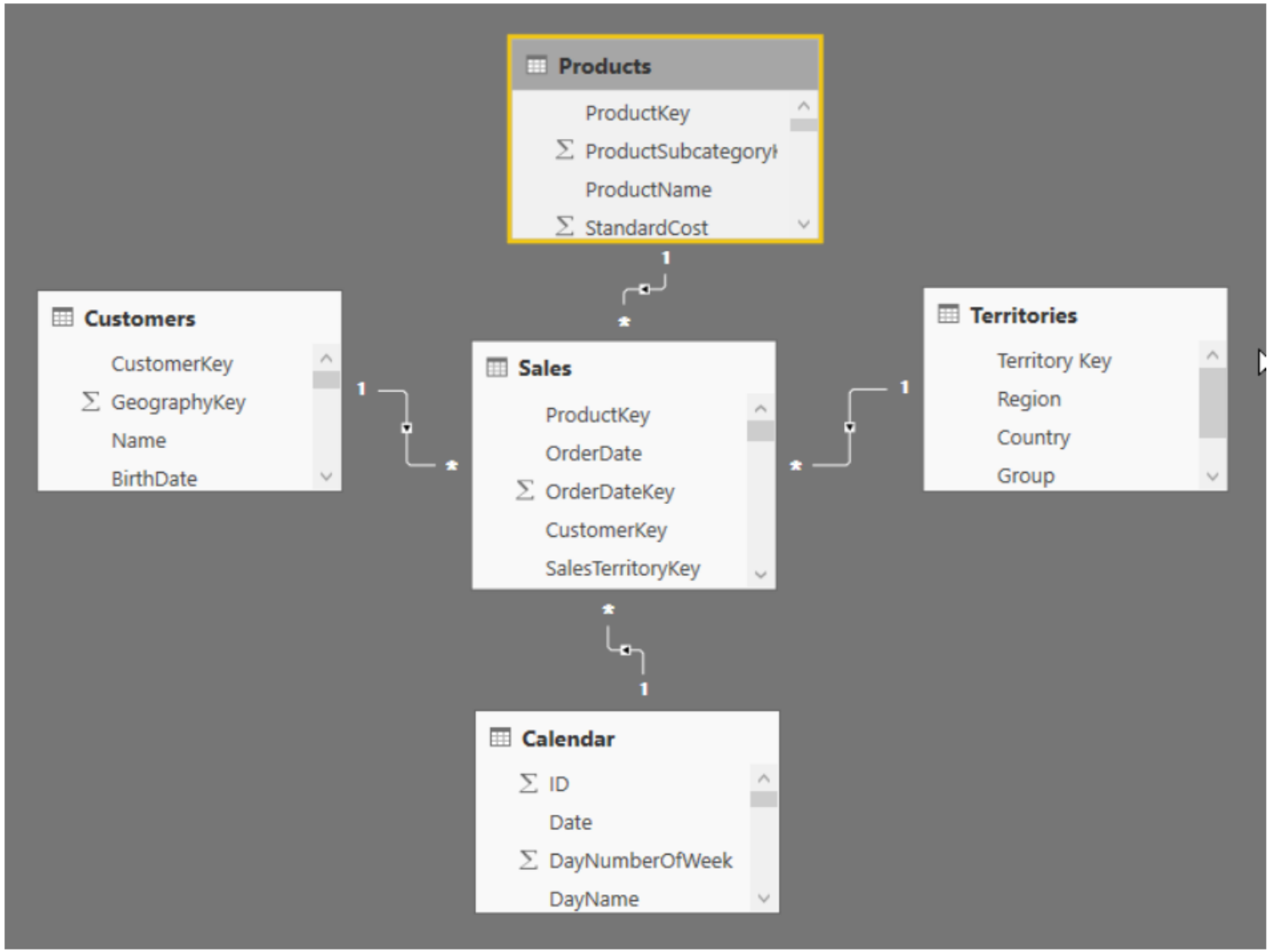


Snowflake Schema



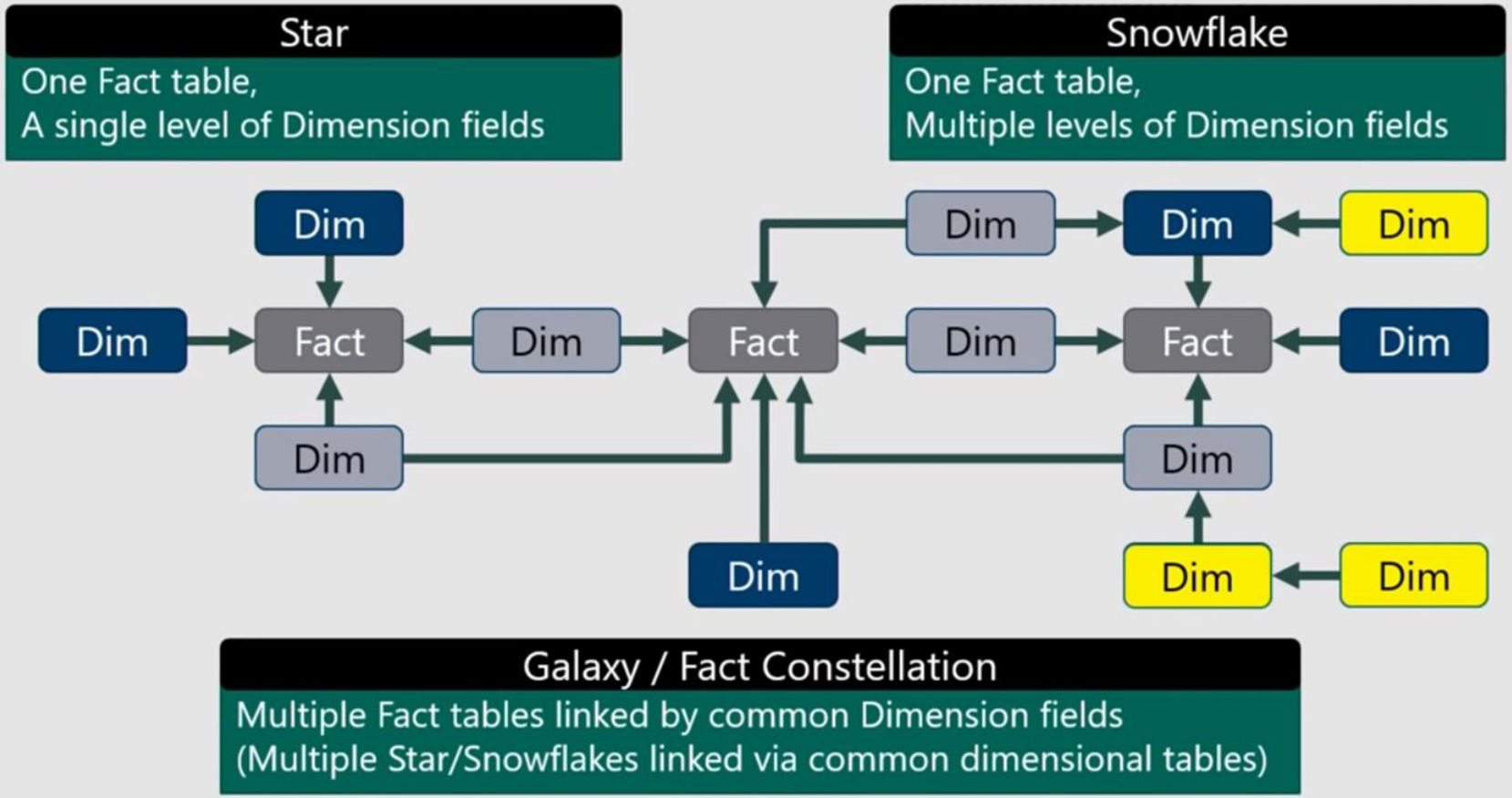
DATA MODEL
Schema

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.



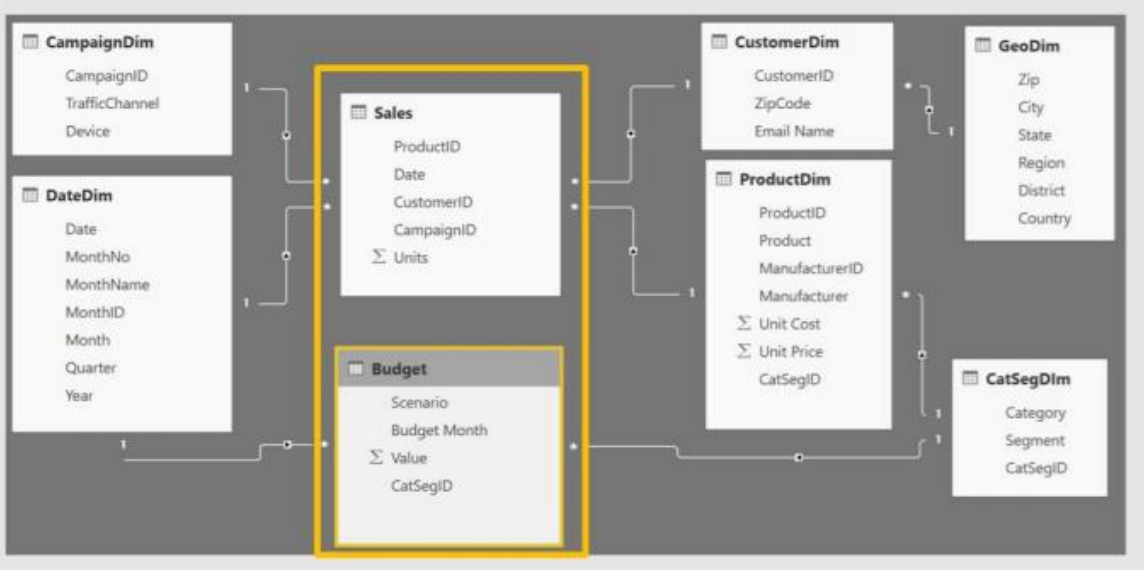
MASTERING DATA ANALYTICS

Relationship Schemas



MASTERING DATA ANALYTICS

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.

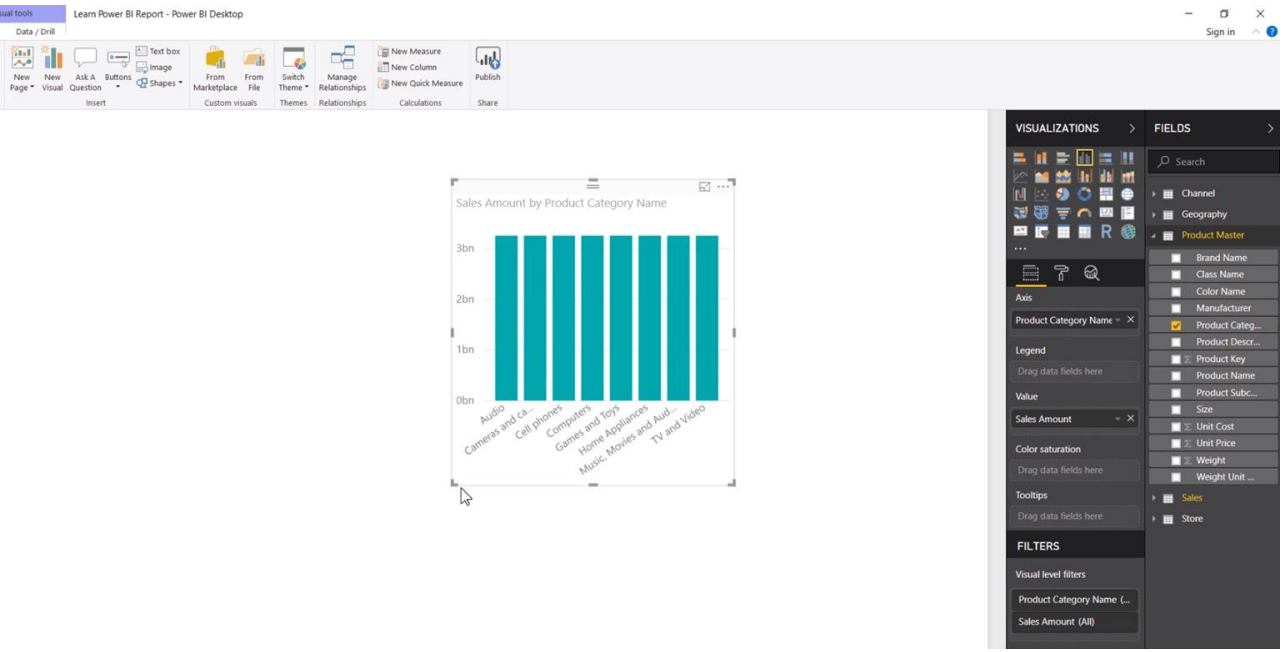
MASTERING DATA ANALYTICS

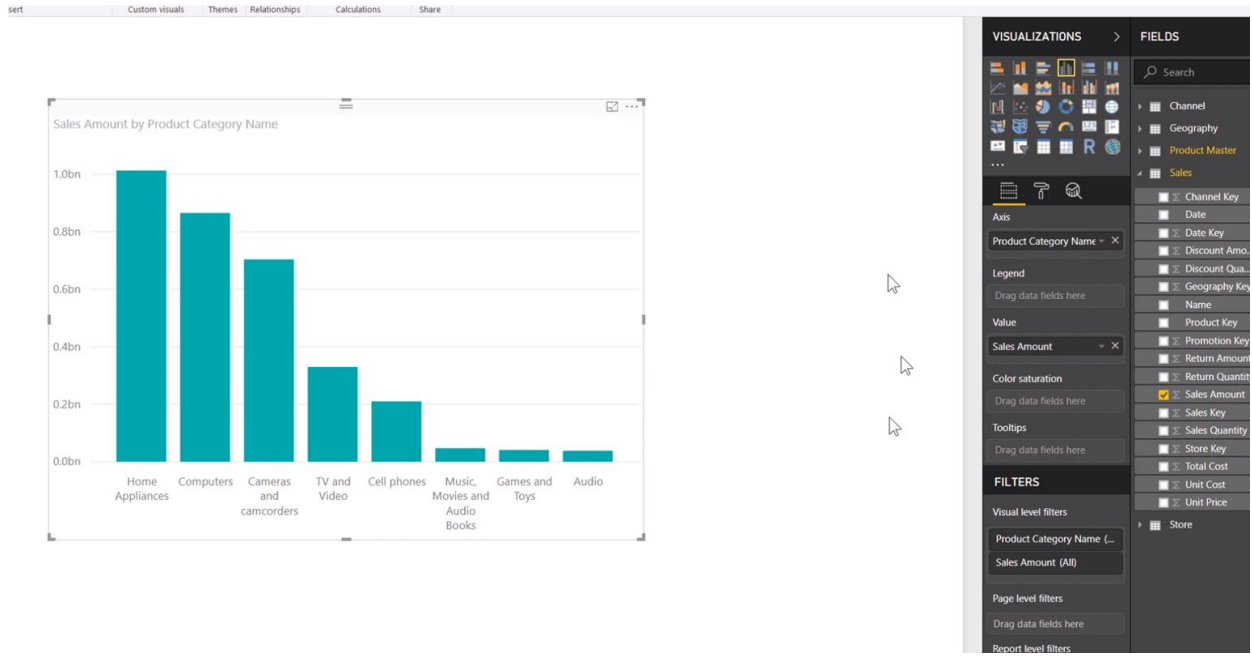
- Fact tables and dimension tables
- Schema
- **Relationships**
- Cardinality
- Cross filter direction
- Hierarchies

