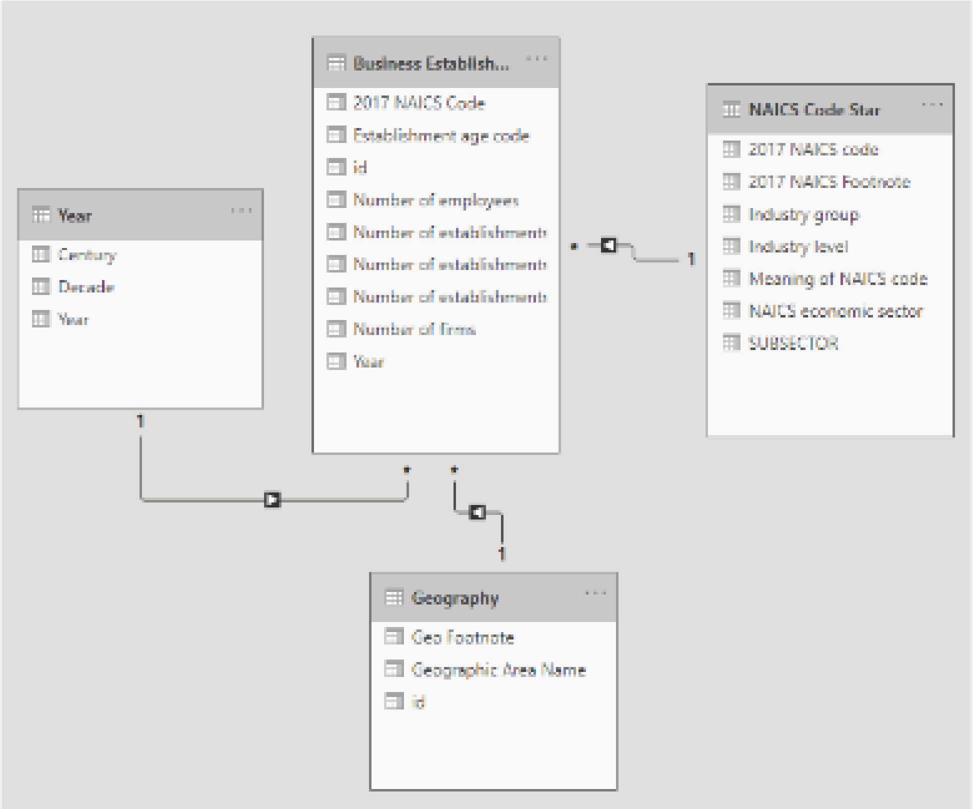
Data Modeling

Dr. Ernesto Lee

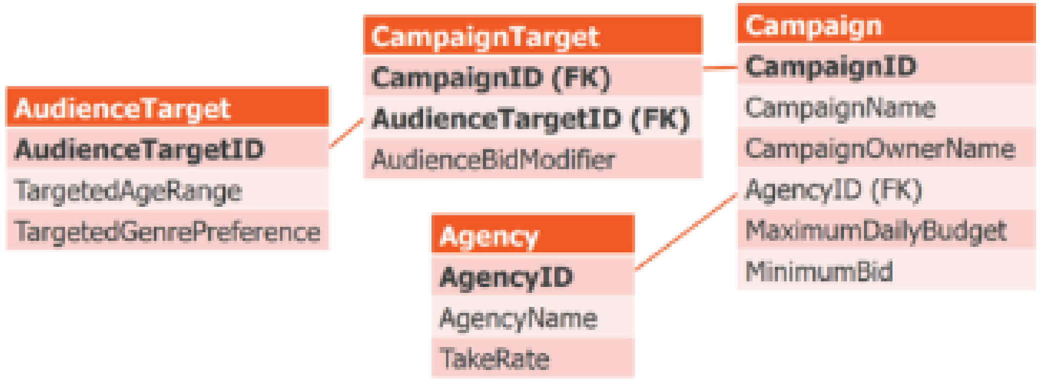
What is a data model?

- Conceptual view of data elements
- Typically a visual representation
- Data models include:
  - Tables
  - Columns
  - Relationships between tables
  - Data types
  - Keys



Data modeling

- The process of creating a data model
- Why model data?
  - Data ≠ perfect
  - Reshape data for analysis
  - Compress data usage
  - Easier to understand model



AudienceTarget
- **AudienceTargetID**
- TargetedAgeRange
- TargetedGenrePreference

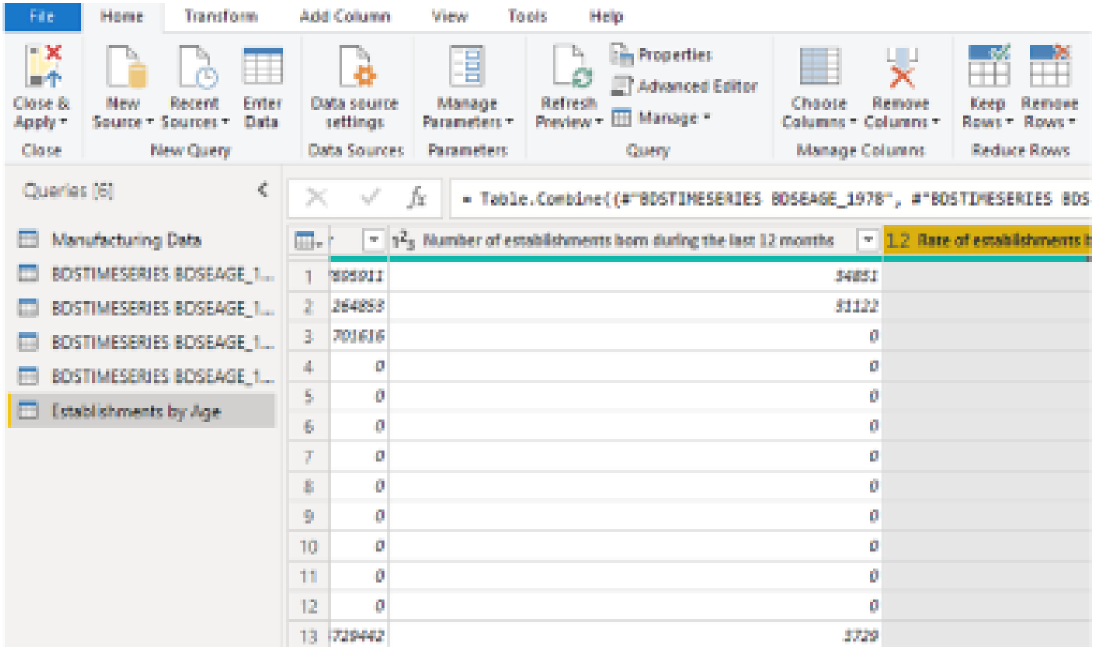
CampaignTarget
- **CampaignID (FK)**
- **AudienceTargetID (FK)**
- AudienceBidModifier

