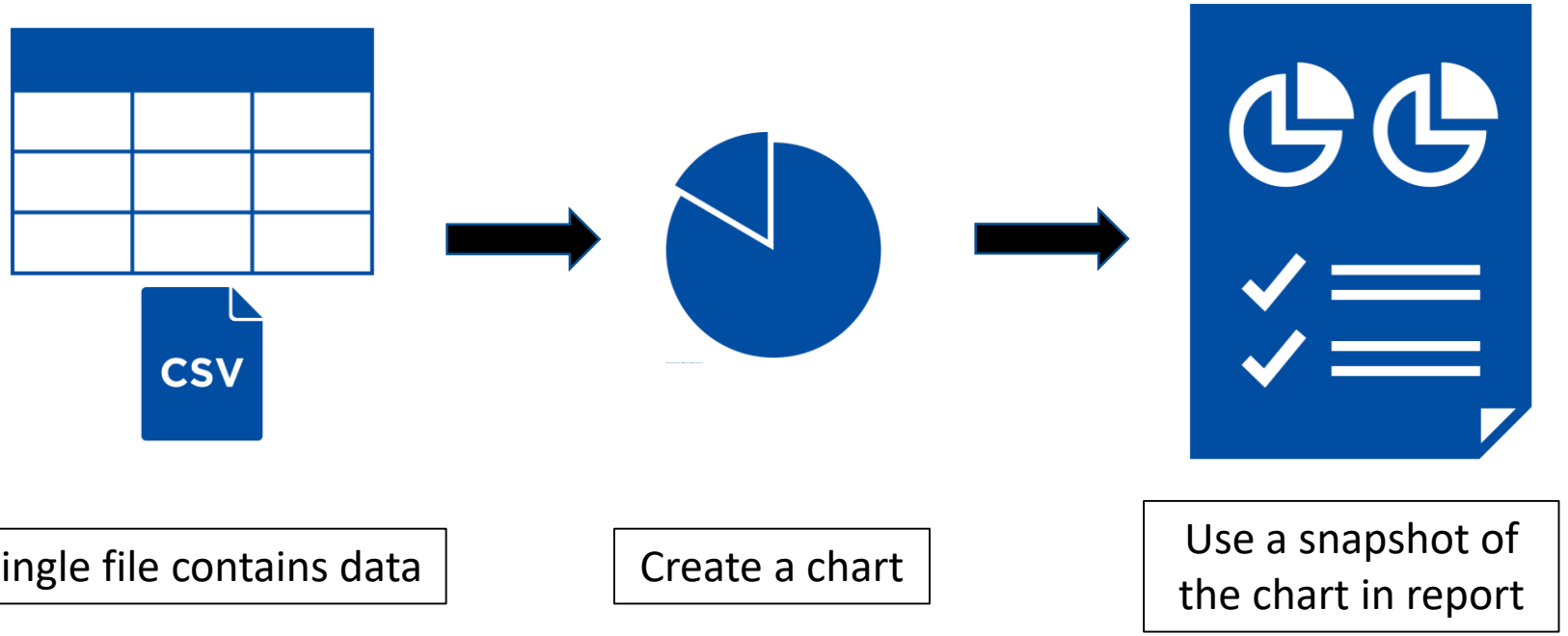
Start-Tech Academy

Data Visualization Process

Too simple?



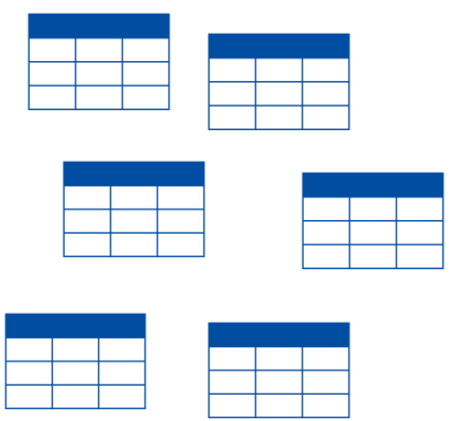
Data Visualization Process

Practical challenges

- Data is in multiple files
- Data is in different types of files
- Data is getting dynamically updated
- Ability to create all popular charts
- Ability to customize charts as required
- Ability to draw multiple charts and create dashboards

Data Visualization Process

Practical challenges



Multiple data sources

Steps

1. Connect with data sources
2. Collate and prepare data
3. Draw all types of charts
4. Customize the charts

Dynamic & Interactive dashboard

Share, Embed & View

Collaborate

Tableau Products

- Tableau Desktop
- Tableau Prep
- Tableau Server
- Tableau Online
- Tableau Public

Tableau Desktop

1. Windows/ Mac application for PC
2. The mail visualization tool by Tableau
3. Can connect with Tableau Online/ Tableau Public for collaboration with team members

Tableau Products

- Tableau Desktop
- Tableau Prep
- Tableau Server
- Tableau Online
- Tableau Public

Tableau Prep

1. Used for preprocessing the data
2. Does data cleaning/ validation and integration using a visual interface
3. Tableau desktop creator license grants access to Tableau Prep also

Tableau Products

- Tableau Desktop
- Tableau Prep
- Tableau Server
- Tableau Online
- Tableau Public

Tableau Server/ Tableau Online

- Used for connecting and collaborating with team members

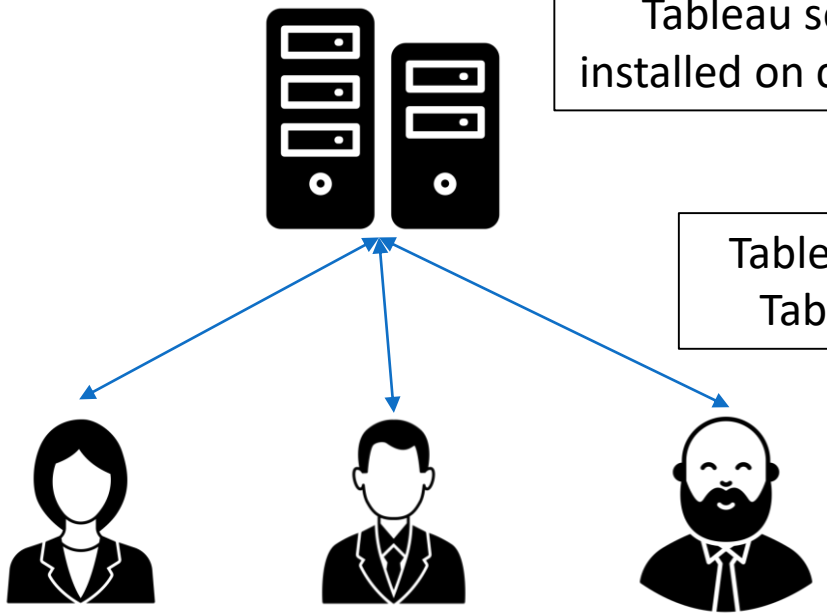


Tableau server will be installed on company server

Tableau Desktop will talk with Tableau server via Intranet

Individuals will be using Tableau Desktop on their PC

Tableau Products

- Tableau Desktop
- Tableau Prep
- Tableau Server
- Tableau Online
- Tableau Public