DATA MODEL
Relationship

MASTERING DATA ANALYTICS

No Relationships

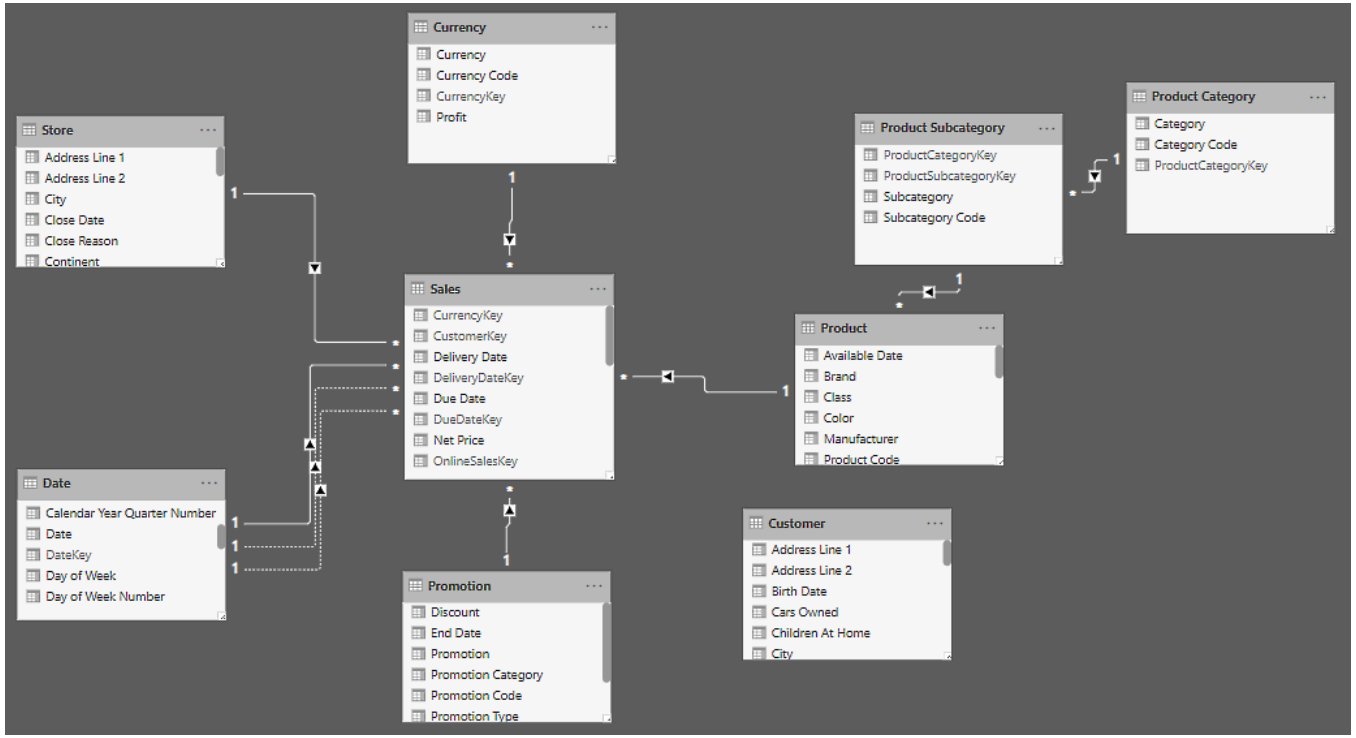




DATA MODEL
Relationship

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.

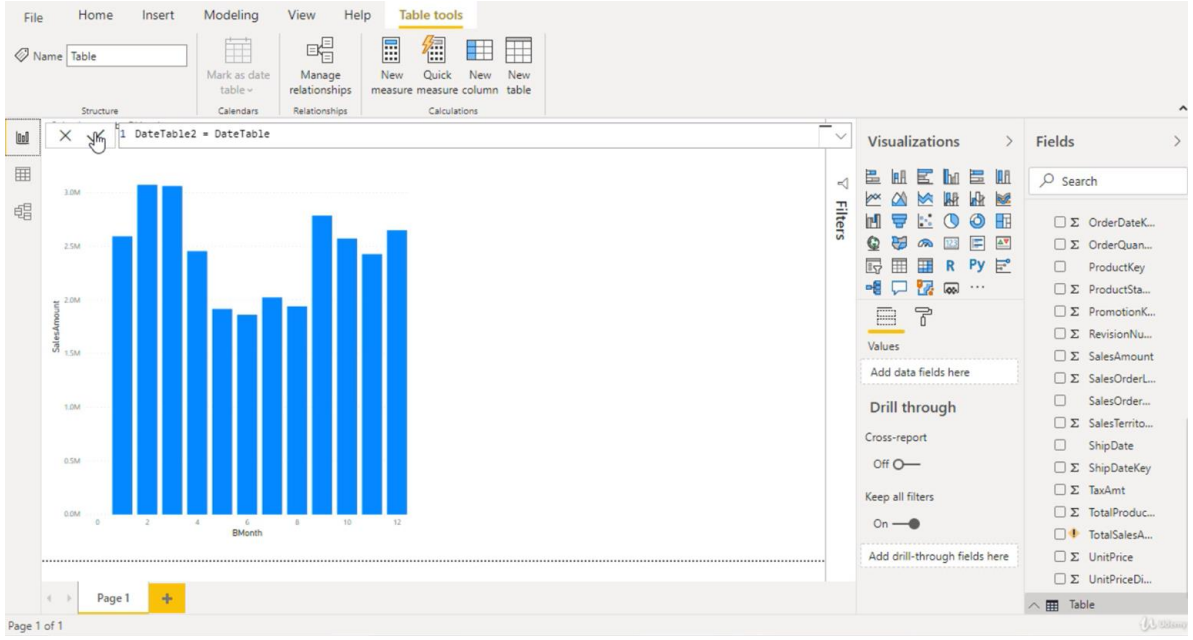
A relationship is defined by a **single column** from each table. You **cannot use multiple columns to define a relationship**, but you can create new columns in each table that consist of multiple columns concatenated together. You can then use these to create the relationships.