Campaign
- **CampaignID**
- CampaignName
- CampaignOwnerName
- AgencyID (FK)
- MaximumDailyBudget
- MinimumBid

Agency
- **AgencyID**
- AgencyName
- TakeRate

Data modeling in Power BI & Power Query

- Power Query is the data preparation tool of different Microsoft products, including Power BI
- Main goals:
  - Manage queries
  - Data modeling
- Data modeling: 80% in Power Query, 20% in Power BI



```
= Table.Combine({#"BDSTIMESERIES BDSEAGE_1978", #"BDSTIMESERIES BDS
```

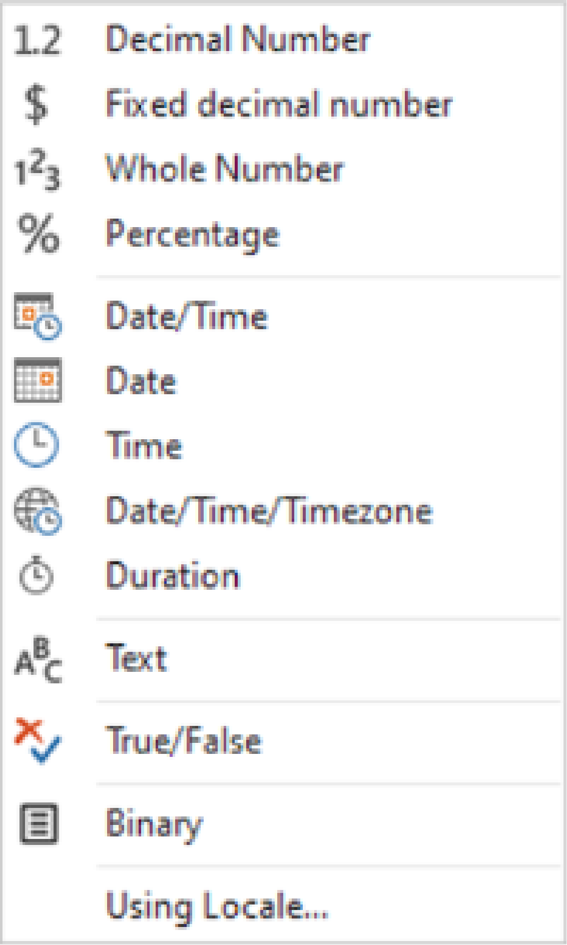
	'	1²3 Number of establishments born during the last 12 months	1.2 Rate of establishments b
1	'895911	34851	
2	264853	31122	
3	701616	0	
4	0	0	
5	0	0	
6	0	0	
7	0	0	
8	0	0	
9	0	0	
10	0	0	
11	0	0	
12	0	0	
13	'729442	3729	

Columns and row management

Operation	Example
Keep or remove specific columns	Remove empty column
Keep or remove specific rows	Keep top row as header
Split a single column in multiple columns	DD/MM/YYYY column split in DD, MM, YYYY columns
Summarize/group rows in a table by the contents of a column	Sum or median of all rows

Data types

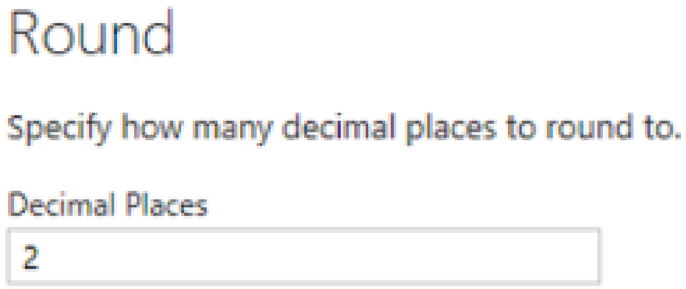
- Choosing the right data type is essential:
  - Constrain data to a specific shape
  - Optimize storage
  - Enable specific functionality
- Power Query infers data type on first few hundred rows



Rounding

Power Query

- Actually *changes* the data, not just formatting
- Typically not the right answer



Round

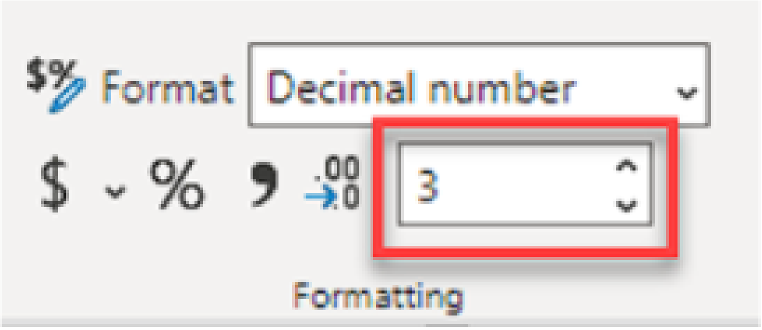
Specify how many decimal places to round to.

Decimal Places

2

Power BI

- Changes how the data *appears*, not how it's stored
- Generally a better answer than rounding in Power Query