Tableau Server/ Tableau Online

- Used for connecting and collaborating with team members

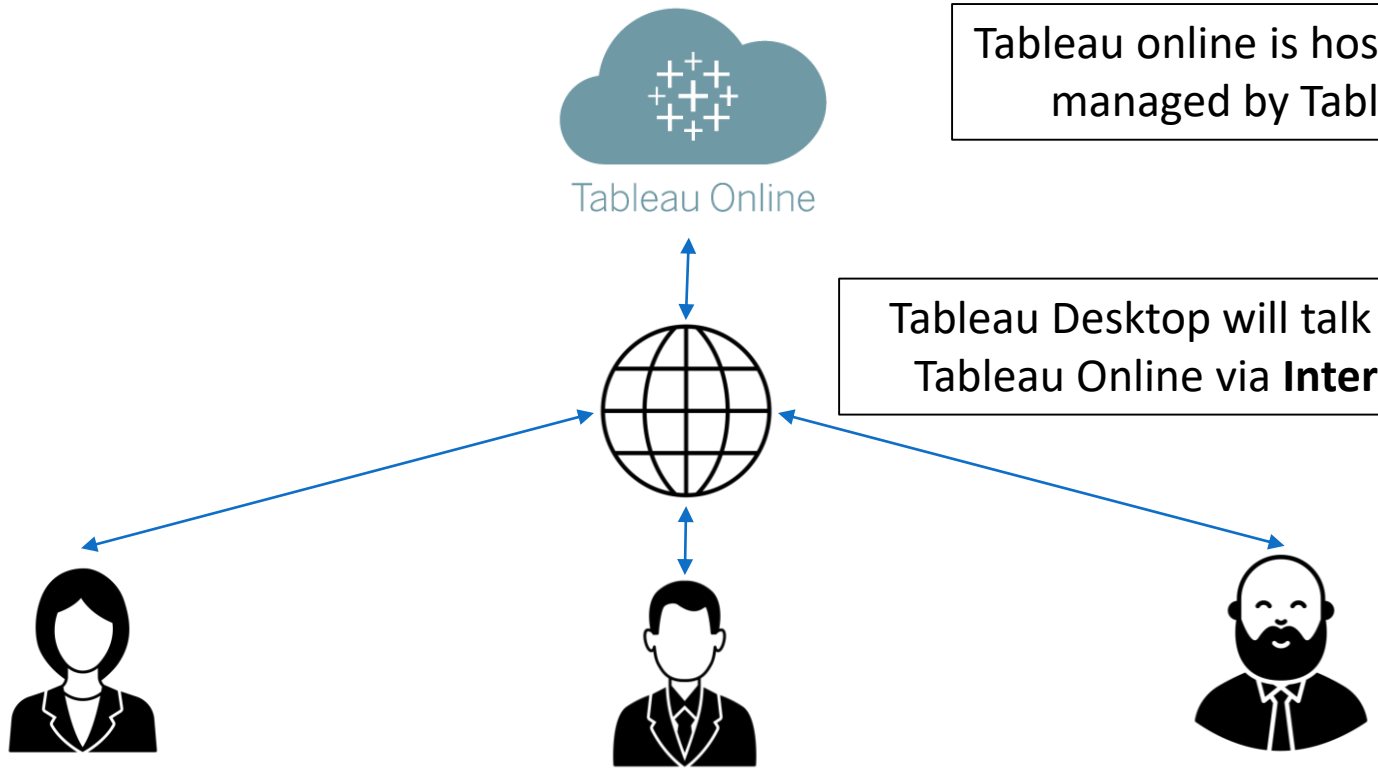


Tableau online is hosted and managed by Tableau

Tableau Desktop will talk with Tableau Online via **Internet**

Tableau Products

- Tableau Desktop
- Tableau Prep
- Tableau Server
- Tableau Online
- Tableau Public

Tableau Public

1. Hosted by Tableau – free for all to use
2. Your work will be open for anyone to see
3. Web based application

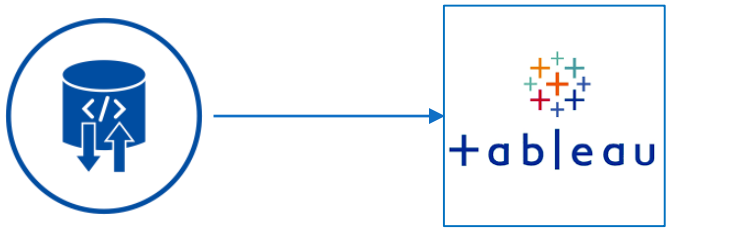
Live vs Extract Connection

Product

Connection
Live Extract

Filters
0 | Add

Real time updates



Data Source

Visualizations

Better performance

Data Source

Snapshot of data
Extract

Visualizations

Live connection

- Whenever any change is made, data is fetched from the data source
- Visualizations are updated in real-time

Extract

- Whenever any change is made, data is fetched from the snapshot (Extract)
- Extracts need to be are updated for any update in data

Joining data from multiple tables

Scenario 1: Joining two different tables

Loan Transaction Table

Loan amount only

Trans. ID	Application Date	Approval date	Type of Loan	Customer ID	Branch ID	Loan Amount	Term of Loan	Interest rate	Lead source
231	10-09-2015	29-09-2015	Home Loan	CC-12145	OFF-SU-10004498	1063320	30	7	Walk-in
287	27-08-2016	16-09-2016	Home Loan	SS-20875	OFF-SU-10004498	1187813	10	2.5	Website
537	11-02-2016	28-02-2016	Gold Loan	PB-19210	TEC-PH-10004897	138323	10	9.7	Website
707	06-08-2016	22-08-2016	Gold Loan	AH-10690	OFF-PA-10000474	90797	5	13.2	Website
301	20-05-2016	04-06-2016	Vehicle Loan	JK-15730	TEC-PH-10001530	271290	30	2.9	Partner Web.

Branch Table

Region values only

Branch ID	Employee Count	Type of Branch	Location type	Region
OFF-PA-10000474	15	Traditional-upgraded	Metro	Central
OFF-SU-10004498	13	Digi-Smart	Metro	West
TEC-PH-10001530	17	Traditional	Metro	East
TEC-PH-10004897	13	Digi-Smart	Urban	East

3 ways of joining data in tableau

Joining

Later

Blending

Relationships

New way of joining data in Tableau

Merging data from multiple tables

Scenario 2: Merging similar tables

Combining similar data using operators like Union, Intersect and Except

Customer ID	Customer Name	Segment	Age	Country	City	State	Postal Code	Region
EB-13870	Emily Burns	Consumer	34	United States	Orem	Utah	84057	West
EH-13945	Eric Hoffmann	Consumer	21	United States	Los Angeles	California	90049	West
TB-21520	Tracy Blumstein	Consumer	48	United States	Philadelphia	Pennsylvania	19140	East
MA-17560	Matt Abelman	Home Office	19	United States	Houston	Texas	77095	Central

Online customers

+

Customer ID	Customer Name	Segment	Age	Country	City	State	Postal Code	Region
ON-18715	Odella Nelson	Corporate	27	United States	Eagan	Minnesota	55122	Central
PO-18865	Patrick O'Donnell	Consumer	64	United States	Westland	Michigan	48185	Central
LH-16900	Lena Hernandez	Consumer	66	United States	Dover	Delaware	19901	East

Walk-in customers

↓

Customer ID	Customer Name	Segment	Age	Country	City	State	Postal Code	Region
EB-13870	Emily Burns	Consumer	34	United States	Orem	Utah	84057	West
EH-13945	Eric Hoffmann	Consumer	21	United States	Los Angeles	California	90049	West
TB-21520	Tracy Blumstein	Consumer	48	United States	Philadelphia	Pennsylvania	19140	East
MA-17560	Matt Abelman	Home Office	19	United States	Houston	Texas	77095	Central
ON-18715	Odella Nelson	Corporate	27	United States	Eagan	Minnesota	55122	Central
PO-18865	Patrick O'Donnell	Consumer	64	United States	Westland	Michigan	48185	Central
LH-16900	Lena Hernandez	Consumer	66	United States	Dover	Delaware	19901	East

Relationship – performance options

Cardinality

Customer table

Customer ID	Postal code
CG-12520	42420
DV-13045	90036
SO-20335	33311
BH-11710	90036

Are postal code values unique in this table?