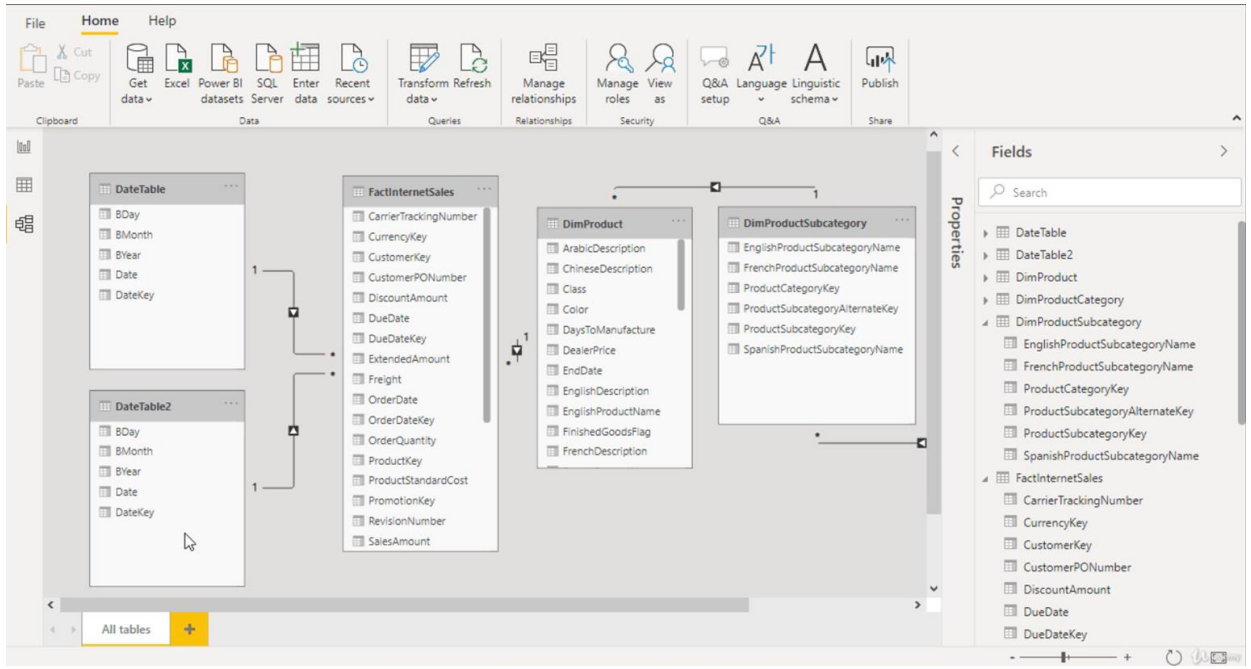


DATA MODEL
Cardinality

MASTERING DATA ANALYTICS

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





MASTERING DATA ANALYTICS

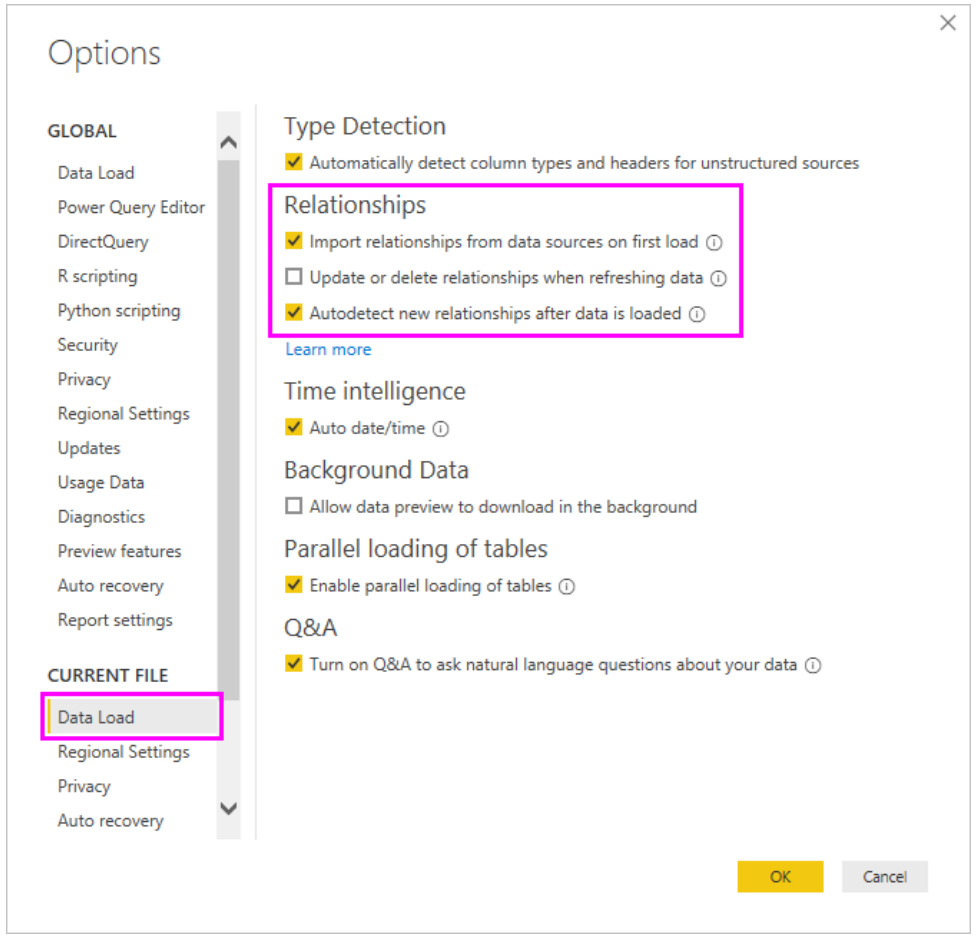
Creating Relationships

- 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

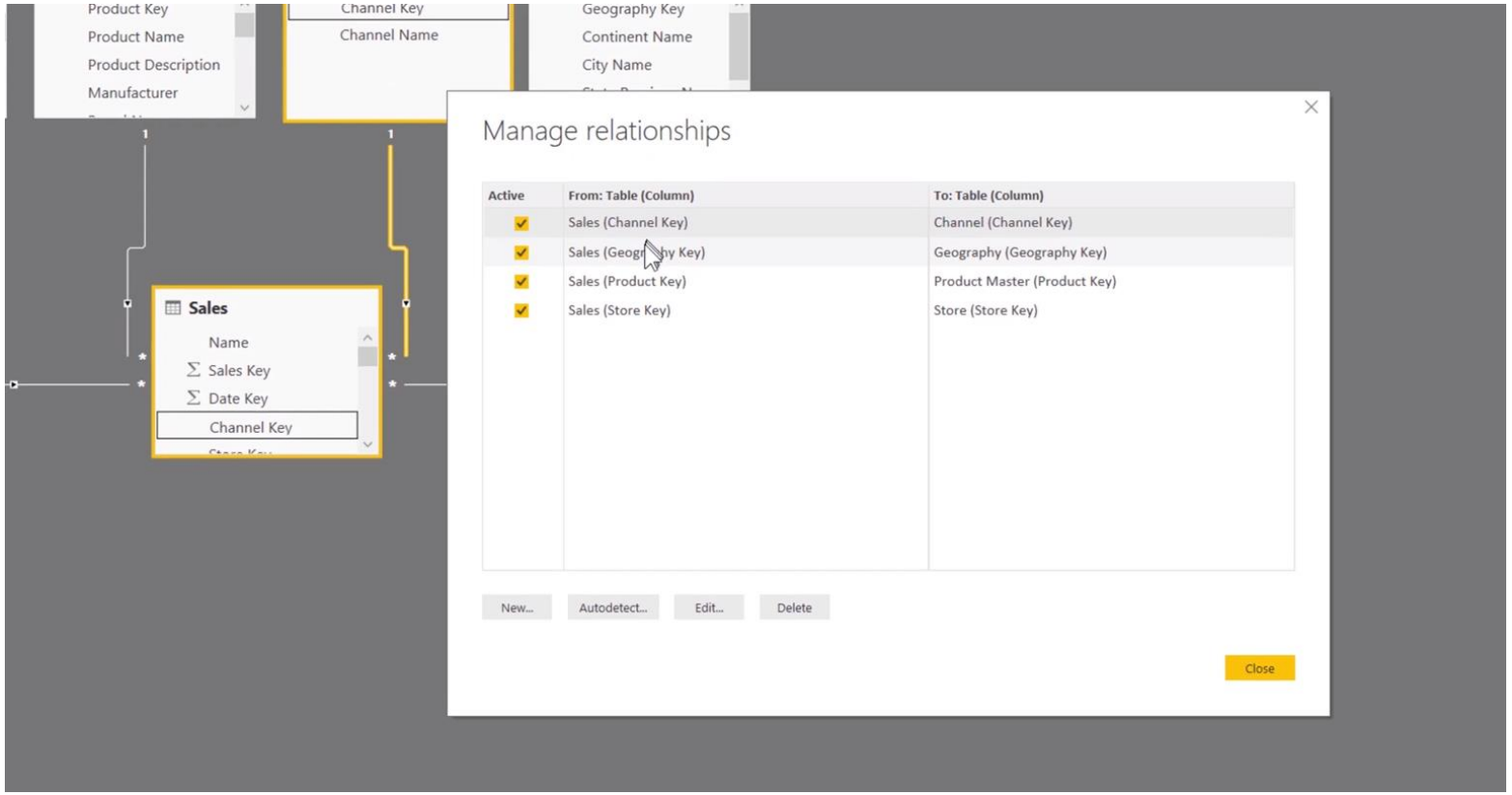
DATA MODEL
Relationship

MASTERING DATA ANALYTICS

Autodetect new relationship



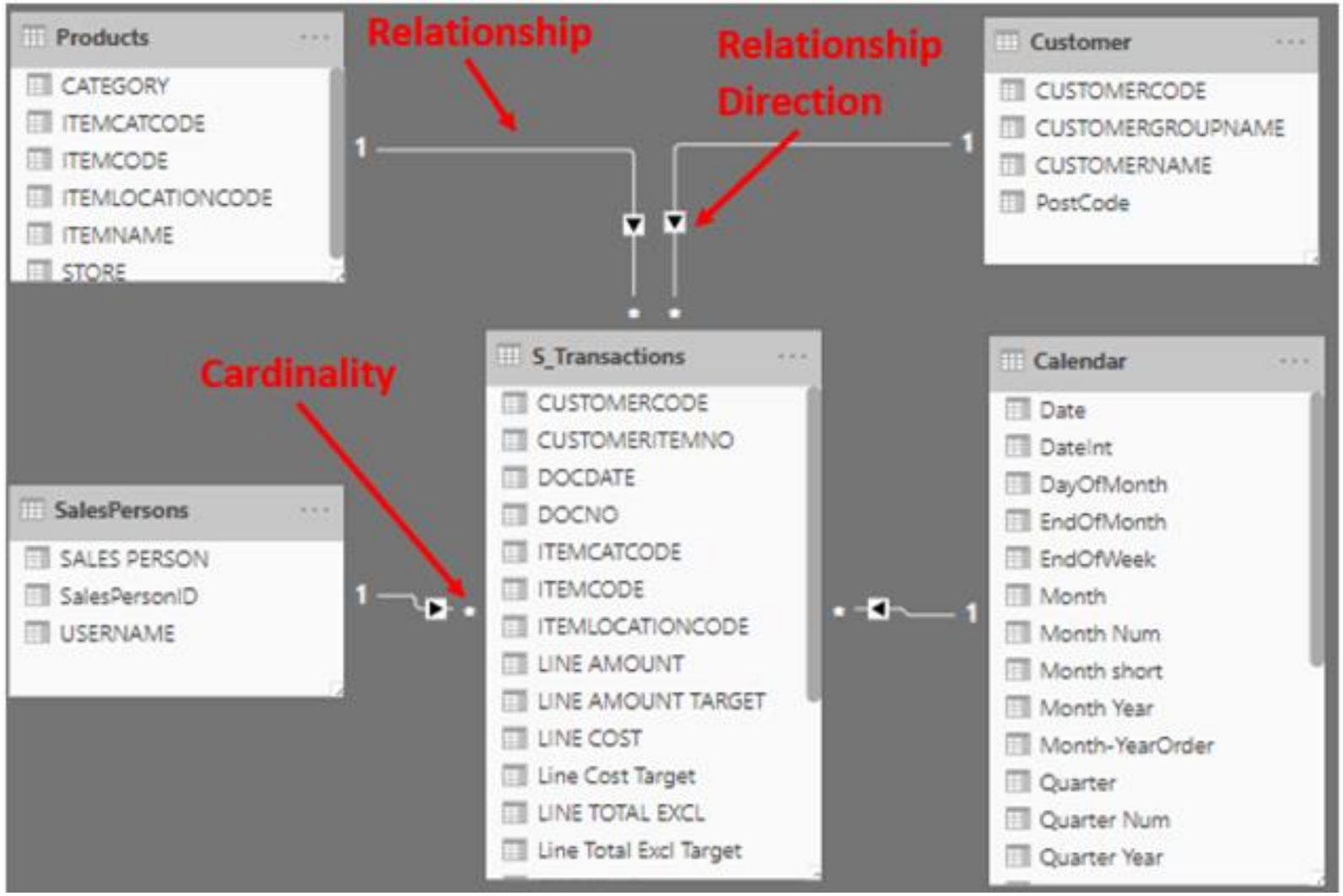
Edit & Manage Relationship (PBI)



DATA MODEL

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

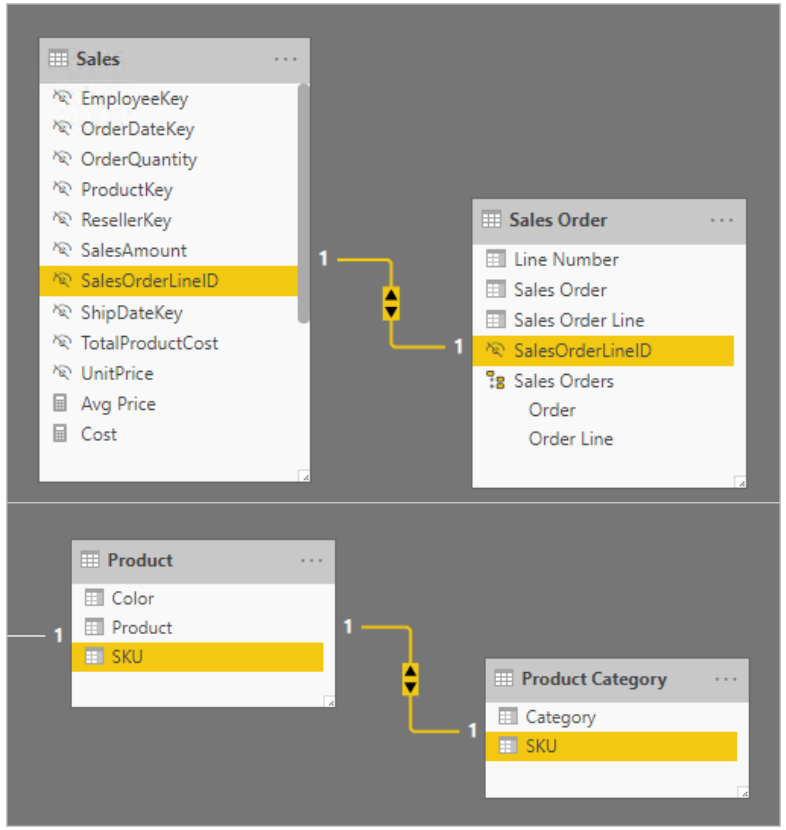
DATA MODEL
Relationship



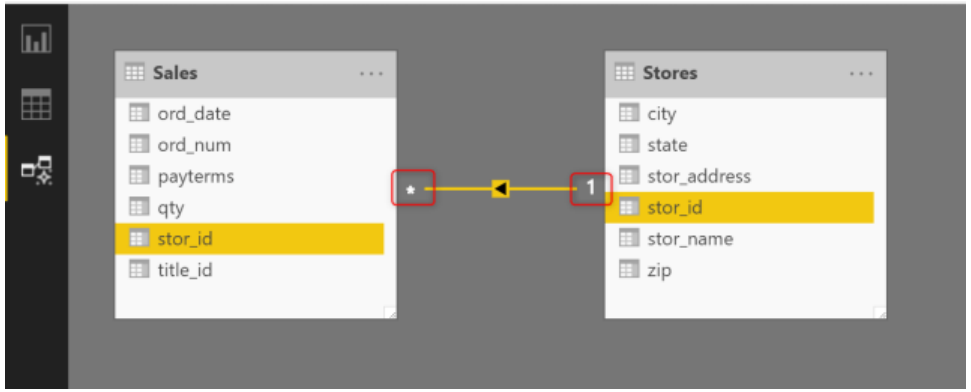
DATA MODEL
Cardinality

MASTERING DATA ANALYTICS

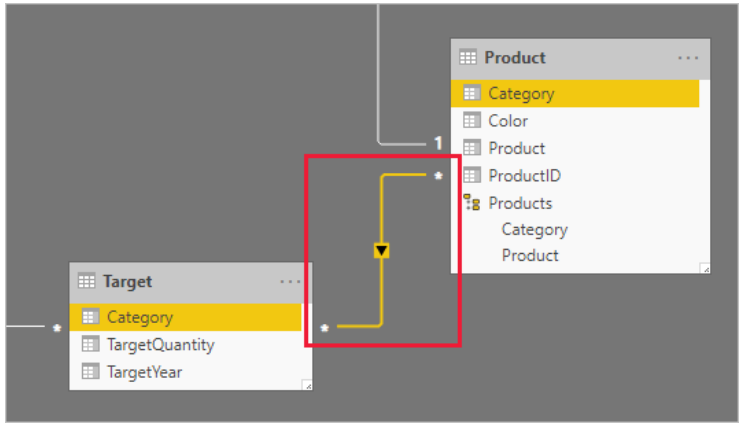
One-to-One



One-to-Many

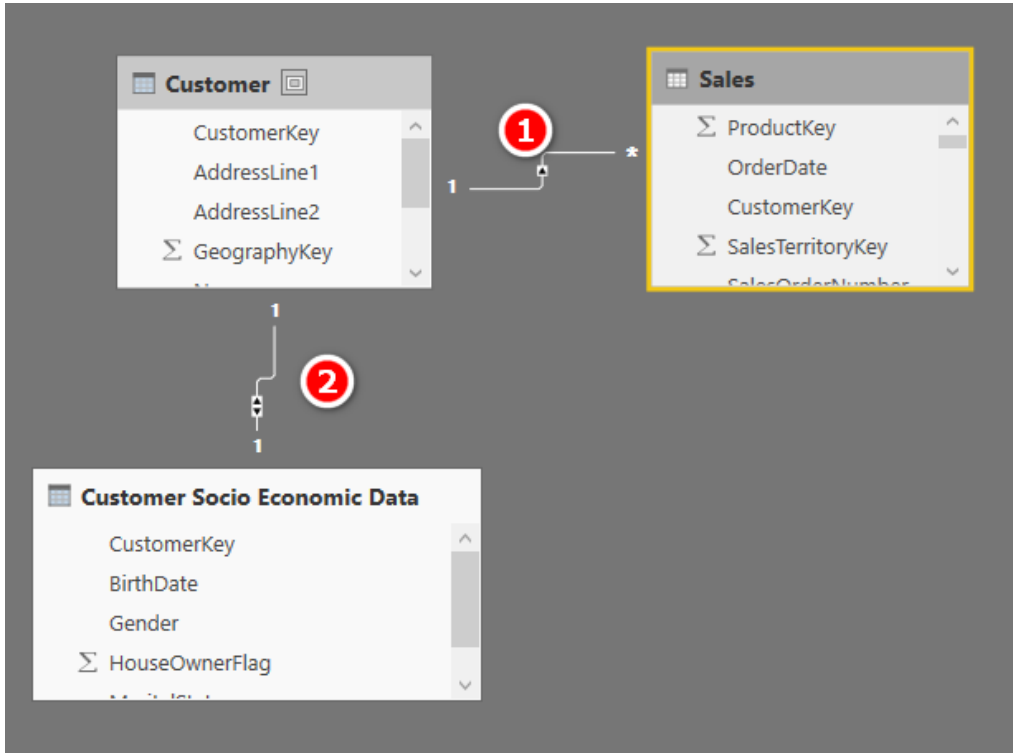


Many-to-Many



DATA MODEL
Cardinality

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.

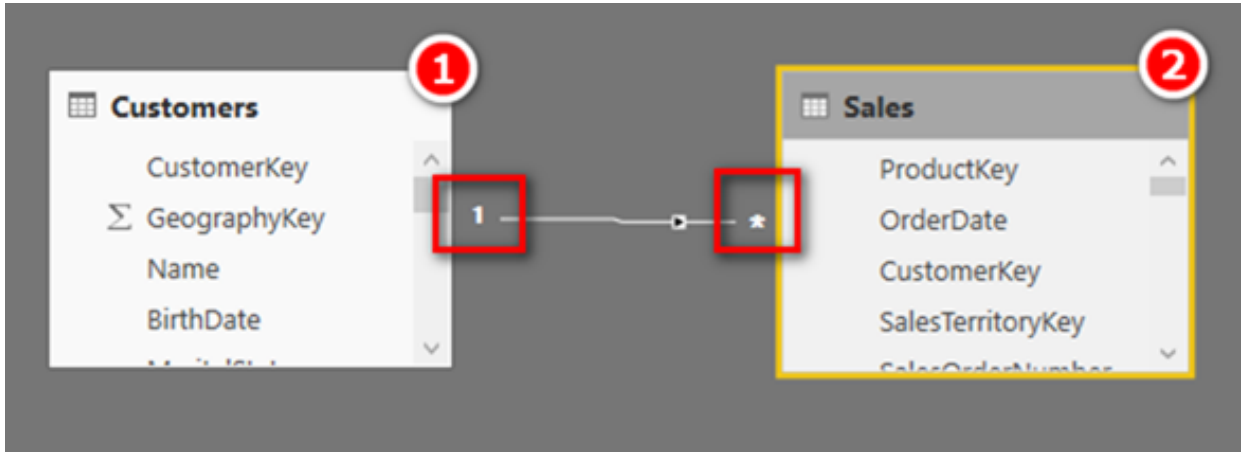
Every relationship has a "Cost". The performance impact may not be noticeable for simple models but may become an issue with very complex models.