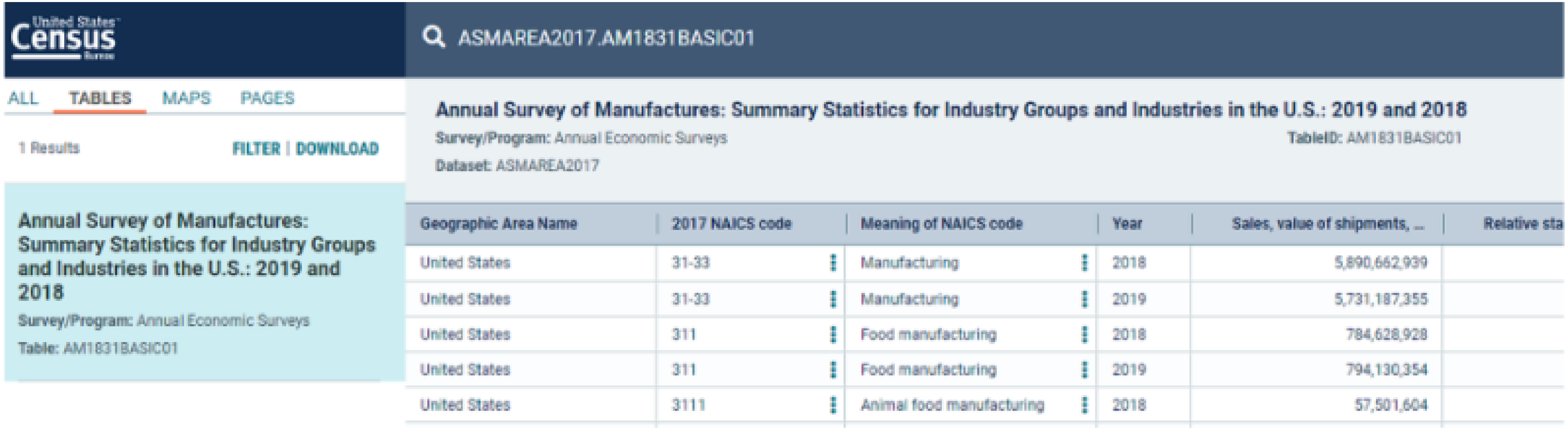


Format Decimal number

3

The dataset

- United States Census Bureau survey data of manufacturers
- Summary statistics for manufacturing firms
- North American Industry Classification System (NAICS)



ASMAREA2017.AM1831BASIC01

1 Results

Annual Survey of Manufactures: Summary Statistics for Industry Groups and Industries in the U.S.: 2019 and 2018

Survey/Program: Annual Economic Surveys

Table: AM1831BASIC01

Annual Survey of Manufactures: Summary Statistics for Industry Groups and Industries in the U.S.: 2019 and 2018

Survey/Program: Annual Economic Surveys

TableID: AM1831BASIC01

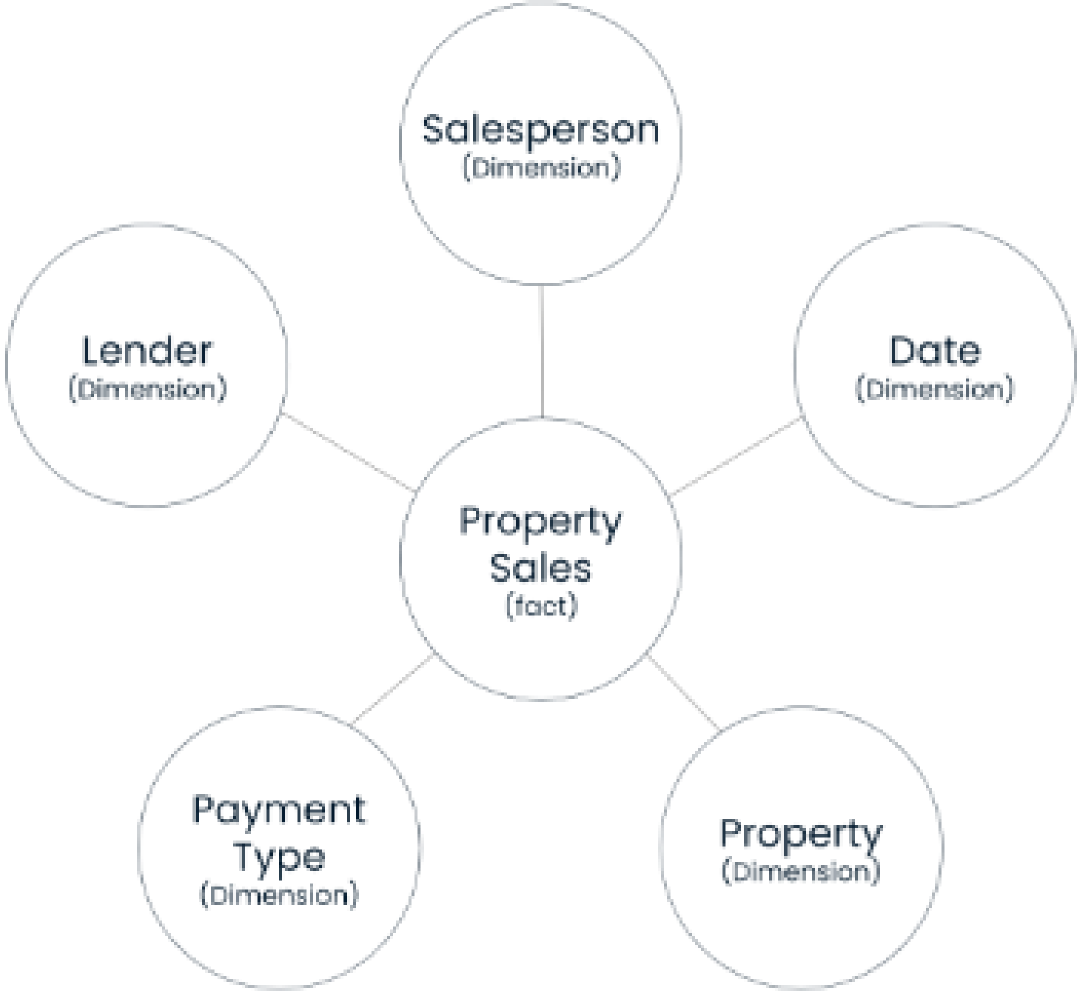
Dataset: ASMAREA2017

Geographic Area Name	2017 NAICS code	Meaning of NAICS code	Year	Sales, value of shipments, ...	Relative sta
United States	31-33	Manufacturing	2018	5,890,662,939	
United States	31-33	Manufacturing	2019	5,731,187,355	
United States	311	Food manufacturing	2018	784,628,928	
United States	311	Food manufacturing	2019	794,130,354	
United States	3111	Animal food manufacturing	2018	57,501,604	