Not unique -> **'Many'**

Postal code master table

Postal code	City	State	Region
32303	Tallahassee	Florida	South
32725	Deltona	Florida	South
32935	Melbourne	Florida	South
33012	Hialeah	Florida	South
33311	Fort Lauderdale	Florida	South
.	.	.	.
.	.	.	.

Are postal code values unique in this table?

Unique -> **'One'**

'Many to One'

We are joining customer table with postal code master - postal code is the matching key

Relationship – performance options

Cardinality

Customer table

Customer ID	Postal code
CG-12520	42420
DV-13045	90036
SO-20335	33311
BH-11710	90036

Is customer ID unique in this table?

Unique -> **'One'**

Reference table

Customer ID	Ref Name	Ref Contact
CG-12520	Cindy Stewart	10897310
CM-11935	Dan Campbell	16589278
CM-12385	Darren Koutras	95721837
CG-12520	Denny Ordway	16507437
CS-12355	Evan Bailliet	76772276
CS-12460	Erica Hackney	69524187

Is customer ID unique in this table?

Not unique -> **'Many'**

'One to Many'

We are joining customer table with reference table – customer ID is the matching key

Relationship – performance options

Cardinality

'Many to Many' is the default setting in Tableau

It will not give wrong visualizations

We can improve performance slightly if we specify the exact cardinality

Change only if you know what you are doing

Relationship – performance options

Referential Integrity

Customer table

Customer ID	Postal code
CG-12520	42420
DV-13045	90036
SO-20335	33311
BH-11710	90036

Reference table

Customer ID	Ref Name	Ref Contact
CG-12520	Cindy Stewart	10897310
CM-11935	Dan Campbell	16589278
CM-12385	Darren Koutras	95721837
CG-12520	Denny Ordway	16507437
CS-12355	Evan Bailliet	76772276
CS-12460	Erica Hackney	69524187

Are all customer IDs in customer table also present in the reference table?

No -> **'Some records match'**

Are all customer IDs in reference table also present in the customer table?

Yes -> **'All records match'**

We are joining customer table with reference table – customer ID is the matching key

Relationship – performance options

Referential Integrity

'Some records match' is the default setting in Tableau

It will not give wrong visualizations

We can improve performance if we specify the exact referential integrity

Change only if you know what you are doing

Physical vs Logical layer

Observations

Physical Layer

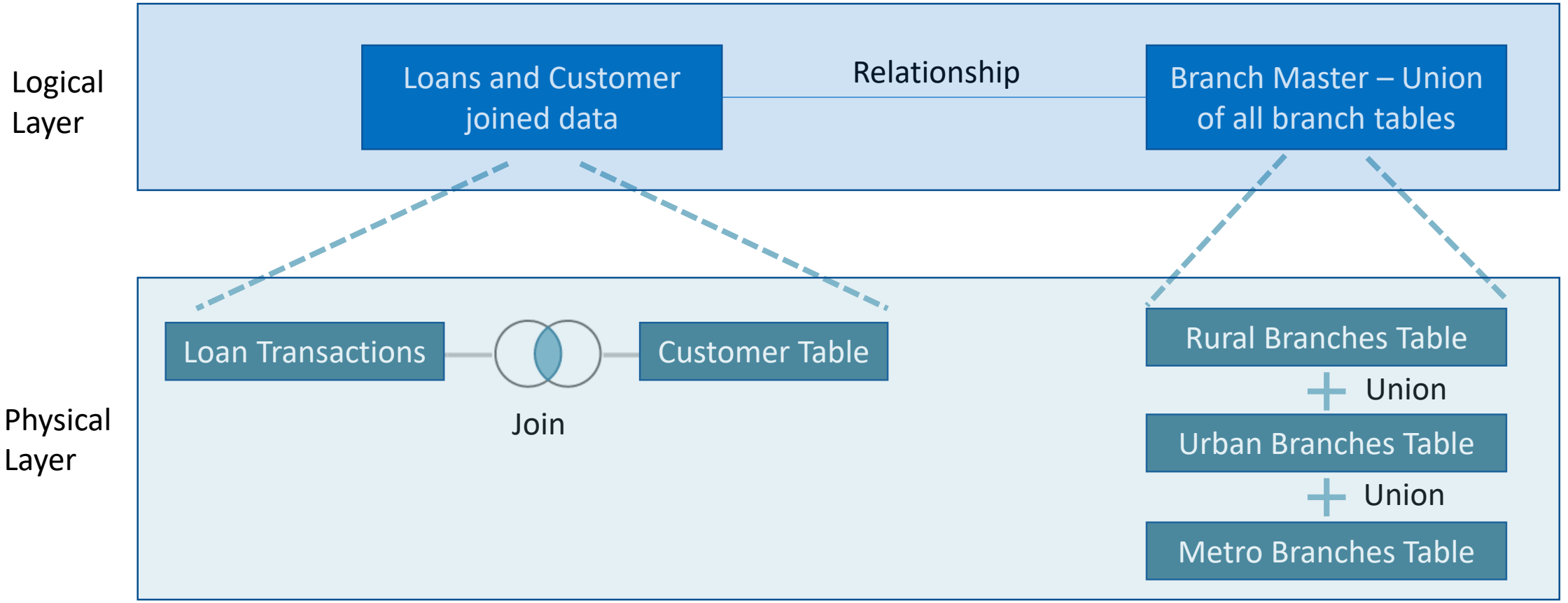
- We can do Join and Union here
- Result of joining and union is a single table