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).

MASTERING DATA ANALYTICS

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).

Data Model

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

MASTERING DATA ANALYTICS

CROSS FILTER DIRECTION

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

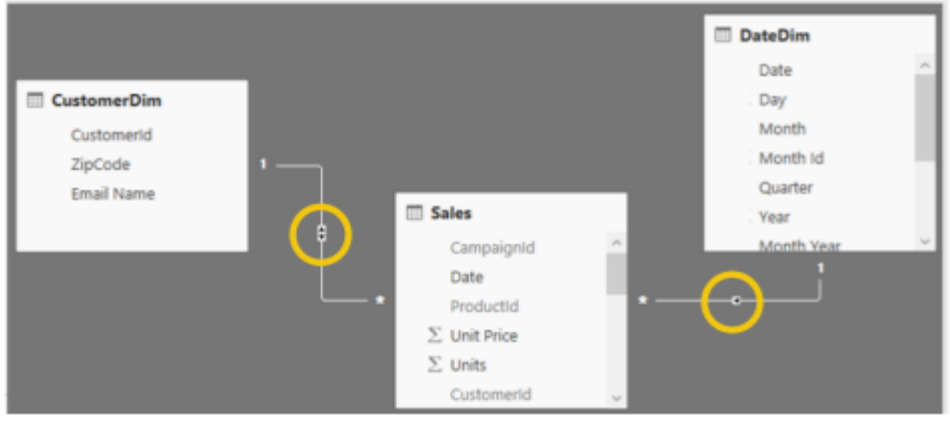
- **Single direction**: If you filter records in the table on the **one side of a relationship**, the filtering choices are carried through to the table on the many side of the relationship. However, if you filter records in the table on the many side of the relationship, these are not carried through 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**.

MASTERING DATA ANALYTICS

Filter Context and Multiple Tables

Filter Context and Multiple Tables



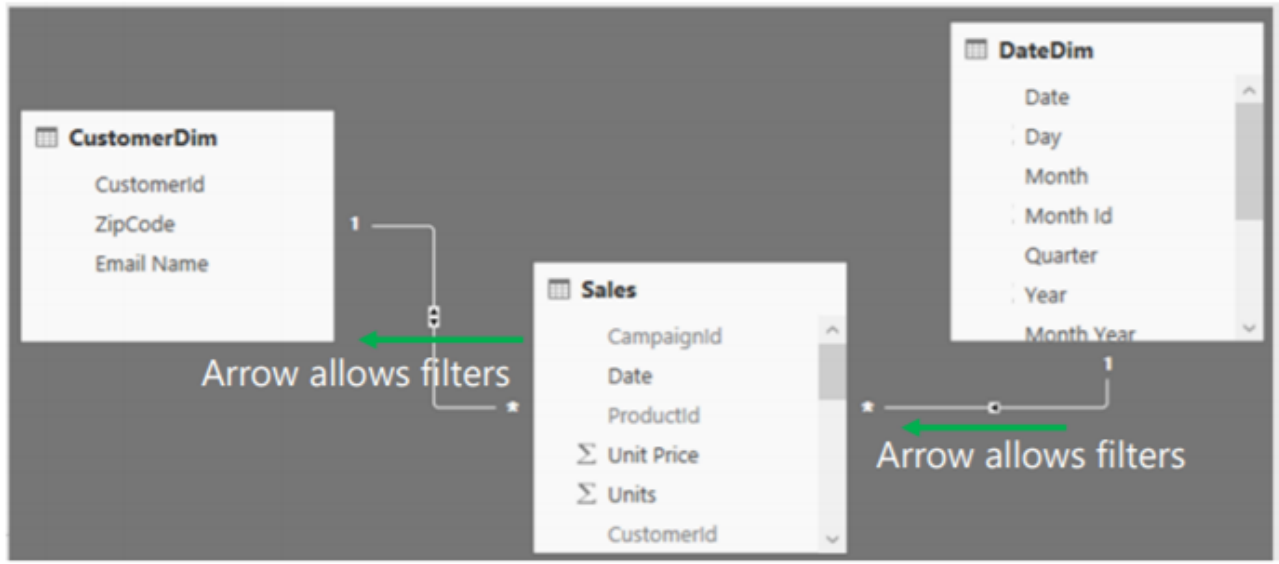
- Filters (Filter context) automatically propagate based on direction of arrows in relationships

- Examples
  - Filter goes from DateDim to CustomerDim
  - Filter does not go from CustomerDim to DateDim

MASTERING DATA ANALYTICS

Filter Context and Multiple Tables

Filter Context and Multiple Tables – Right Arrow Direction



Cross filtering works properly

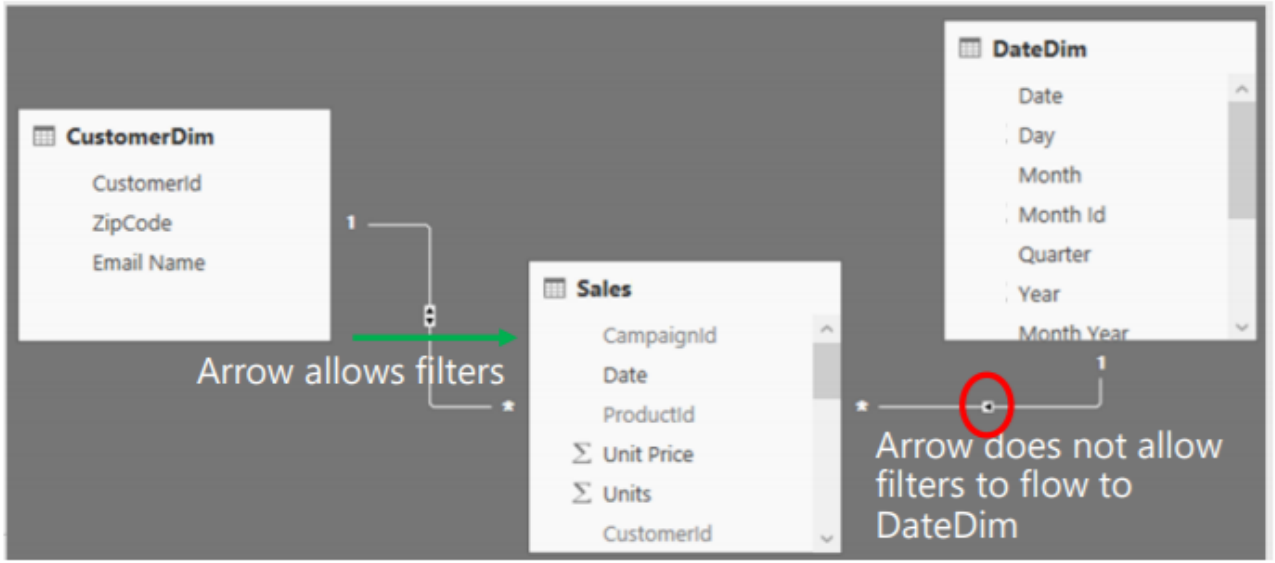
Month	Total Sales M	Count of CustomerId
Jan	$1,673,394.03	7132
Feb	$431,531.13	2820
Mar	$690,671.10	4017
Apr	$852,018.76	4629
May	$972,018.47	5185
Jun	$907,703.04	4854
Jul	$608,678.35	3680
Aug	$1,355,530.22	6242
Sep	$720,851.83	4186
Oct	$1,117,087.73	5728
Nov	$2,372,763.71	8242
Dec	$2,003,261.11	7683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

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

MASTERING DATA ANALYTICS

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**

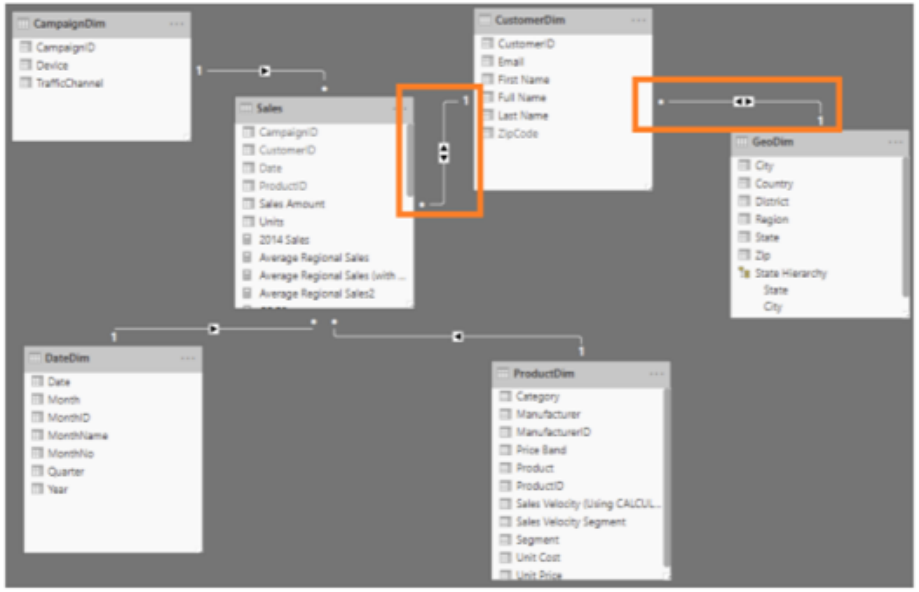
- Filter goes from DateDim to CustomerDim
- This is why the Count of Month in the table above is incorrect

MASTERING DATA ANALYTICS

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

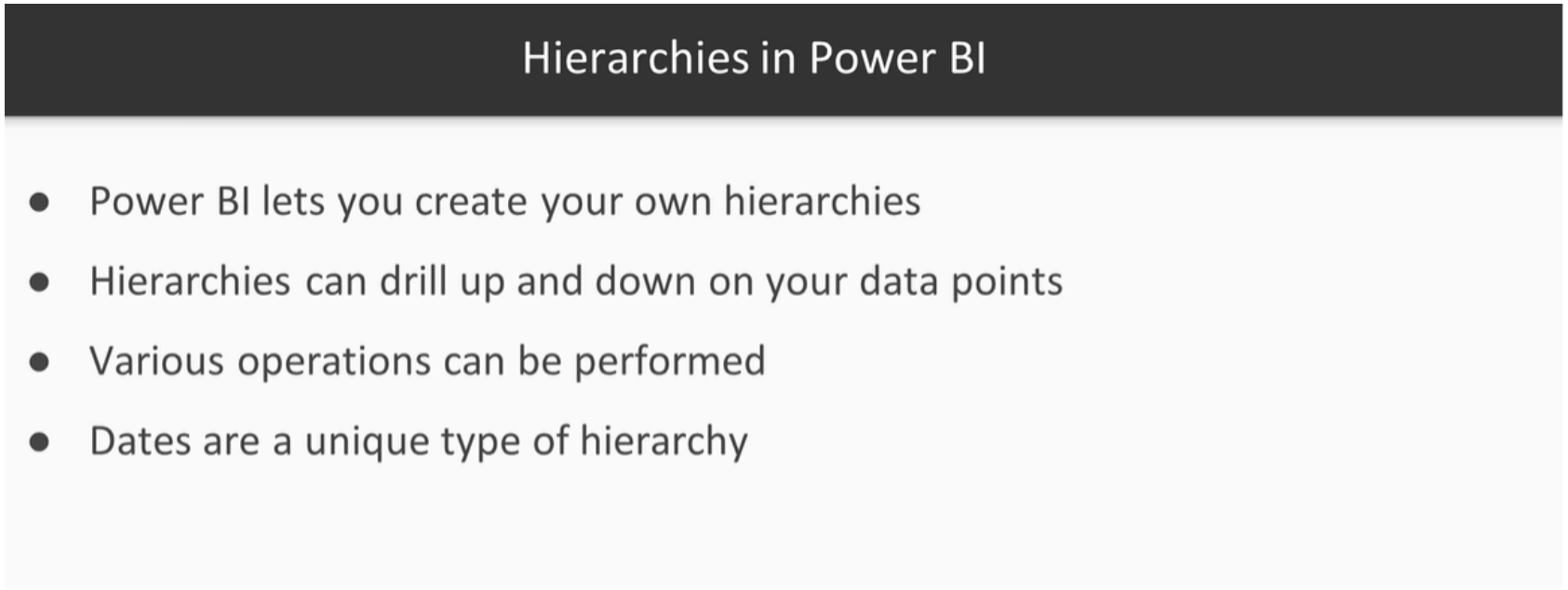
DATA MODEL

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

MASTERING DATA ANALYTICS

Hierarchies

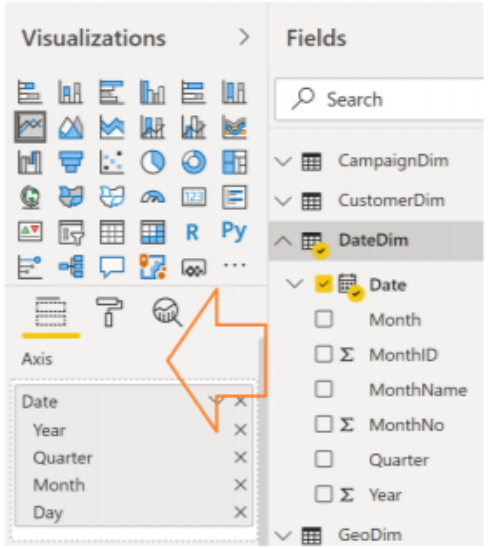
A hierarchy is a set of **nested columns** that are **grouped together** in a way that allows you to drill up and down a report visual using a single object from the field list. A **typical example** of a hierarchy is usually found in a **date table**, where a date hierarchy might consist of the year, 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.