Load and Clean the Data

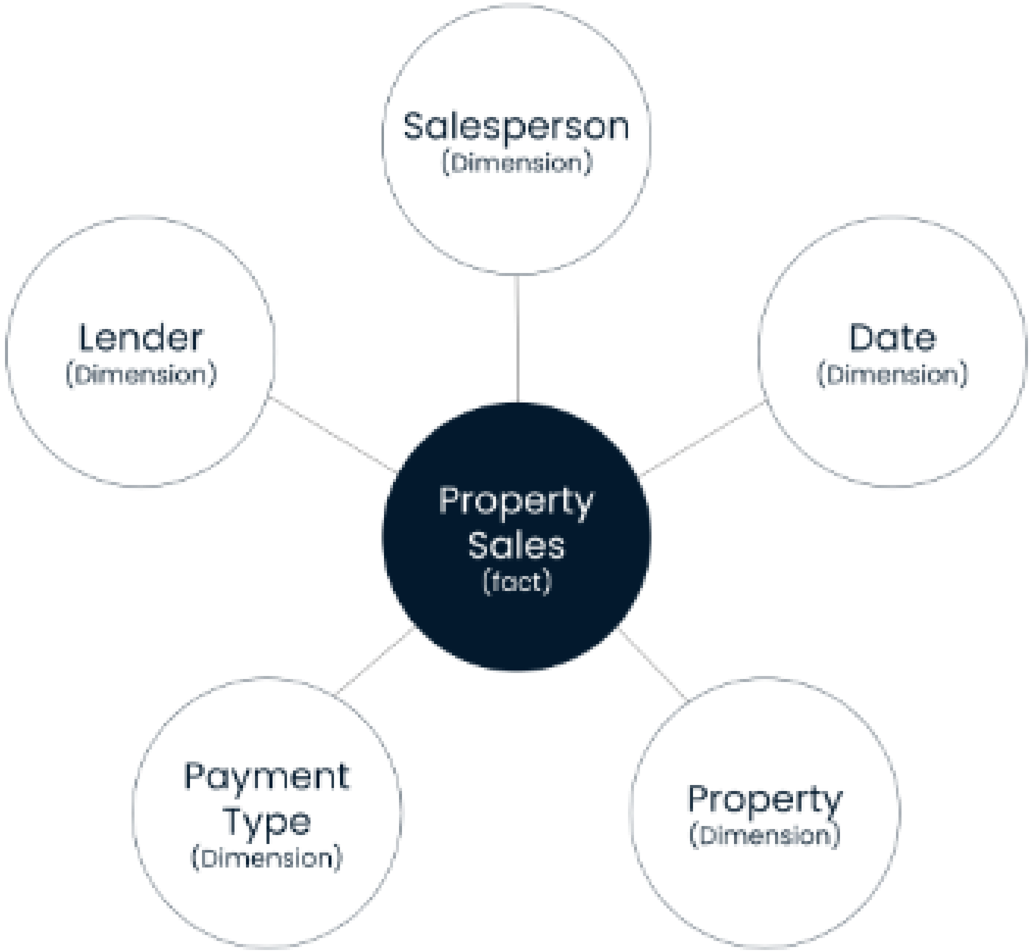
- https://github.com/fenago/cts2451/blob/main/modelling/Datasets/manufacturing_data.csv

- Key concepts
  - **Facts**: metrics from a business process
  - **Dimensions:** context surrounding a business process
  - Combine to form a **star schema**
- Star schemas are used in data warehouses
- Power BI is optimized for star schemas



Fact tables

- Made up of
  - **Facts (measures)**
    - Measurements or metrics from your business process
  - **Keys**
    - Used to establish relationships between fact and dimension tables
- Fact tables are long and narrow
  - Lots of rows
  - Fewer columns



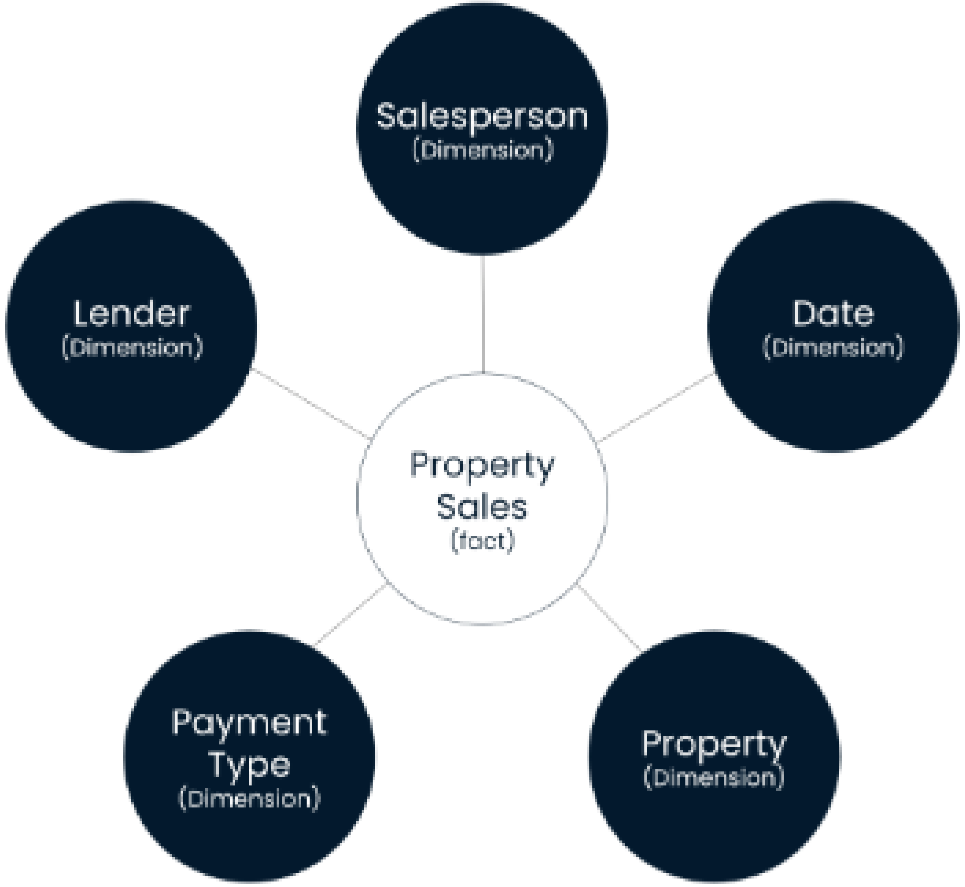
Fact tables: an example

Property Sales table

LenderID	StartDateID	PropertyID	PaymentTypeID	SalesPersonID	Rent	Duration
CO76	20200624	PG14	P2	SA9	750	24
CO56	20200907	PG4	P4	SA12	1250	12
CO62	20201201	PG16	P3	SA5	3000	36
CO43	20200201	PG6	P3	SA6	500	24
CO76	20200530	PG20	P2	SA6	5000	12
CO76	20200115	PG11	P2	SA2	2000	24
CO32	20201201	PG15	P2	SA9	450	36
...	...	...	...	...	...	...