Logical Layer

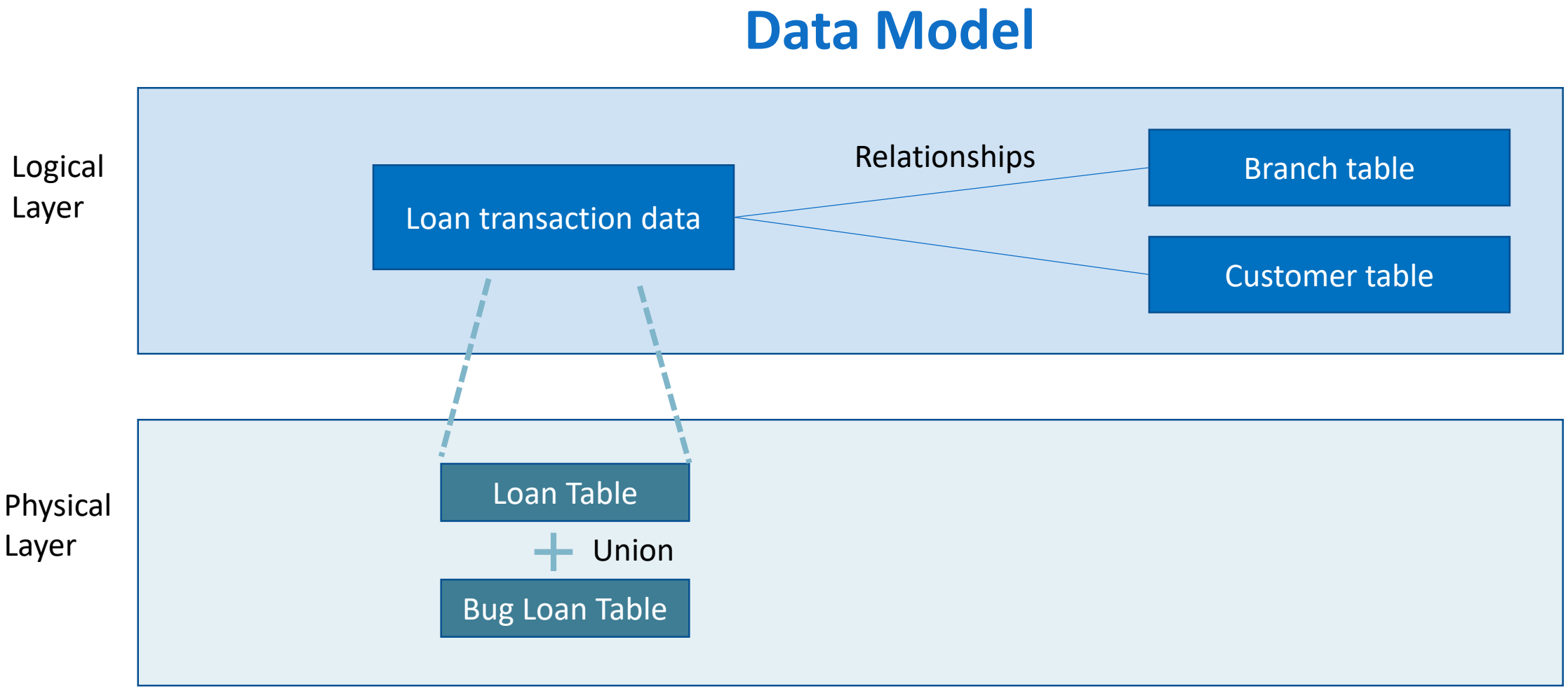
- Single logical table can have multiple joined and Union tables
- Relationships can be defined – ‘Noodles
- Related tables remain individual tables

Physical vs Logical layer

Data Model - Example



Physical vs Logical layer



Types of Data in Tableau

Dimensions vs measures

- Dimensions are columns containing categories/ segments based on which aggregation will be done
- Measures are numeric columns for which we wish to get the aggregate values

Student ID	Gender	Age	Hours studied	Marks scored	Year of exam
S101	Male	19	18	73	2021
S102	Female	20	15	85	2020
S103	Female	16	21	71	2023
S104	Male	19	23	89	2022
S105	Female	19	25	94	2022
S106	Female	20	27	70	2020
S107	Male	21	15	95	2019
S108	Male	21	20	70	2021
S109	Female	16	17	79	2019
S110	Female	18	26	73	2020
…	…	…	…	…	…

Gender	Sum of Hours studied
Female	131
Male	76

Year of Exam	Average of Marks scored
2019	87
2020	76
2021	71.5
2022	91.5
2023	71

Sum, Average, Min, Max, count are some aggregate functions

Types of Data in Tableau

Dimensions vs measures

- Same column can act as a dimension in some scenarios and as a measure in other scenarios

Student ID	Gender	Age	Hours studied	Marks scored	Year of exam
S101	Male	19	18	73	2021
S102	Female	20	15	85	2020
S103	Female	16	21	71	2023
S104	Male	19	23	89	2022
S105	Female	19	25	94	2022
S106	Female	20	27	70	2020
S107	Male	21	15	95	2019
S108	Male	21	20	70	2021
S109	Female	16	17	79	2019
S110	Female	18	26	73	2020
…	…	…	…	…	…

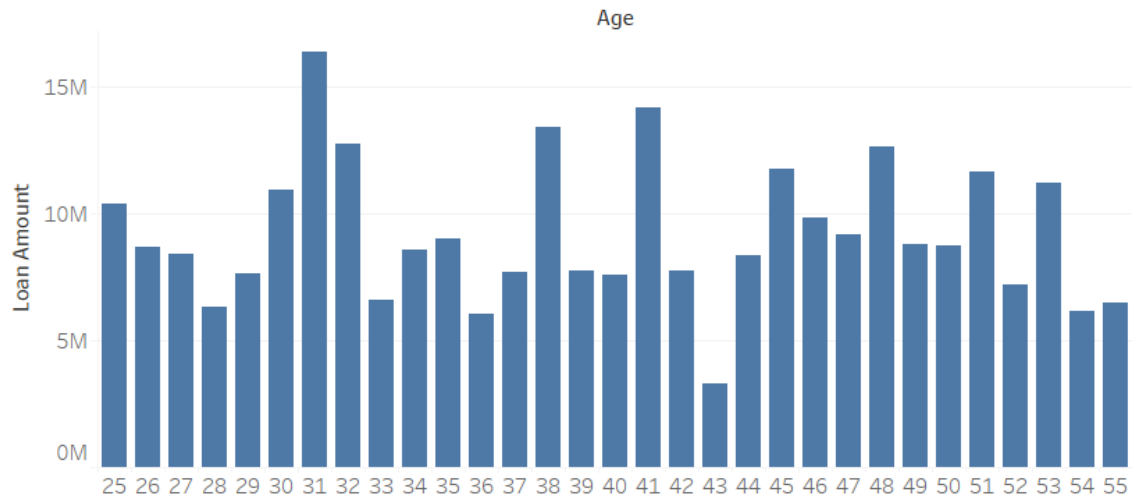
Age	Average of Hours studied
16	19
18	26
19	22
20	21
21	17.5

Gender	Average of Age
Female	18.2
Male	20

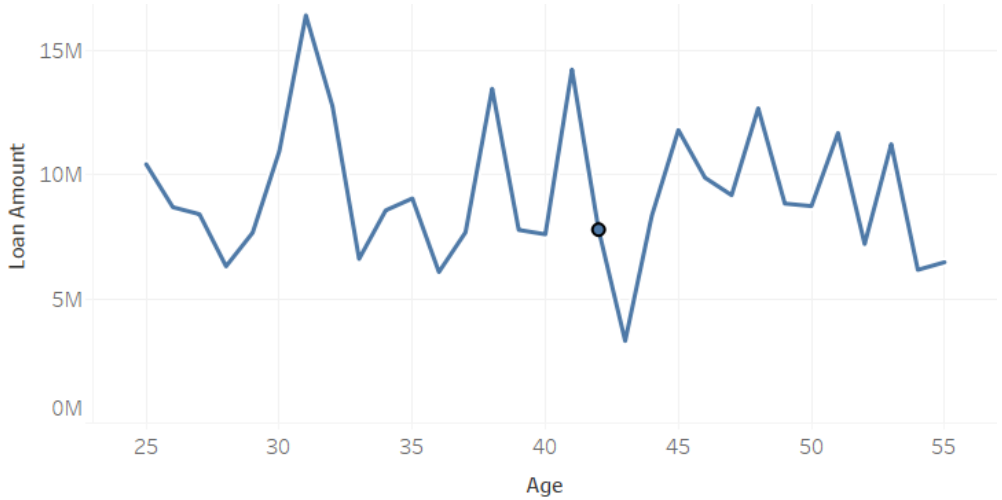
Types of Data in Tableau

Discrete vs continuous

- Discrete - A set of finite values
  - Will add headers on the axes
- Continuous – Infinite range
  - Will add infinite range on the axes