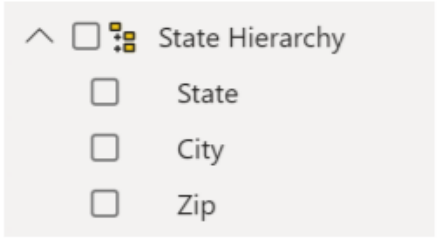
MASTERING DATA ANALYTICS

Hierarchies

- Power BI generates Date hierarchies when dates are added to visuals, this allows the end user to drill from Year, Quarter, Month & Day.



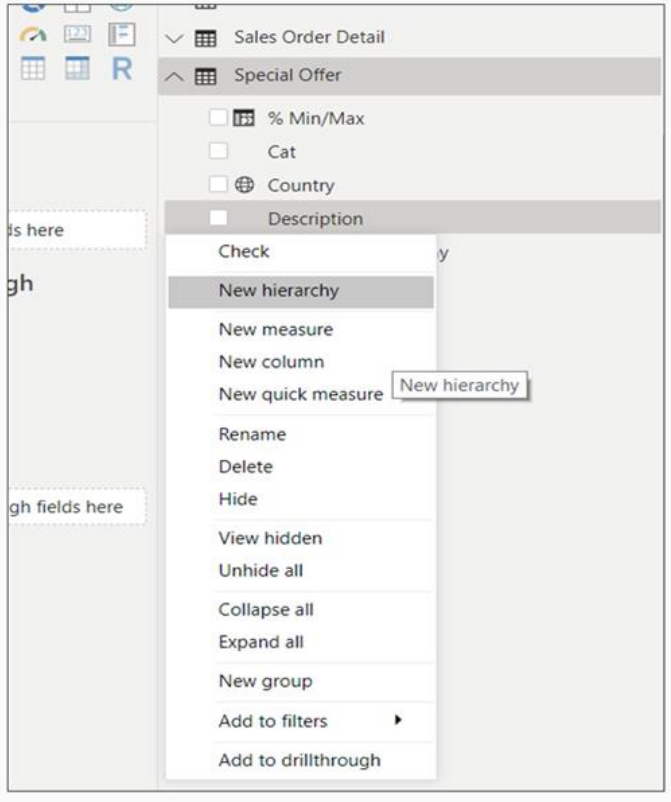
- Users can also create custom hierarchies in the model by dragging a lower level field onto the parent.



MASTERING DATA ANALYTICS

Creating Your Own Hierarchy

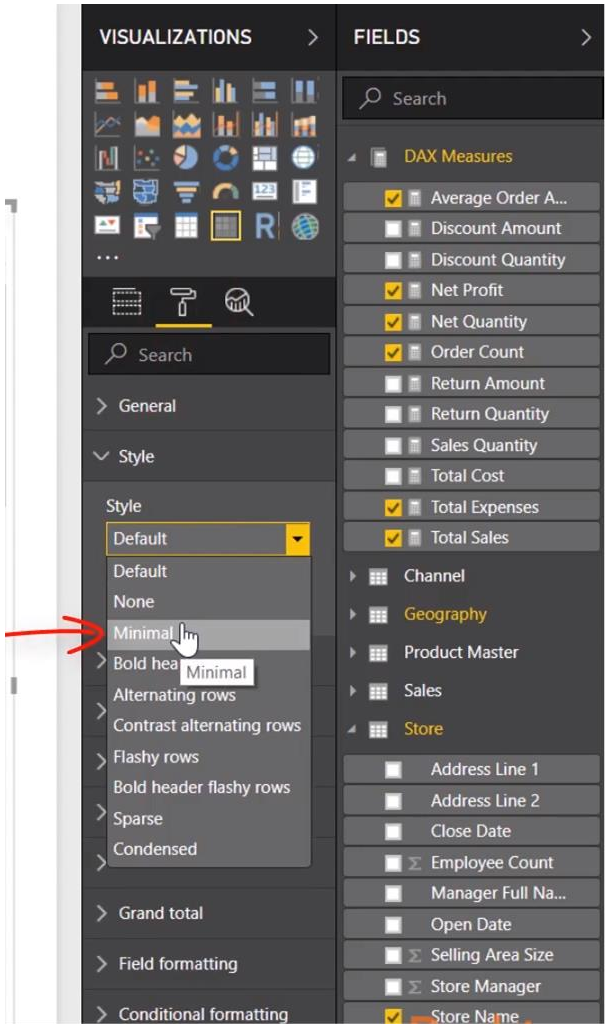
- There are two ways to create a hierarchy:
  - Drag and drop one field on the another
  - Select the field that you want to add and right-click on it; this will open the context menu, where you can either create a new hierarchy or add to the existing hierarchy



DATA MODEL
Hierarchies

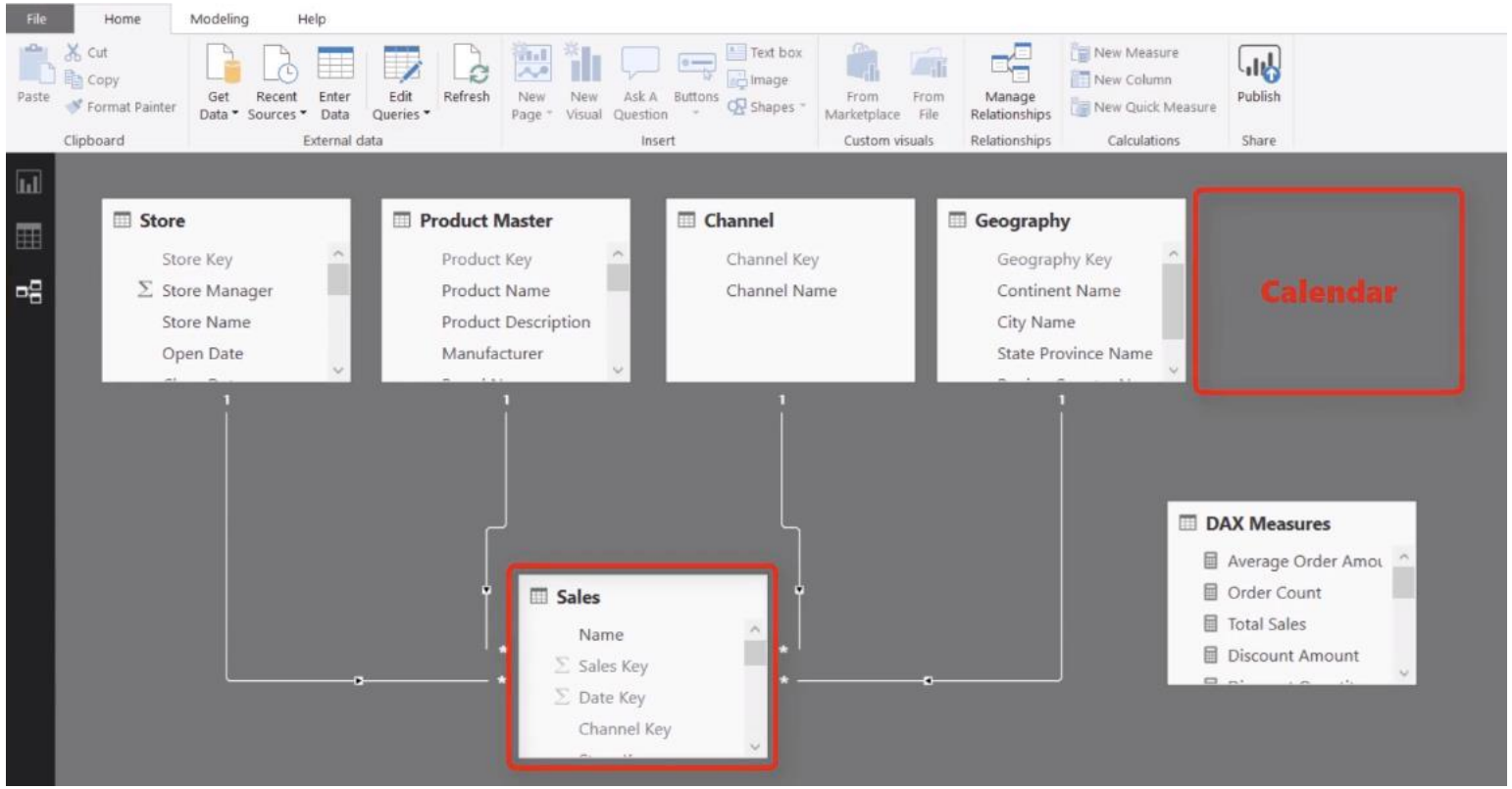
Continent Name	Total Sales	Total Expenses	Net Profit	Net Quantity	Order Count	Average Order Amount
Asia	$703,151,234	$322,629,302	$380,521,932	3,107,004	173,110	$4,062
Armenia	$9,556,092	$4,396,610	$5,159,482	39,548	2,731	$3,499
Armenia	$9,556,092	$4,396,610	$5,159,482	39,548	2,731	$3,499
Australia	$29,467,554	$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					14	$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	$9,735,461	$4,494,604	$5,240,857	39,994	2,787	$3,493
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.



DATA MODEL

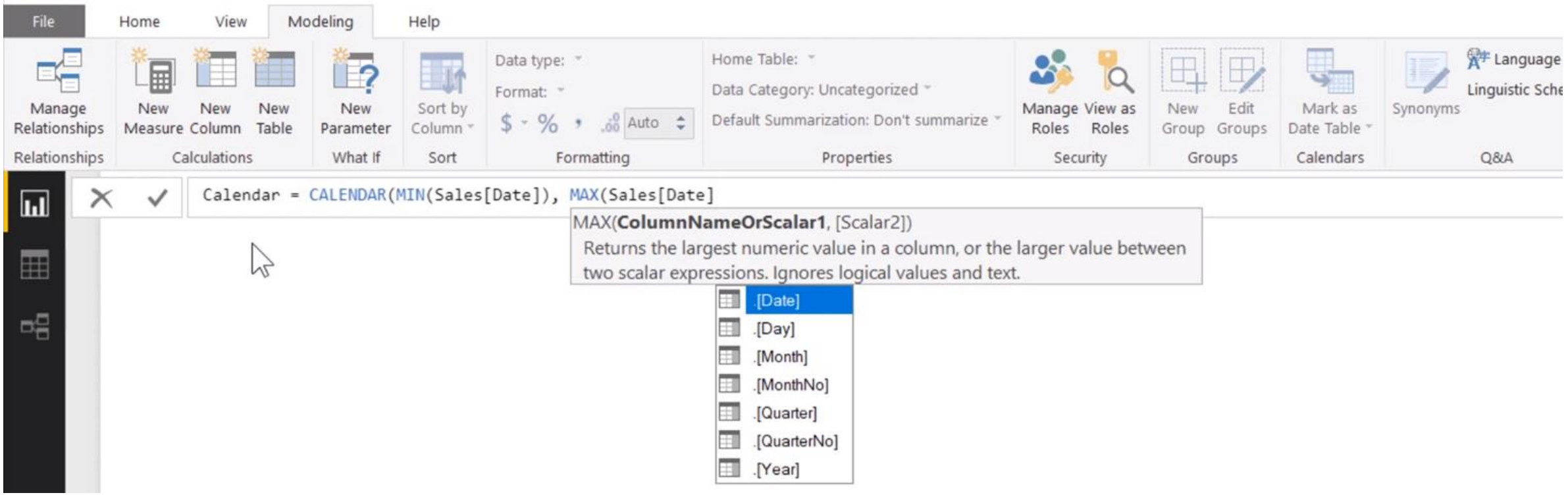
Duplicate vs Reference Tab
New Table Calendar in Model



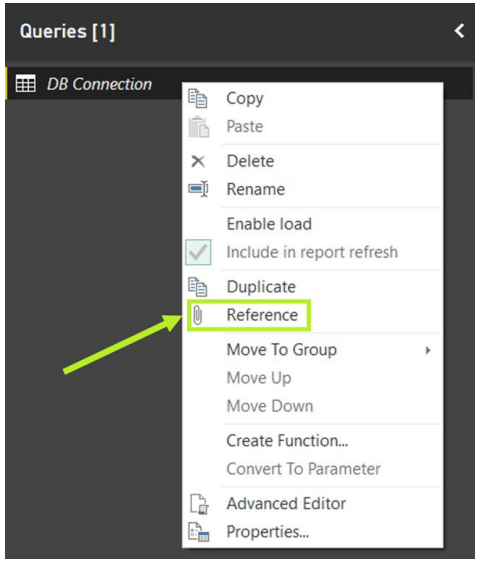
DAX can help us create a Calendar table that we'll anchor to our sales data that will be able to grab the start date for sales and the end date for sales from our Sales table, and then create a Calendar table for exactly that range.