Dimension tables

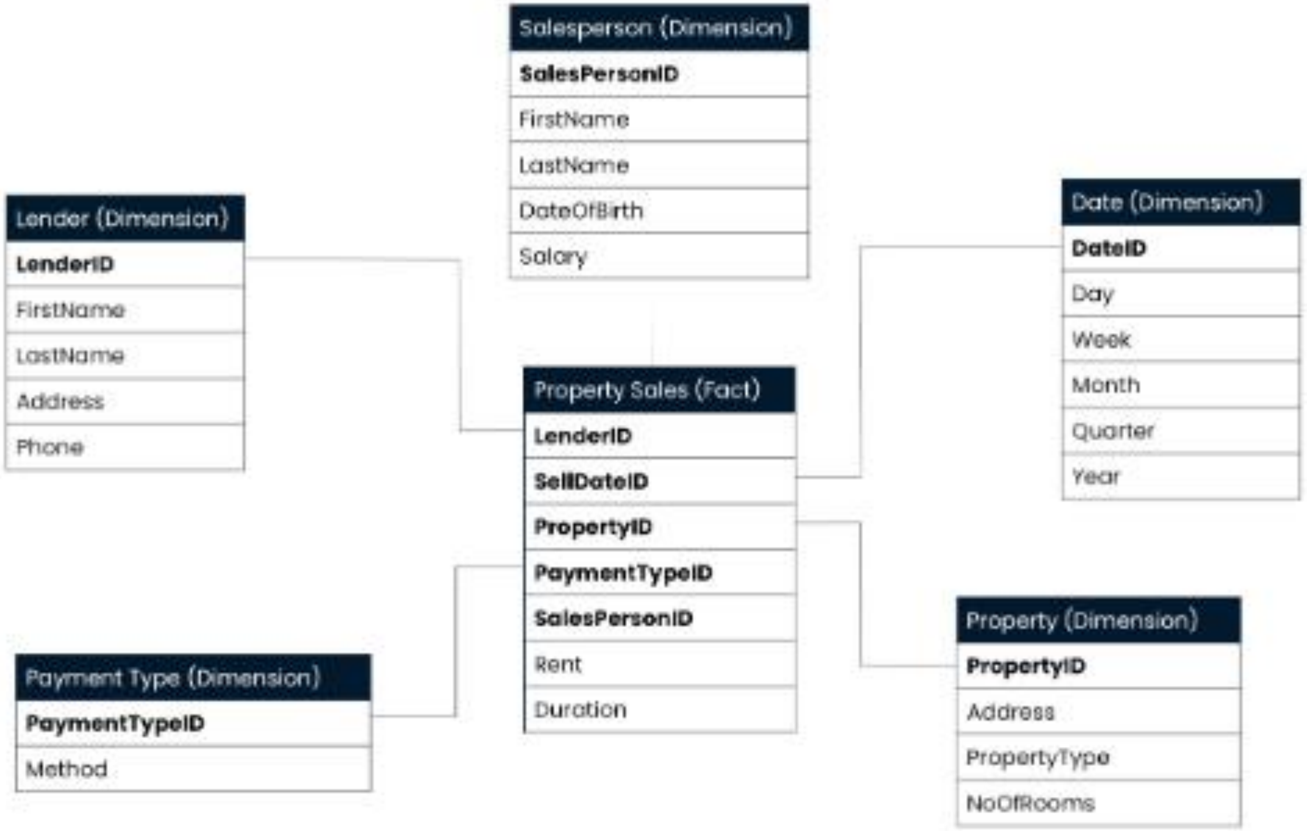
- Provide context
  - Who, what, when, where, why?
- Shared business concepts
  - E.g., person, employee, customer, vendor
- Contain static or "slowly changing" data
  - E.g., name, date of birth, height
- Dimension tables are short and wide
  - Few rows
  - Lots of columns



Dimension tables: an example

Salesperson table

SalesPersonID	FirstName	LastName	DateOfBirth	Salary
SA9	Mary	Howe	1990-02-19	24000
SA12	David	Ford	1978-03-24	18000
SA5	Ann	Beech	1980-11-10	12000
SA6	Julie	Lee	1985-06-13	30000
SA9	John	White	1965-10-01	9000
...	...		...	



Salesperson (Dimension)
SalesPersonID
FirstName
LastName
DateOfBirth
Salary

Lender (Dimension)
LenderID
FirstName
LastName
Address
Phone

Date (Dimension)
DateID
Day
Week
Month
Quarter
Year

Property Sales (Fact)
LenderID
SellDateID
PropertyID
PaymentTypeID
SalesPersonID
Rent
Duration