Age is set as discrete



Age is set as continuous

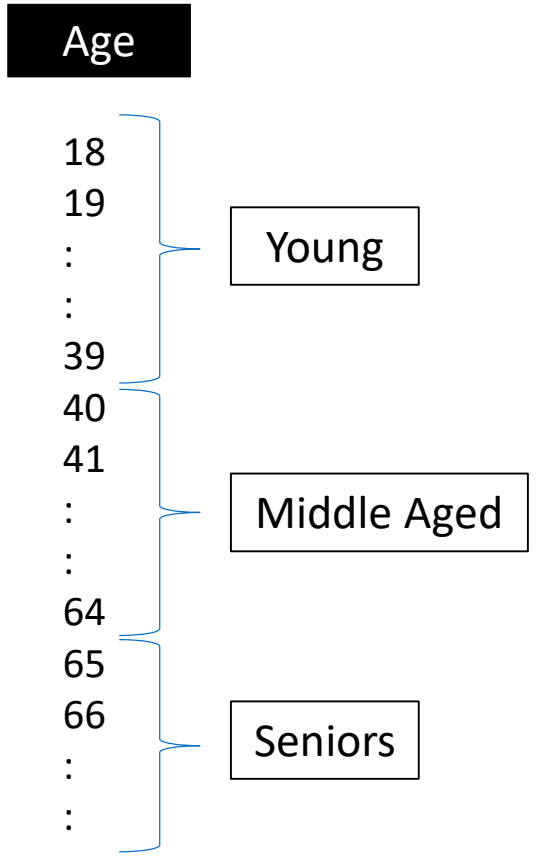
Types of Data in Tableau

	Discrete **(Blue)** **Finite – adds headers**	**Continuous** **(Green)** **Infinite range – adds axes**
Dimension String, date, numeric	Eg. - Name, Gender, category, etc. Common	Eg. Year(transaction date) Rare
Measure Numeric - aggregation	Eg. Sum(profit) Rare	Eg. Sum(profit), Average(age) Common

Binning data

Bins

Converting continuous numeric data into bins/ groups



Grouping data

Groups

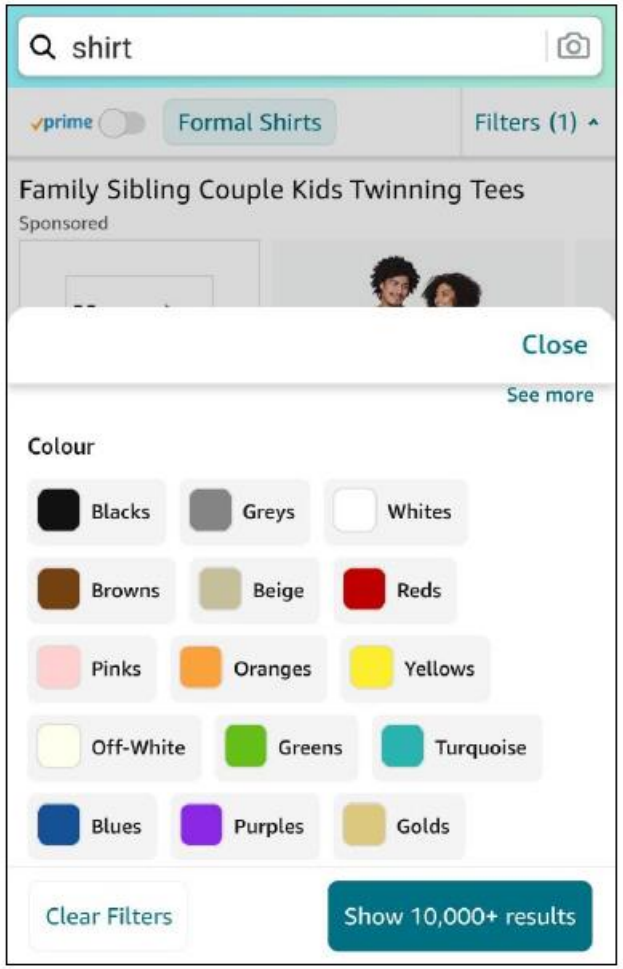
Clubbing similar categories together into groups

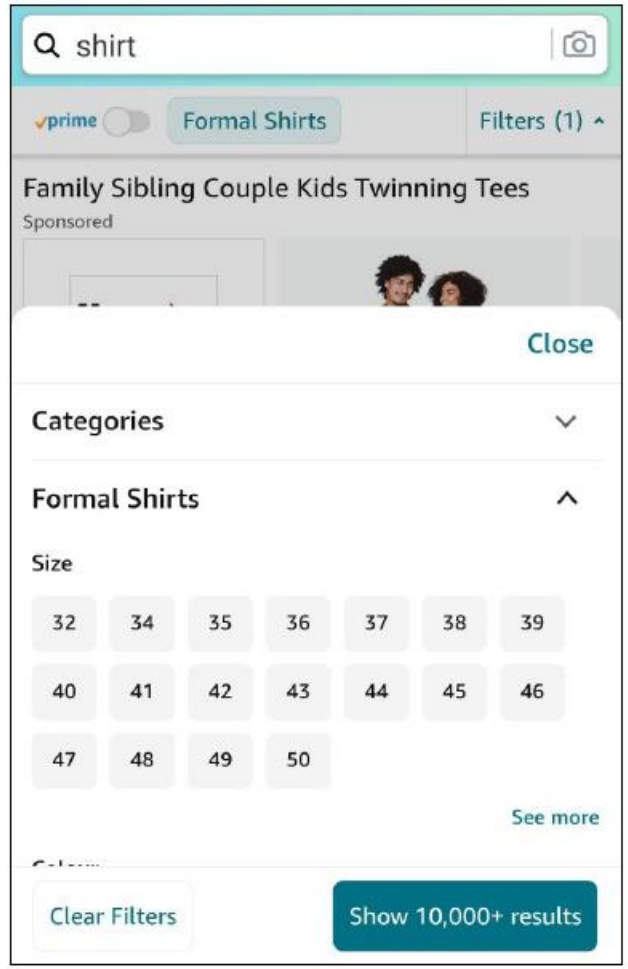
Sub-categories	Groups
Accessories	Phones & Acc.
Appliances	Others
Bookcases	Others
Chairs	Others
Copiers	Others
Envelopes	Others
Furnishings	Others
Labels	Others
Machines	Others
Paper	Others
Phones	Phones & Acc.
Storage	Others
Tables	Tables & Chairs