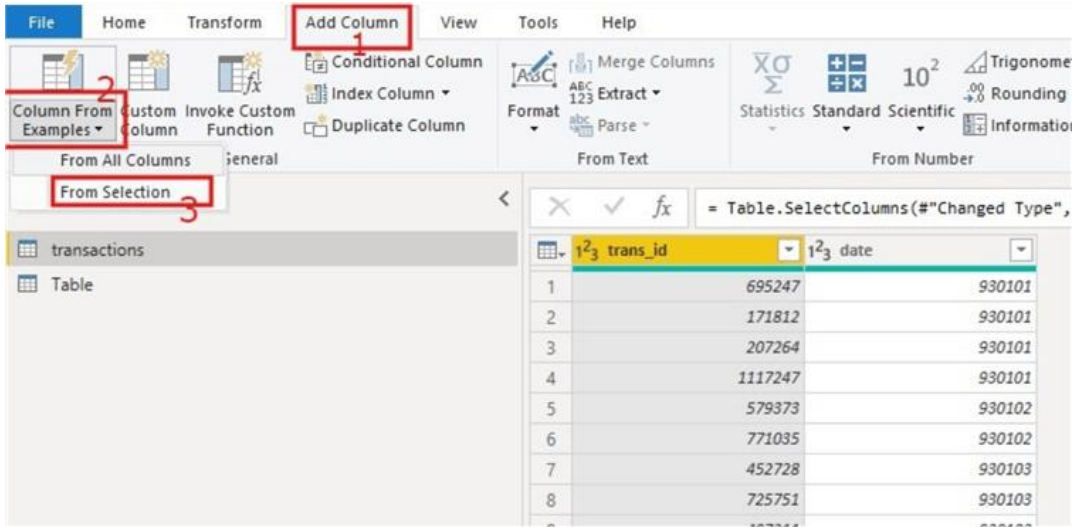
DATA MODEL

Duplicate vs Reference Tab
New Table Calendar in Model



DATA MODEL
Duplicate vs Reference Tab
New Table Calendar in Model

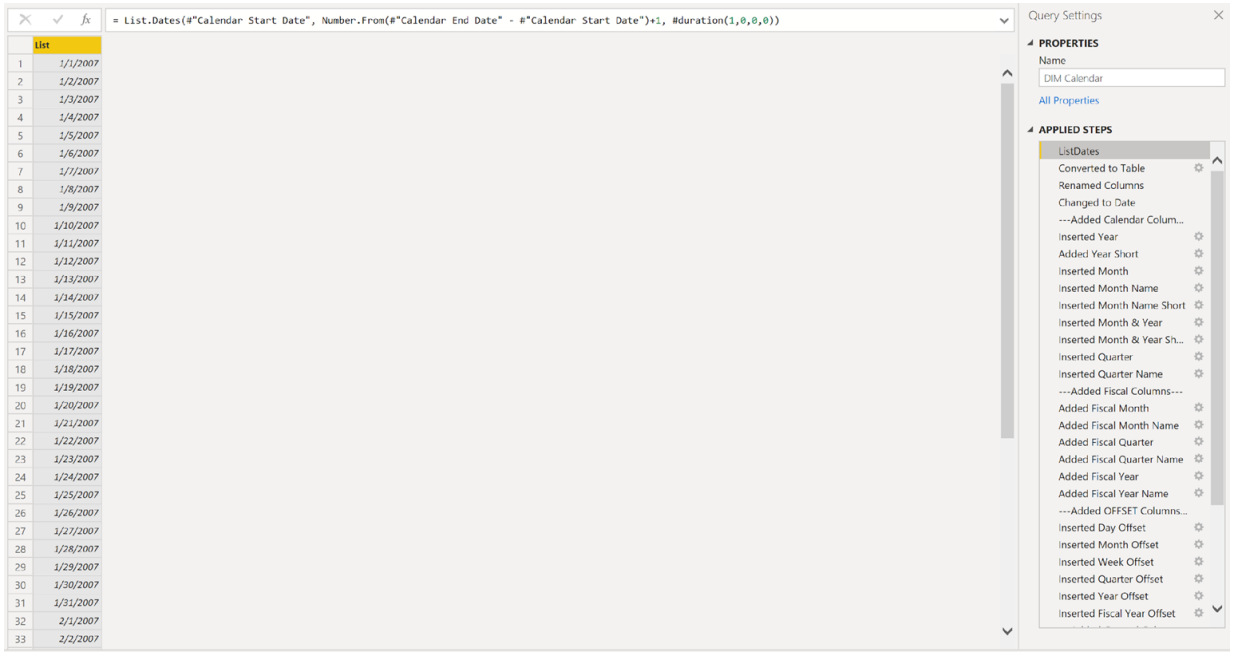


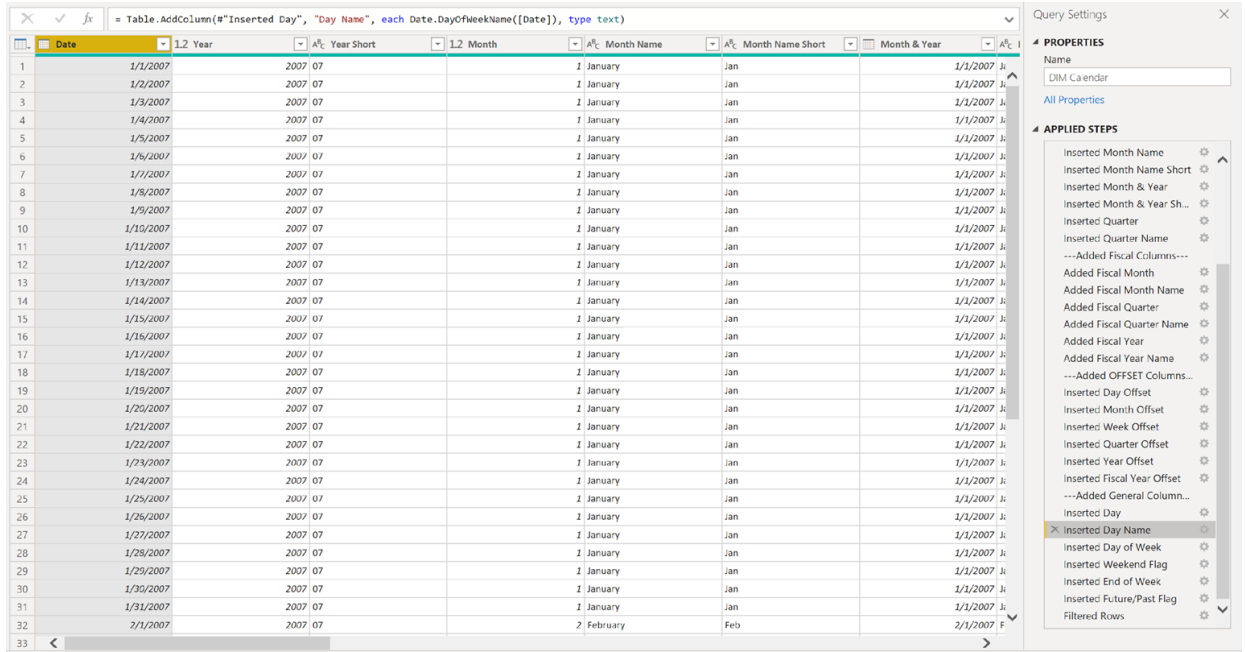


DATA MODEL
Duplicate vs Reference Tab
New Table Calendar in Model

MASTERING DATA ANALYTICS

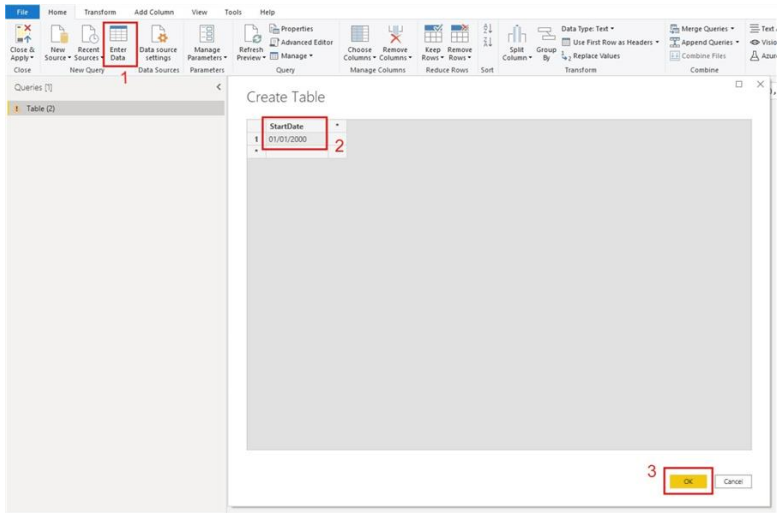
The 3rd way: **M Language**

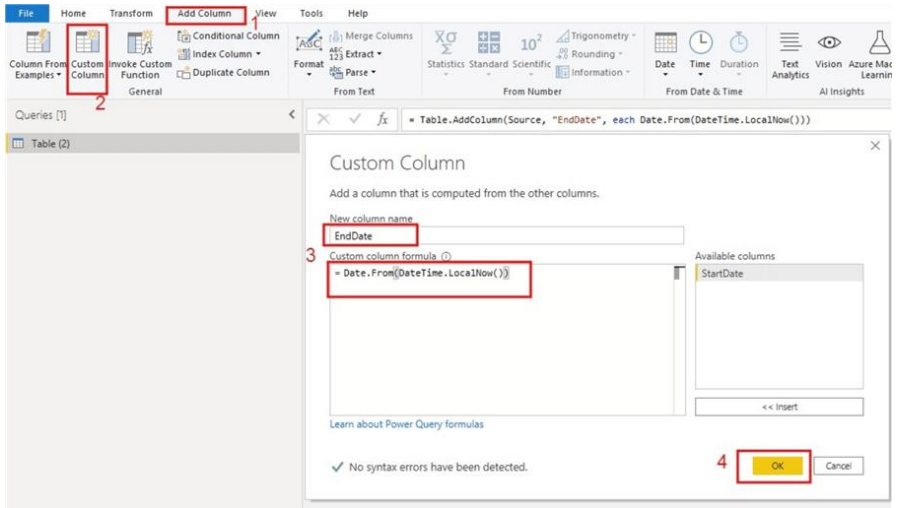


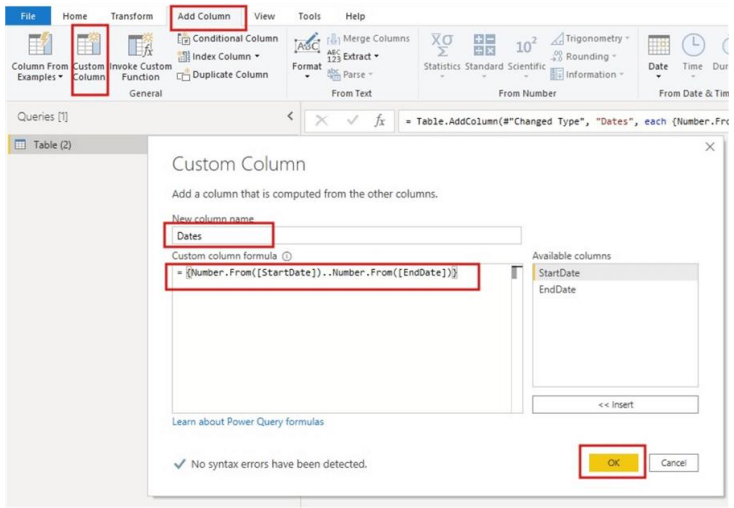


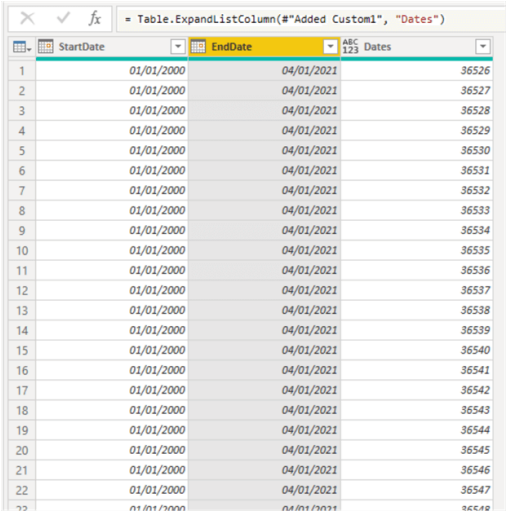
DATA MODEL

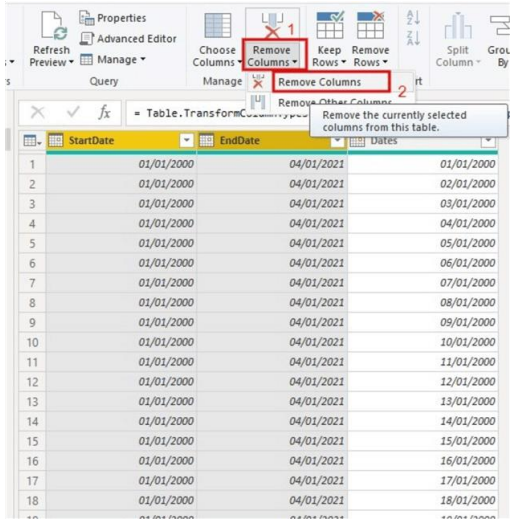
Duplicate vs Reference Tab
New Table Calendar in Model