Payment Type (Dimension)
PaymentTypeID
Method

Property (Dimension)
PropertyID
Address
PropertyType
NoOfRooms

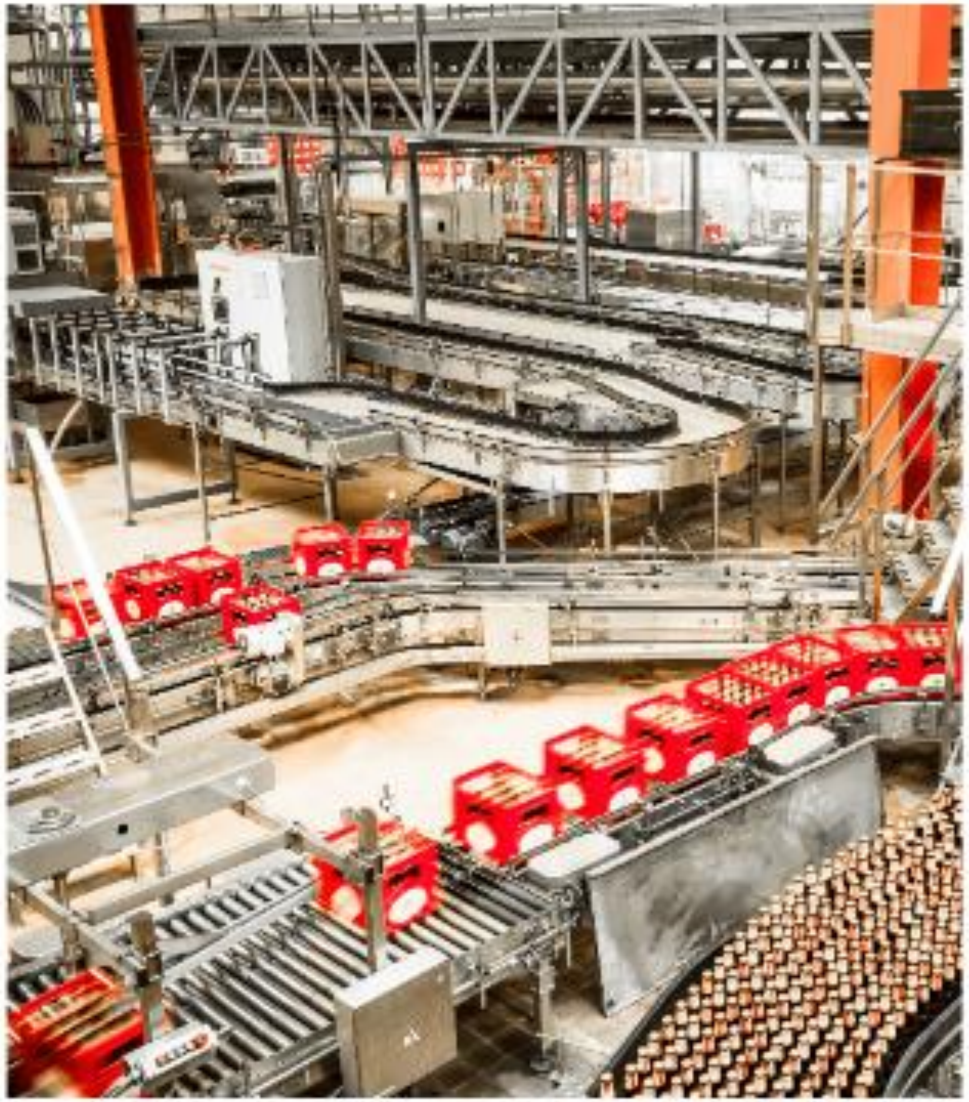
- Dimensions are used in multiple facts
- Dimensions do not link to other dimensions

Fact

- **Establishment Survey:** number of employees, number of firms, ...

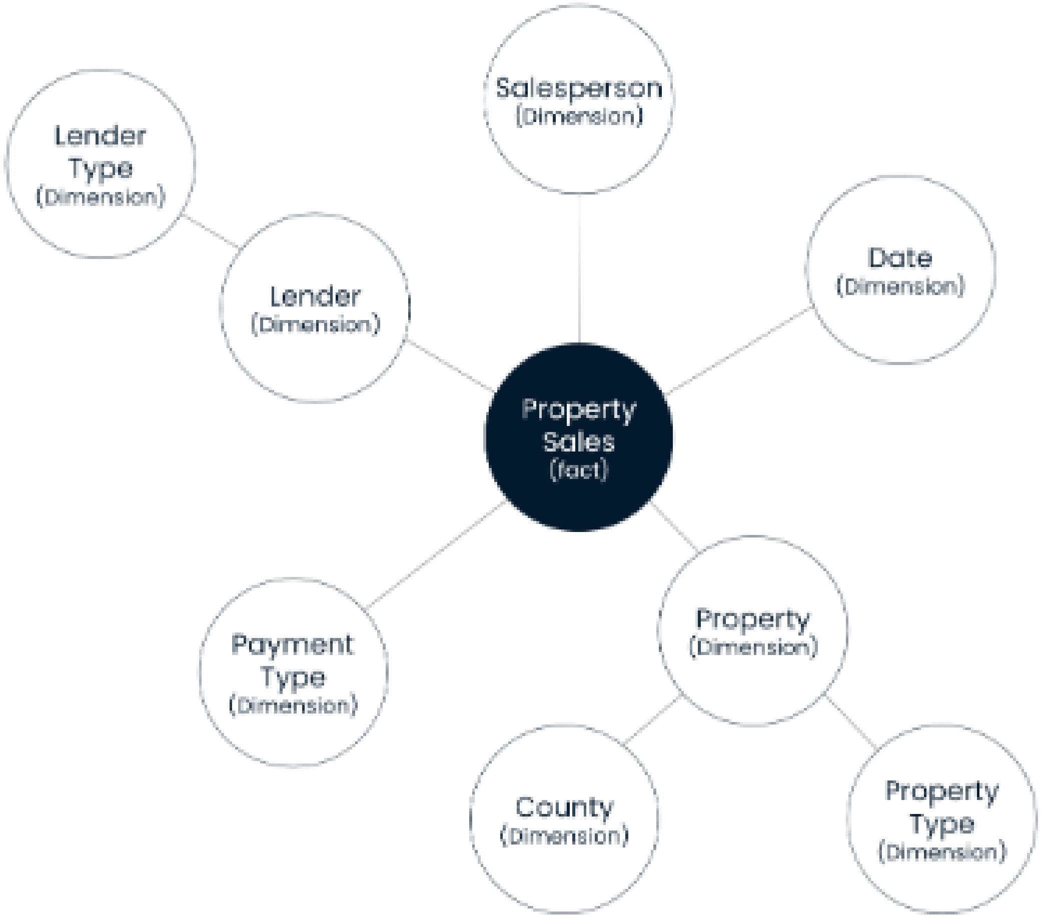
Dimensions

- **Industry:** NAICS code, industry group, subsector, sector
- **Time**: year, decade, century
- **Age**: establishment age
- **Geography**: country, state



Snowflake schema

- Allows relationships between dimensions



A closer look

Star schema

ProductKey	Name	SubCategory	Category
P1	Gloves	Hand	Clothing
P2	Shoes	Foot	Clothing
P3	Laptop	Computers	Electronics
P4	Mittens	Hand	Clothing