Filtering

Examples

Showing only relevant data/ hiding irrelevant information





Filters

Order of operation

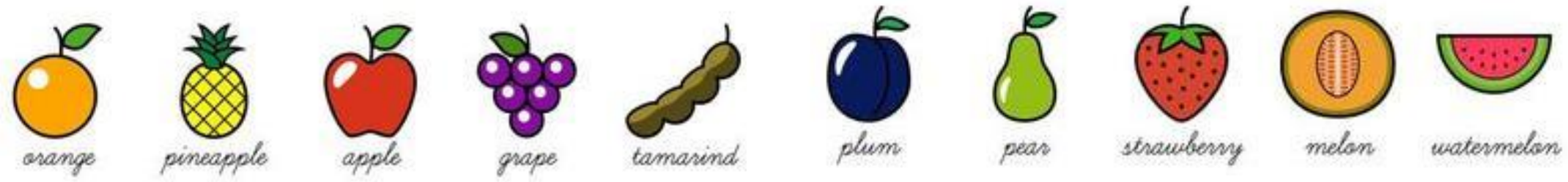
Order of operation is the order in which different types of filters are executed in Tableau

In Tableau, filters are executed in the following order:

1. Extract filters
2. Data source filters
3. Context filters
4. Filters on dimensions
5. Filters on measures

Order of execution is important because it can impact the final output and the performance when multiple filters are applied

Filters



Output of two dimension filters

Two filters

Filter 1 – Only red fruits (inside or outside)

Filter 2 – Top 2 fruits by weight

Calculation

Output of filter 1 – Apple, Strawberry and Watermelon

Output of filter 2 – Watermelon and melon

Final output – Only Watermelon

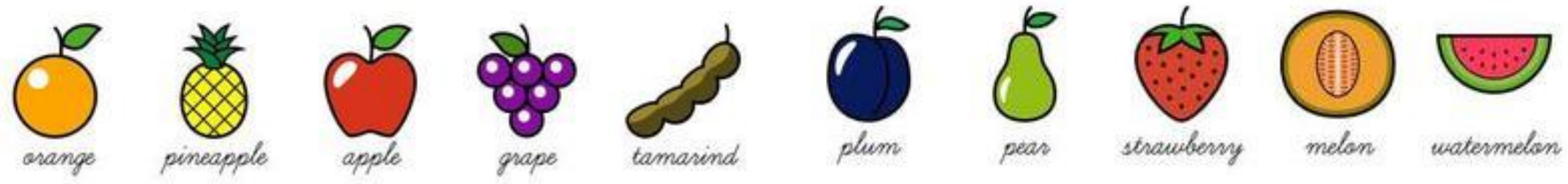
Steps

Filter 1 checks 10 fruits

Filter 2 checks 10 fruits

Common/ intersection of two results is shown finally

Filters



Output of one context and one dimension filter

Two filters

Filter 1 (Set as context filter) – Only red fruits (inside or outside)

Filter 2 – Top 2 fruits by weight

Calculation

Output of filter 1 – Apple, Strawberry and Watermelon

Output of filter 2 – Watermelon and Apple

Final output – Watermelon and Apple

Different Result

Steps

Filter 1 checks 10 fruits

Filter 2 checks 3 fruits only

Less calculation

Result of filter 2 is final result

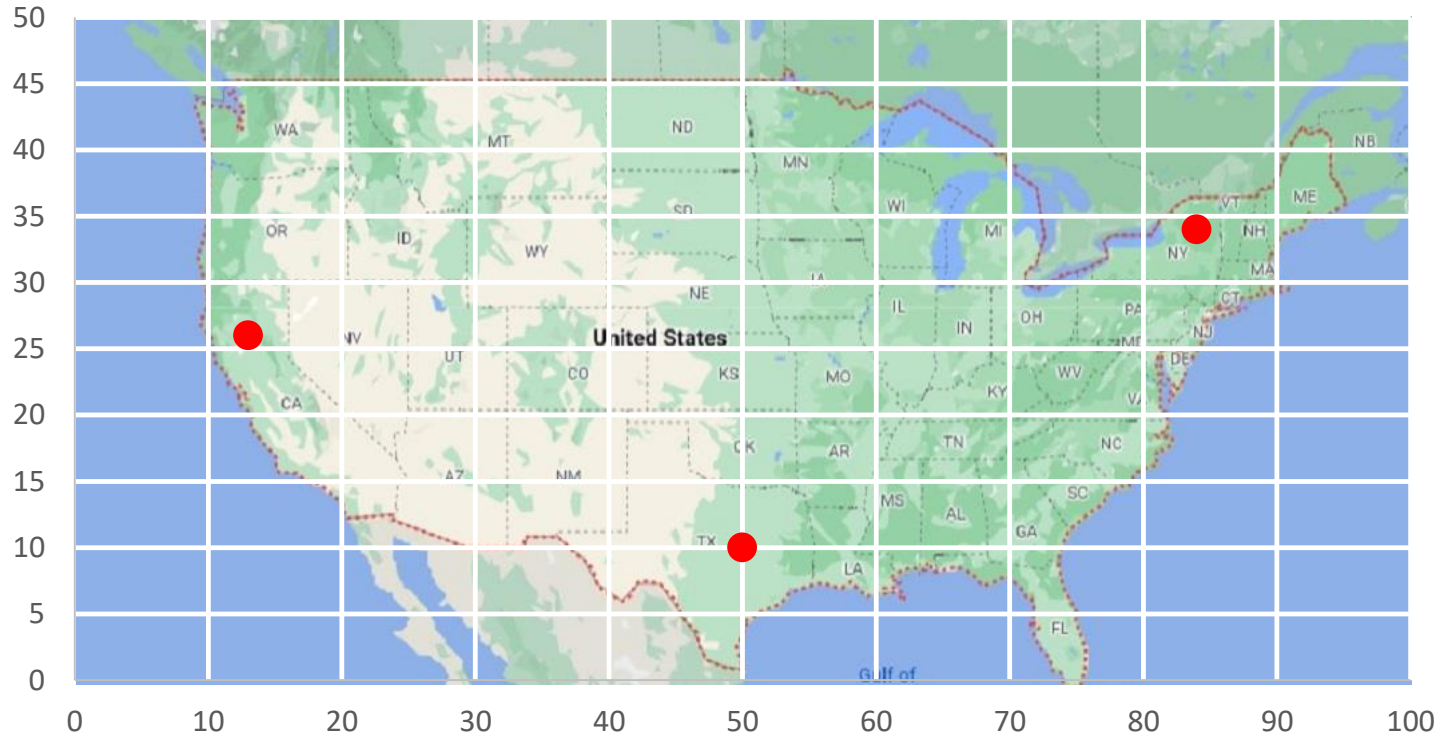
Maps

To plot data points on custom background images

Custom background image

Maps

Custom background image



State	X coordinate	Y coordinate
California	13	26
New York	84	34
Texas	50	10

Maps

Territories

Territory	States
Territory 1	California, Texas, New York
Territory 2	Washington, Pennsylvania, Illinois
Territory 3	All other states

Blending for missing geocoding

Joining

Blending

Relationships

Blending is used when the data tables are present in different data sources

Data Analysis

Functions

Predefined formulas to do a specific calculation

Types of functions:

1. Number functions
2. Date functions
3. Text functions
4. Logical functions
5. Aggregate functions

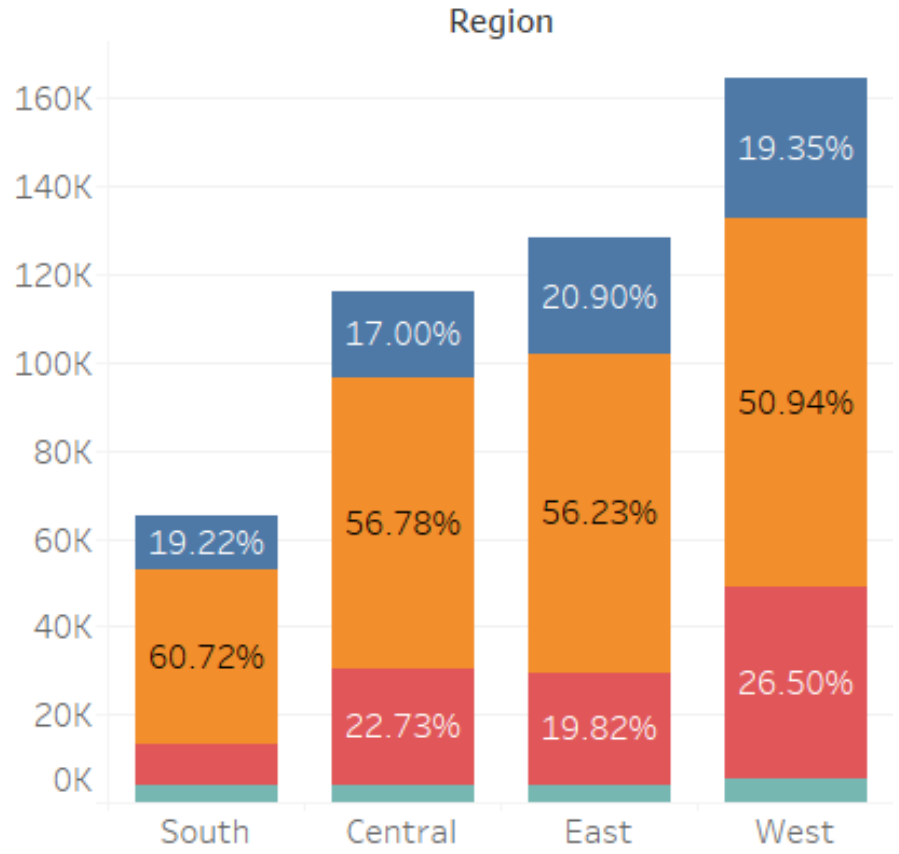
Documentation - https://help.tableau.com/current/pro/desktop/en-us/functions_all_categories.htm

Task – Create the same calculated field (as in last class) using CASE WHEN function

Data Analysis

Table Calculations

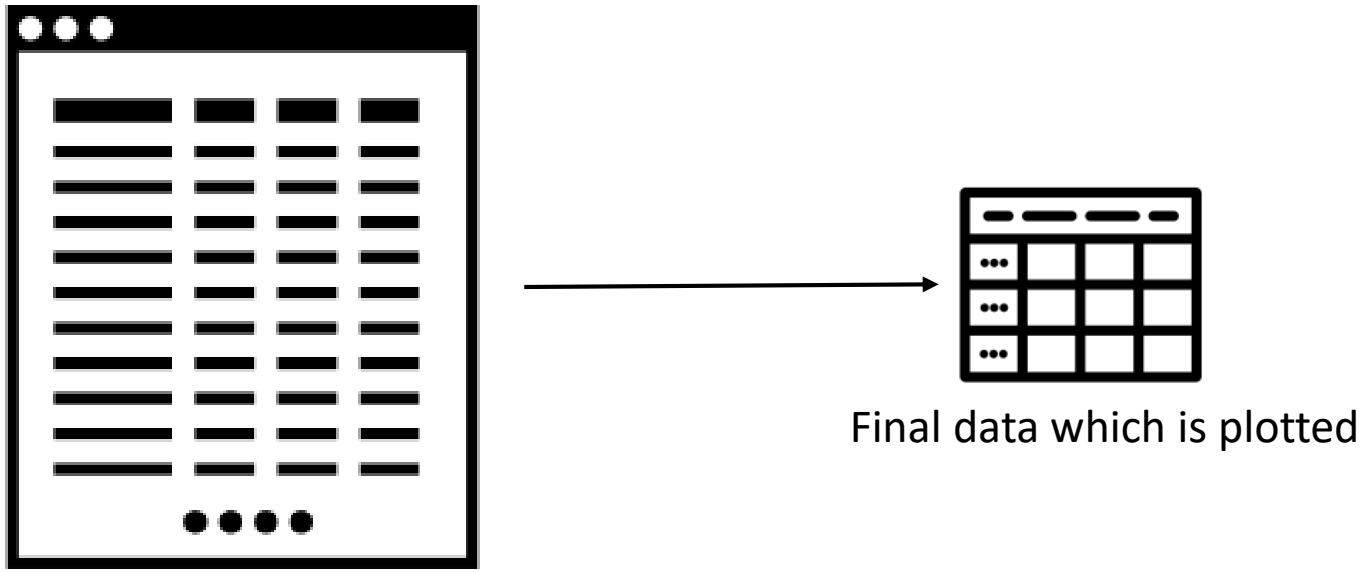
Table calculations – Calculations on plotted data



Data Analysis

Table Calculations

Table calculations only consider the final data plotted for calculations



The complete data

Final data which is plotted

Data Analysis

Table Calculations

Table calculations have two parts – Calculation and Direction

Types of calculations – Differences, percentages, ranks etc.

Direction – Across the table, Down the table, down and across etc.

	Marketing	Finance	HR	Total
Salaries	200	300	100	600
Agency payments	500	100	500	1100
Other expenses	300	100	200	600
Total	1000	500	800	**2300**

Data Analysis

Table Calculations

Table calculations have two parts – Calculation and Direction

	Marketing	Finance	HR	Total
Salaries	200	300	100	600
Agency payments	500	100	500	
Other expenses	300	100	200	
Total	1000	500	800	2300

Across

This total will be used

	Marketing	Finance	HR	Total
Salaries	33%	50%	17%	100%
Agency payments	45%	10%	45%	100%
Other expenses	50%	17%	33%	100%

Data Analysis

Table Calculations

Table calculations have two parts – Calculation and Direction

	Marketing	Finance	HR	Total
Salaries	200	300	100	600
Agency payments	500		500	1100
Other expenses	300	100	200	600
Tot	1000	500	800	**2300**

Down

This total will be used

	Marketing	Finance	HR
Salaries	20%	60%	12.5%
Agency payments	50%	20%	62.5%
Other expenses	30%	20%	25%
Total	100%	100%	100%

Data Analysis

Table Calculations

Table calculations have two parts – Calculation and Direction

	Marketing	Finance	HR	Total
Salaries	200	300	100	600
Agency payments	500	100	500	1100
Other expenses	300	100	200	600
Total	1000	500		2300