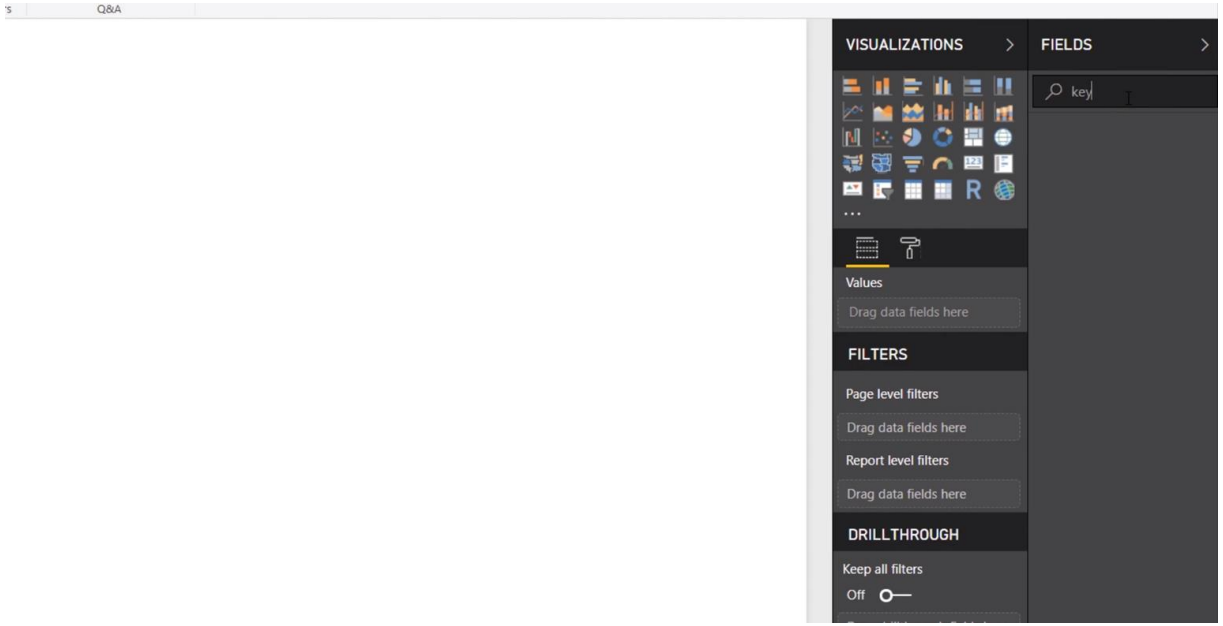


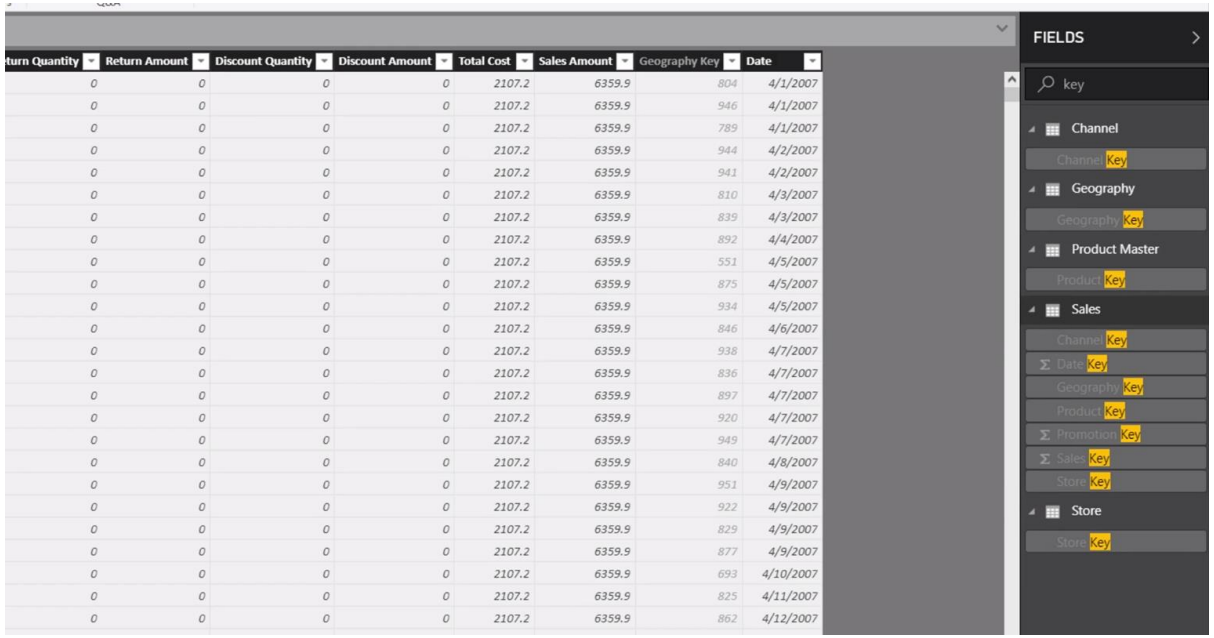
Hide Columns and Tables

- 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.

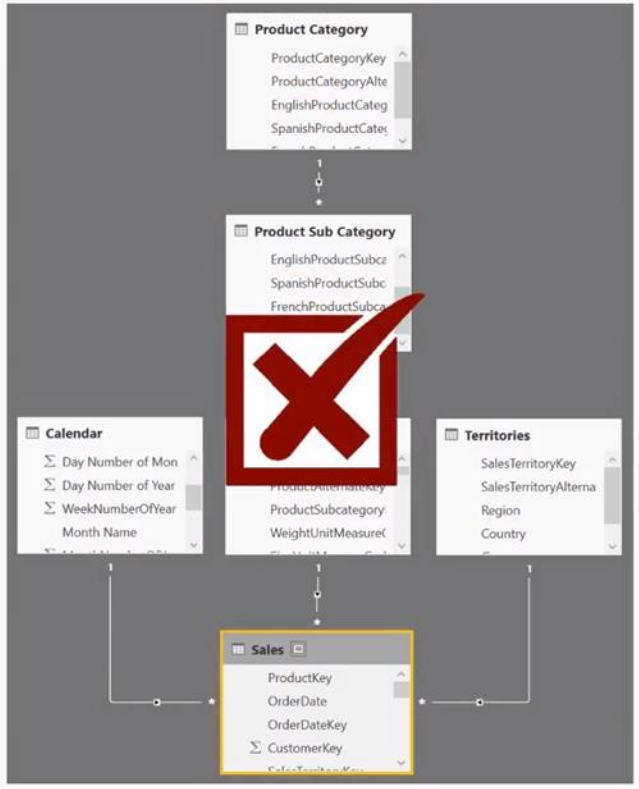




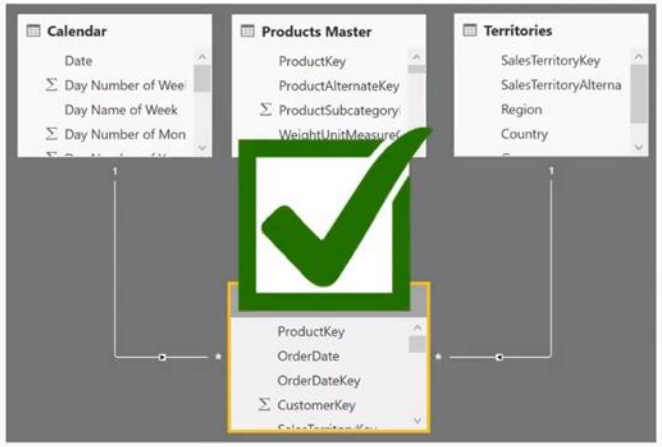
- **Merge Table in Power Query**
- Naming Convention
- VertiPaq Engine
- Optimize Your File tips

MASTERING DATA ANALYTICS

Why do we **Want** to **Merge Tables**?

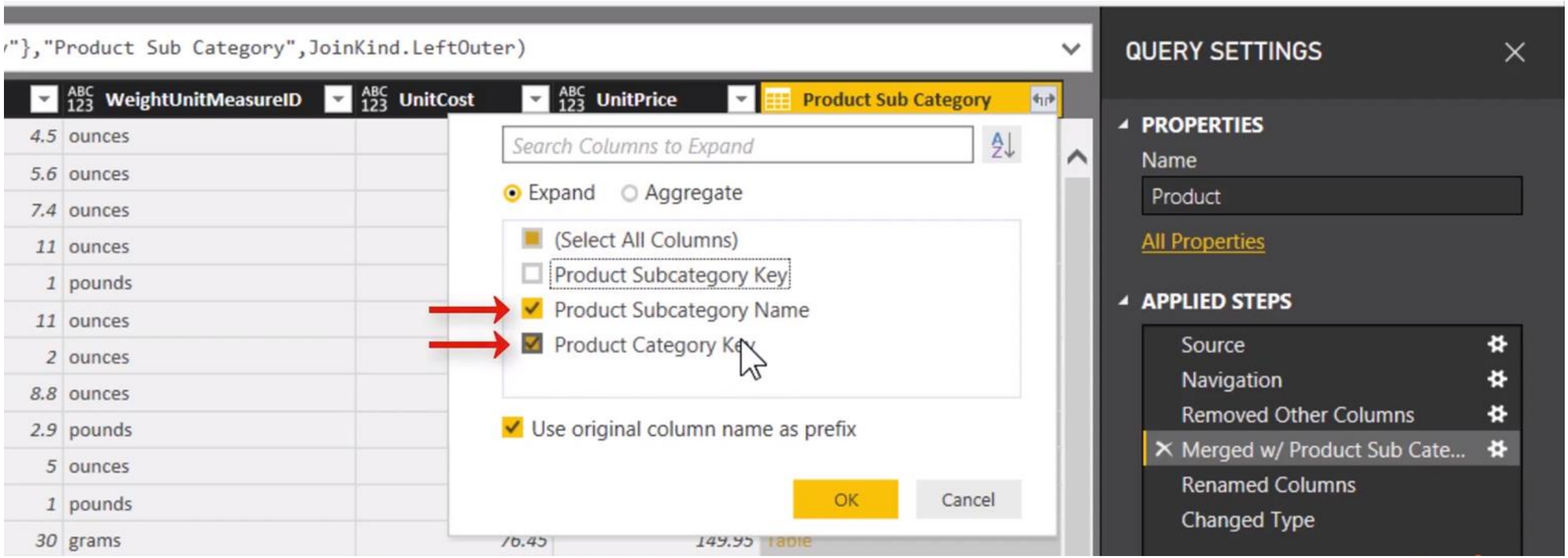


Snowflake Schema

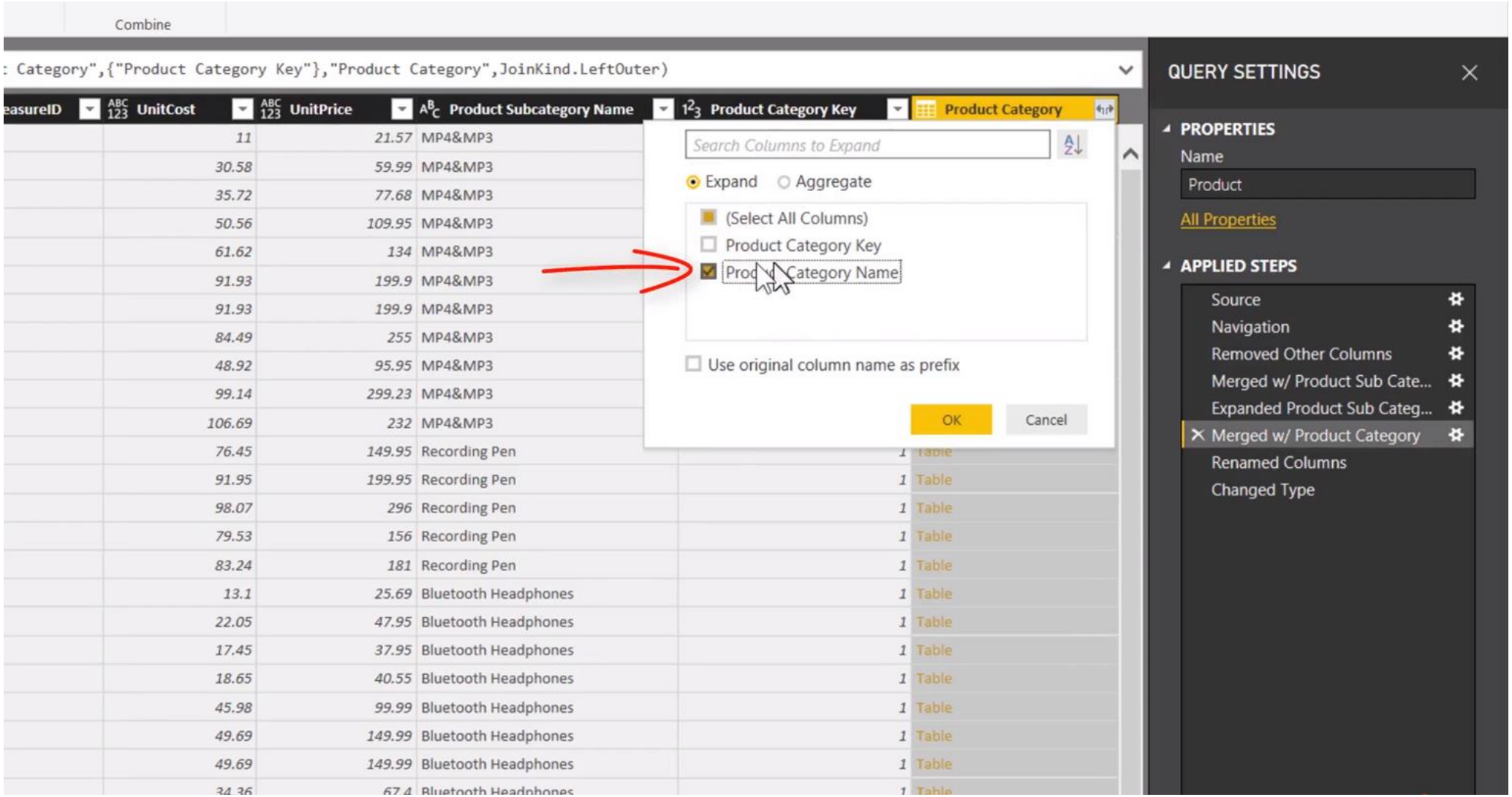


Star Schema

DATA MODEL
Merge



DATA MODEL
Merge



DATA MODEL

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

DATA MODEL
Naming Convention

MASTERING DATA ANALYTICS

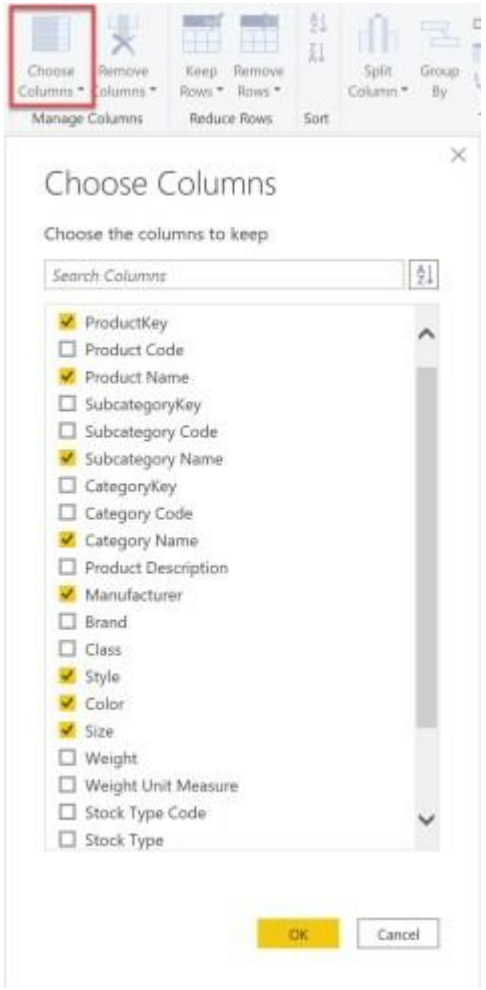
Rename Columns

- 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



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

MASTERING DATA ANALYTICS

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!!