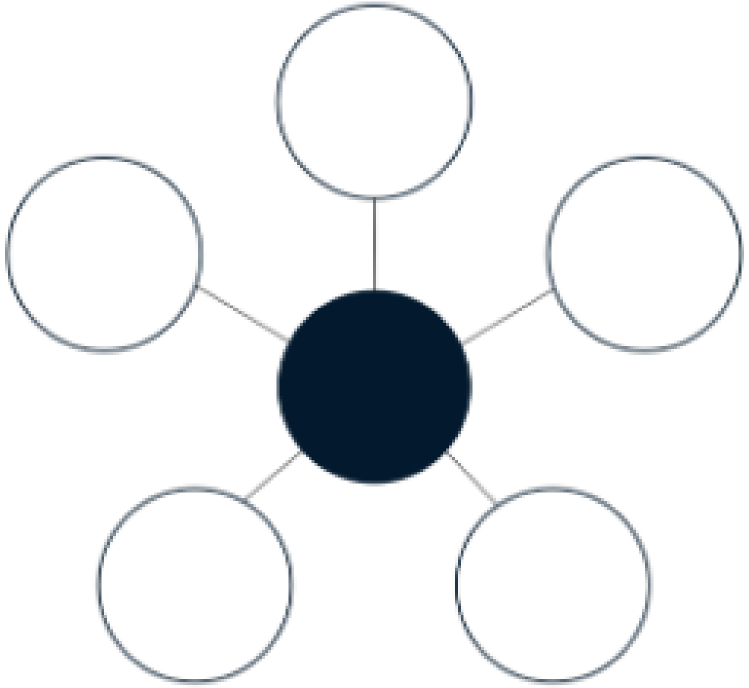
Snowflake schema

ProductKey	Name	SubCategoryKey
P1	Gloves	S1
P2	Shoes	S2
P3	Laptop	S3
P4	Mittens	S1

SubCategoryKey	SubCategory	CategoryKey
S1	Hand	C1
S2	Foot	C1
S3	Computers	C2

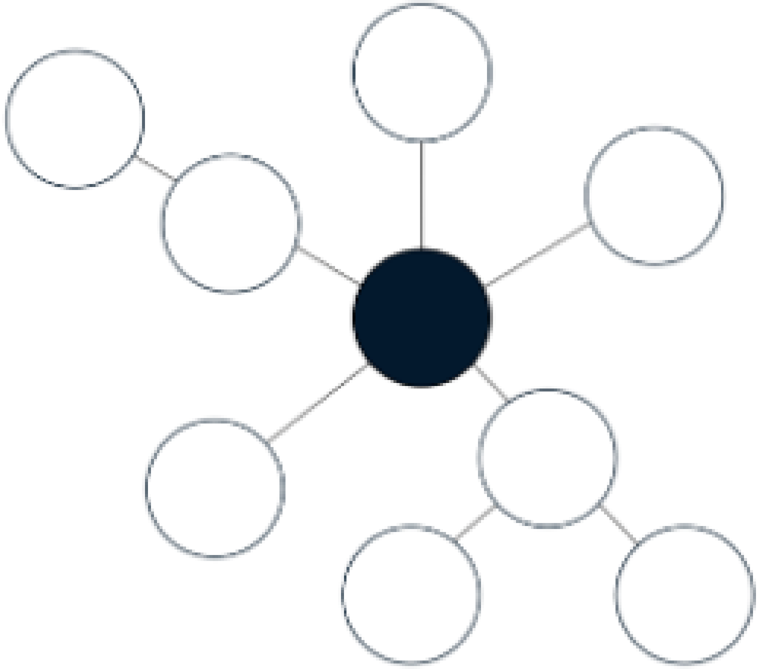
CategoryKey	Category
C1	Clothing
C2	Electronics

Star schema



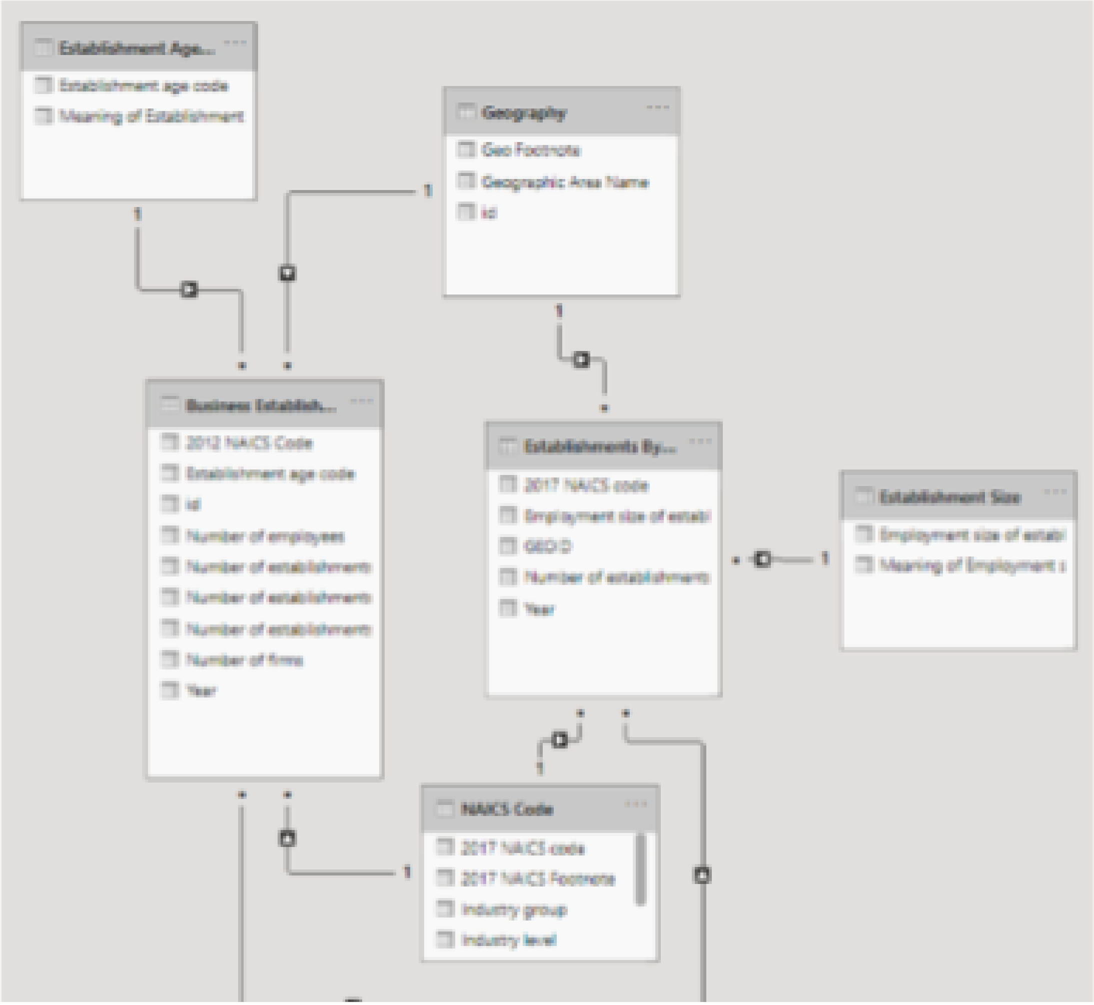
- Preferred approach
- Easy for business users to understand
- Most BI tools optimize for this schema