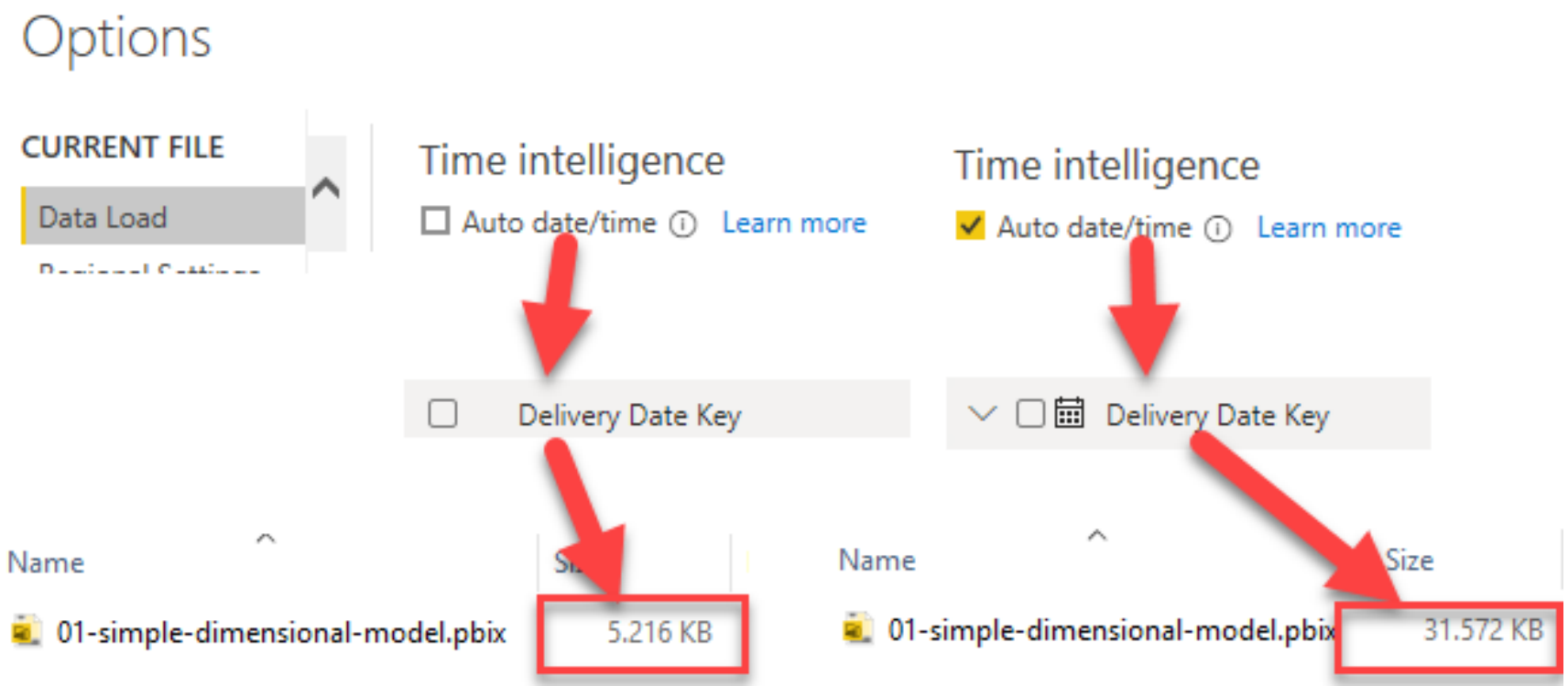
Use integer surrogate keys, pre-sort them

- 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

DATA MODEL

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

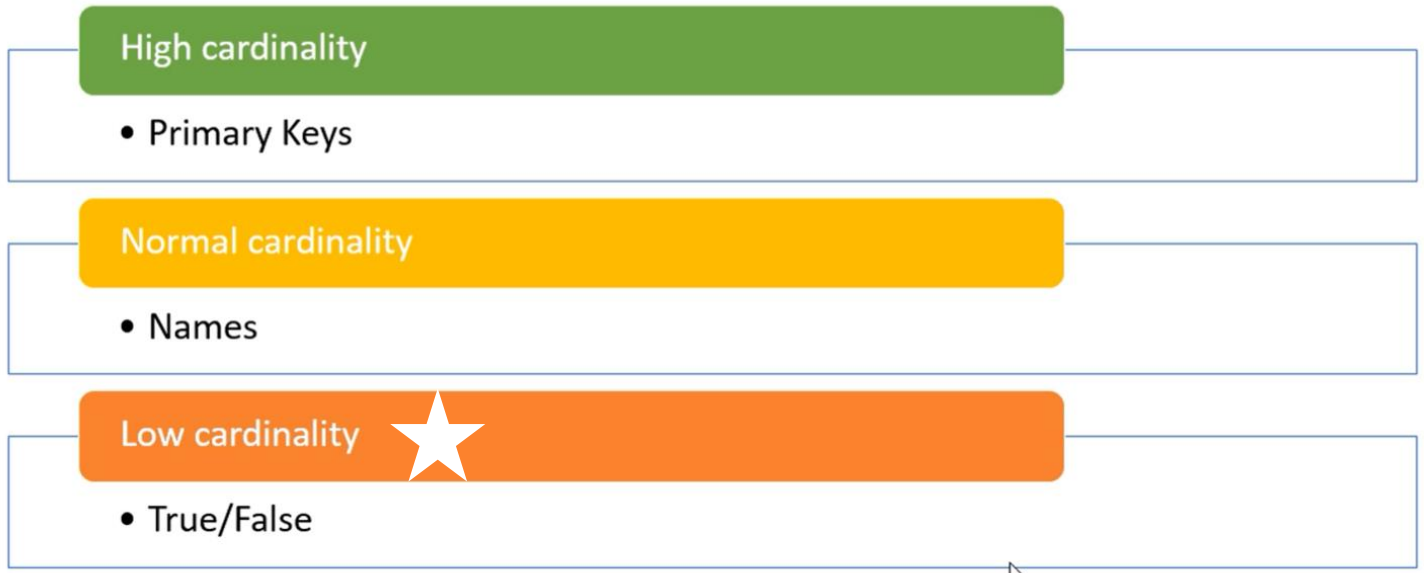
DATA MODEL
Optimization - Time Intelligence Active



DATA MODEL
Optimization - Lower Cardinality whenever possible

MASTERING DATA ANALYTICS

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.

MASTERING DATA ANALYTICS

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

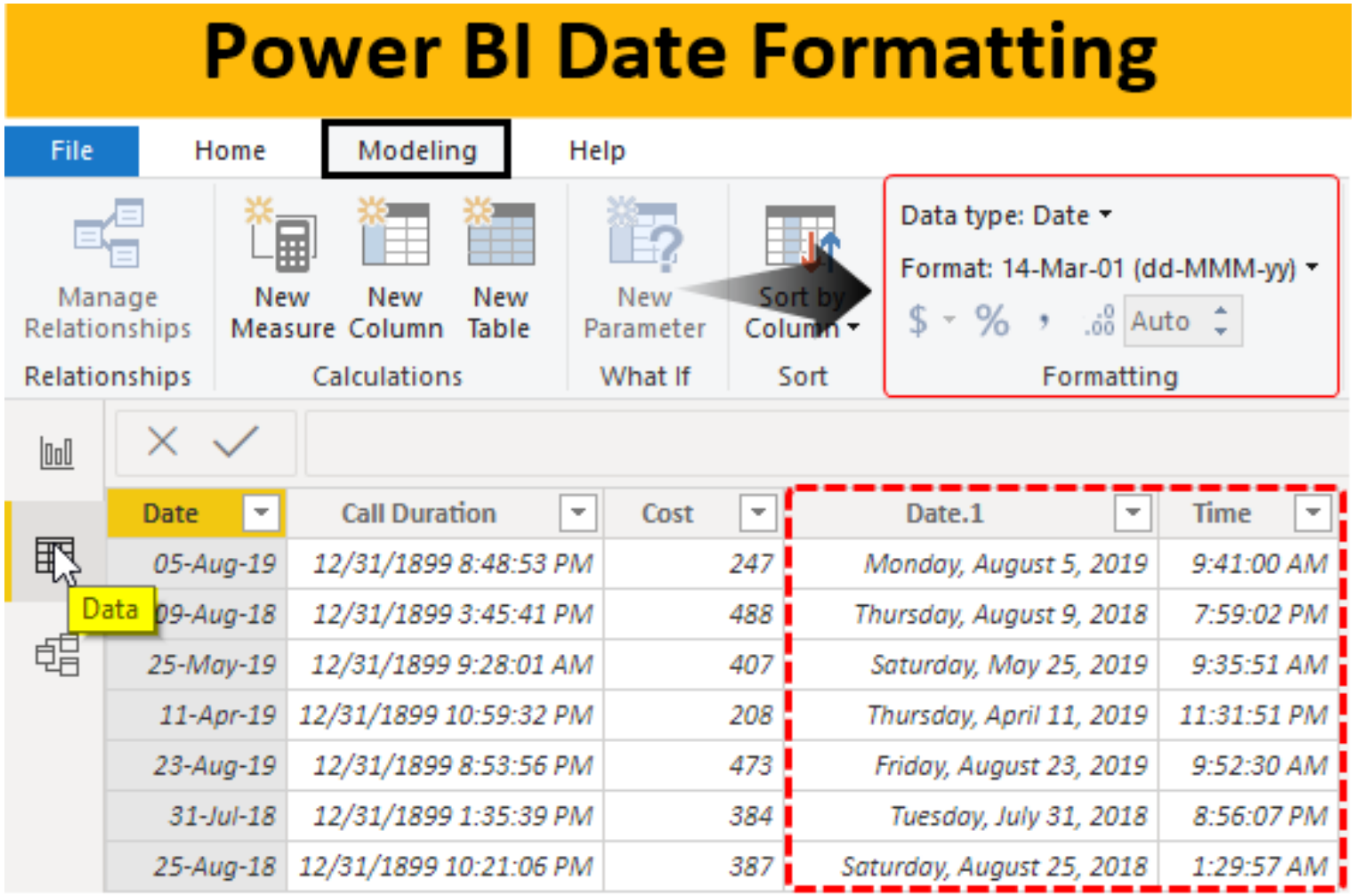
Why is it undesired?

- 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



Use integers instead of strings

Why is it undesired?

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

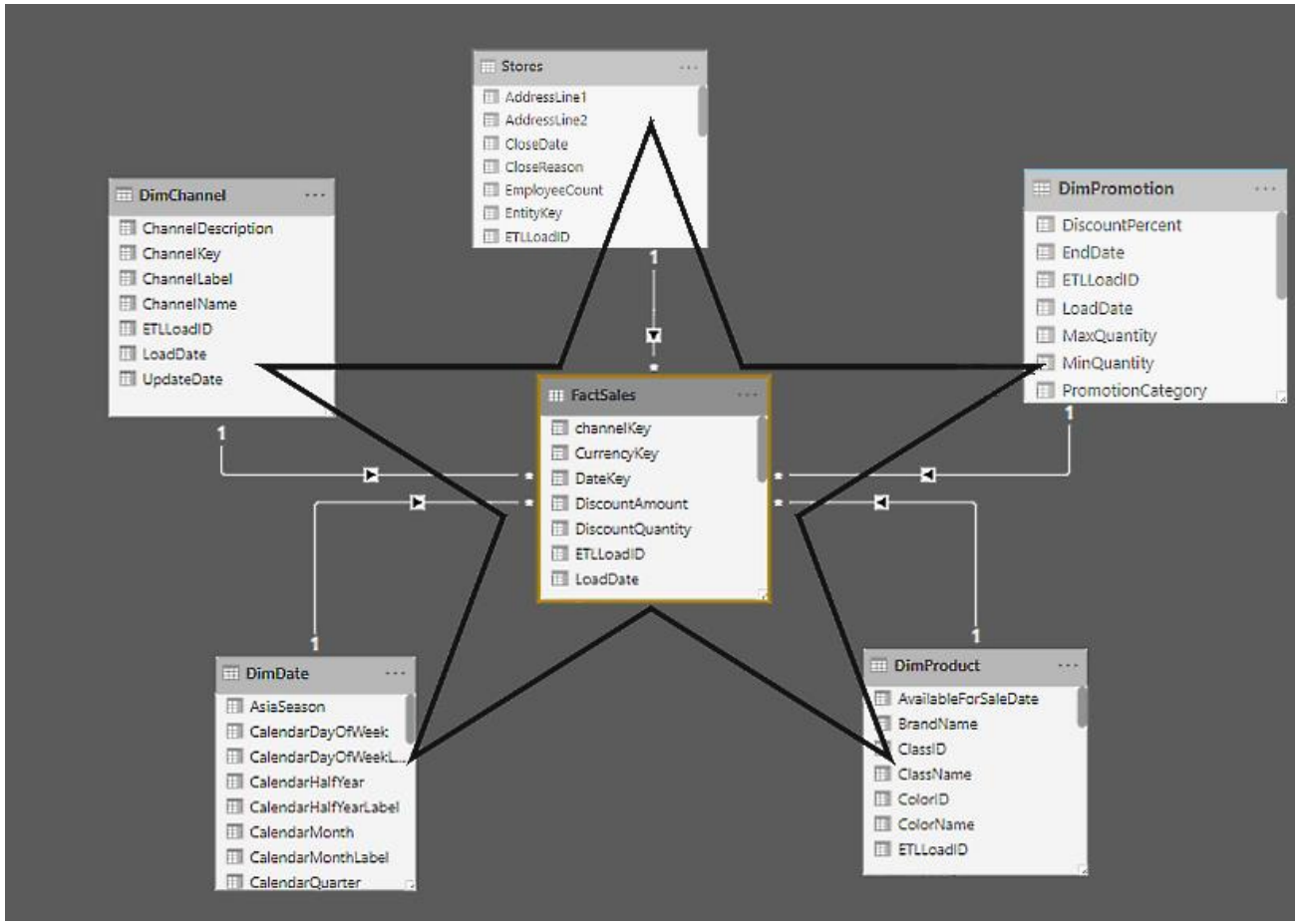
Proposed Solution

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

DATA MODEL
Optimize Your File - Optimize Schema

MASTERING DATA ANALYTICS

Keeping your data model simple



MASTERING DATA ANALYTICS

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

MASTERING DATA ANALYTICS

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

MASTERING DATA ANALYTICS

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

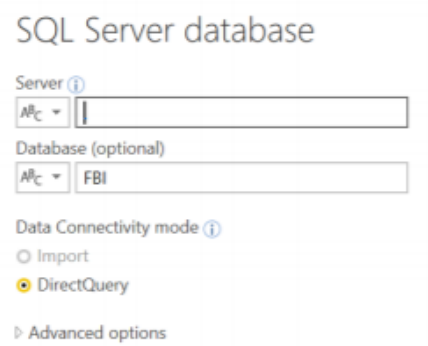
MASTERING DATA ANALYTICS

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

MASTERING DATA ANALYTICS

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

MASTERING DATA ANALYTICS

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 !!

MASTERING DATA ANALYTICS

Columnar Database

Row Based 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

PBI - Columnar Database

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

- 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**

MASTERING DATA 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)

MASTERING DATA ANALYTICS

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