Snowflake schema



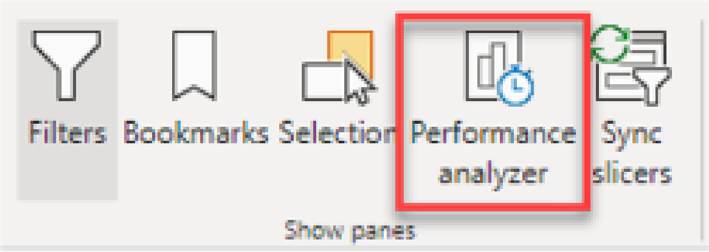
- Used in some data warehouses
- Less duplication
- Updating records is more efficient

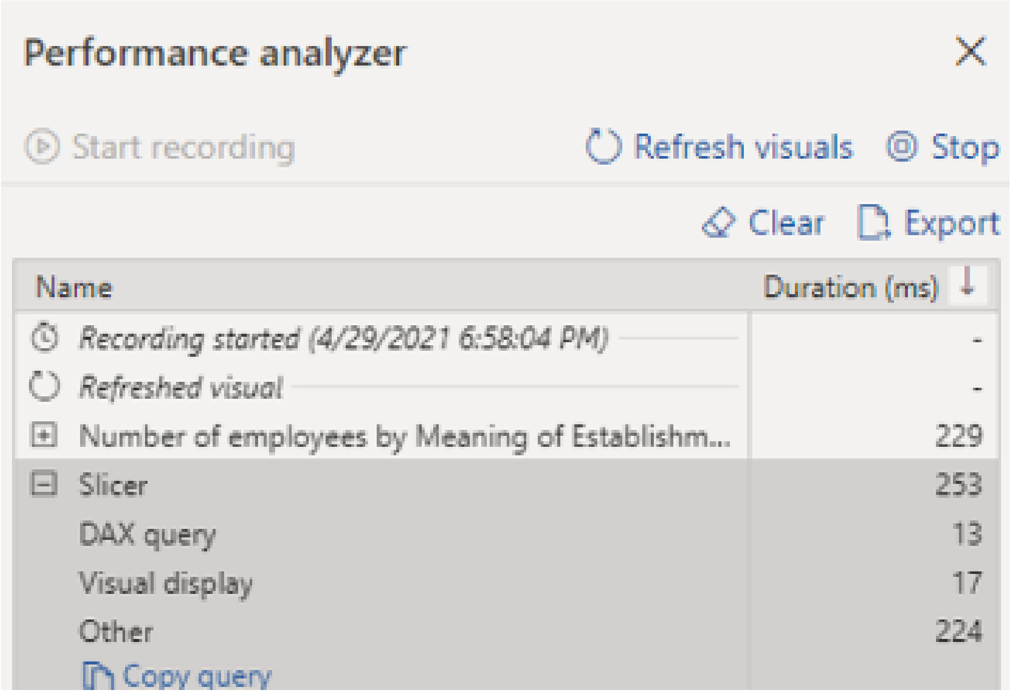
Stars and snowflakes in Power BI



- Both schemas work!
- But Power BI prefers star schemas
  - Easier to understand
  - Performance is less of a concern

The performance analyzer





Name	Duration (ms)
Recording started (4/29/2021 6:58:04 PM)	-
Refreshed visual	-
⊞ Number of employees by Meaning of Establishm...	229
⊟ Slicer	253
DAX query	13
Visual display	17
Other	224

- Built-in performance analysis
- Each visual has three components
  - How long did the DAX query take?
  - How long did the visual take to render?
  - How long did everything else take?

Performance tuning advice

DAX Query slowness

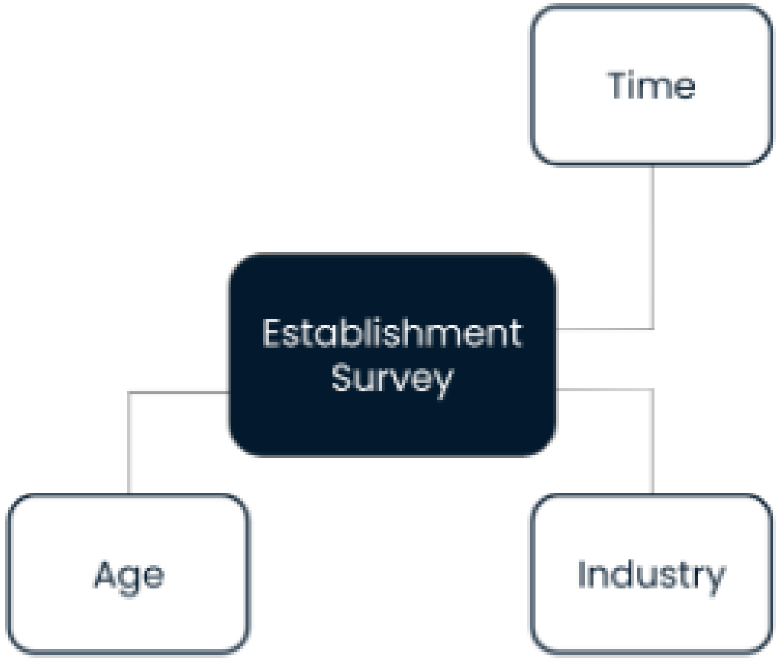
- Tune DAX operations
- Improve data loading performance

Visual display slowness

- Use less complicated visuals
- Show less information on the screen

Other slowness

- Reduce number of visuals on the page



- Which subsector has the highest average number of employees?

  How many employees did the Food Manufacturing subsector count on average in the 90’s?

  How many average employees did 3-year old firms in the Food Manufacturing subsector have in the 90’s?