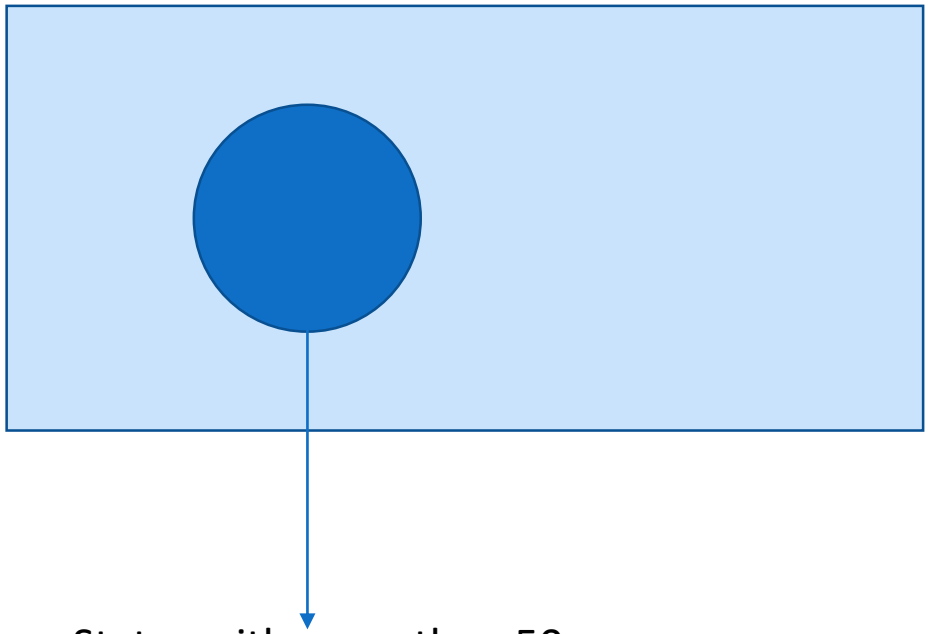
This total will be used

	Marketing	Finance	HR
Salaries	9%	13%	4%
Agency payments	22%	4%	22%
Other expenses	13%	4%	9%

Sets

What

A subset of data based on some conditions



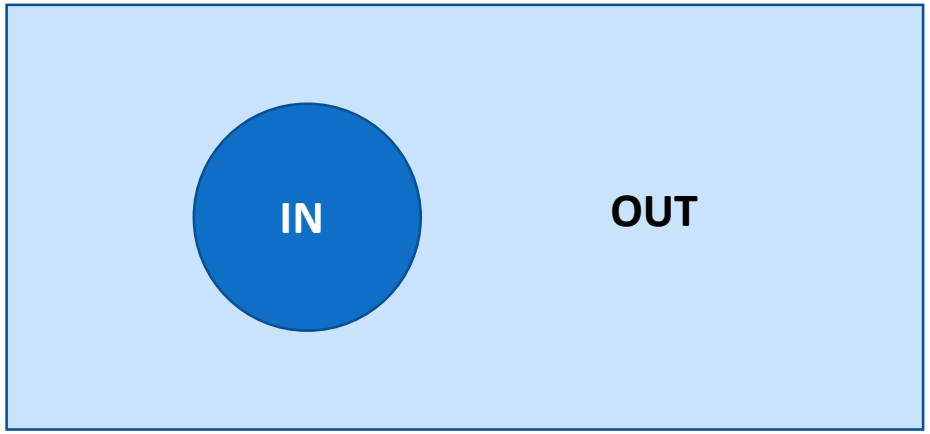
List of all states

States with more than 50 customers/ employees

Sets

Why

Sets are created based on a condition

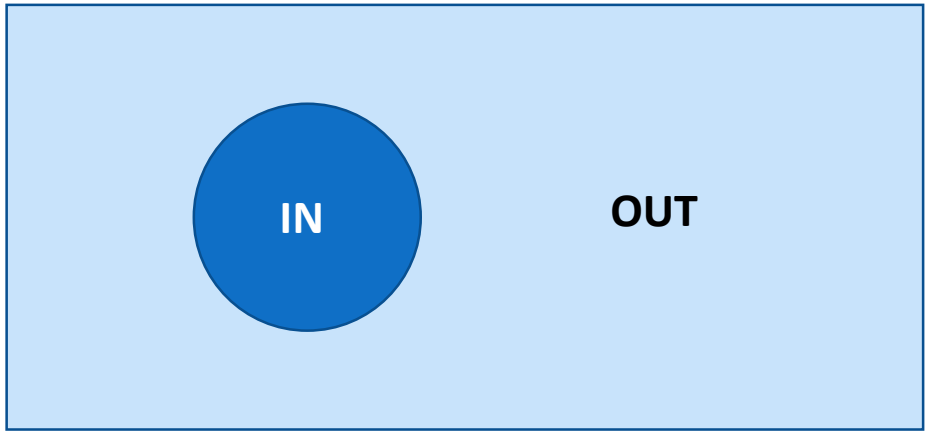


Either select manually or
Specify a condition,
If true – that element is part of Set A or the IN set
If False – that element is part of Set B or the OUT set

Sets

Why

Sets can be used to compare IN vs OUT performance



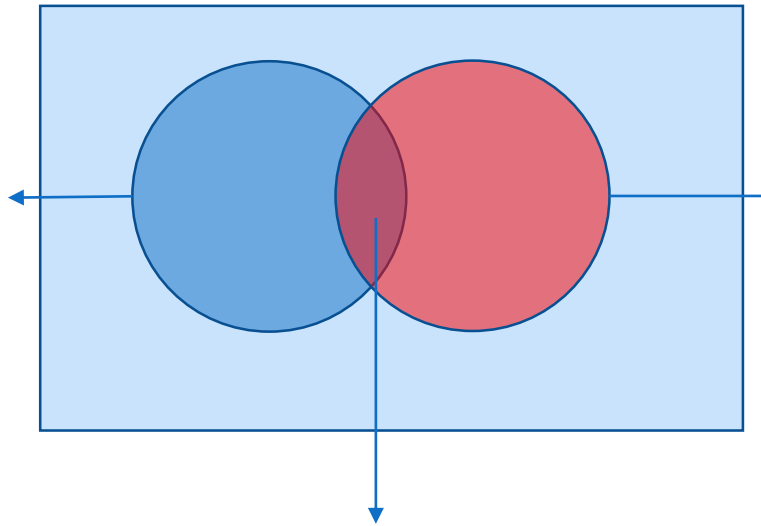
Example: Compare the sum of sales in top 3 states vs all other states

IN – Top 3 states
OUT – All other states

Sets

Why

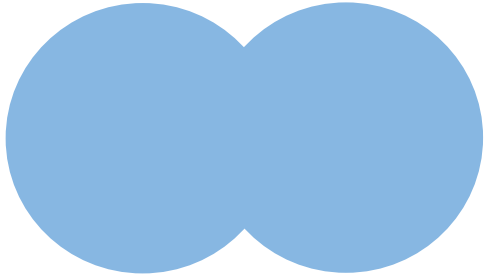
Sets can be used combine sets as per set theory



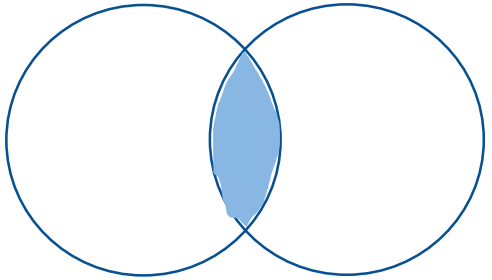
Sets

Why

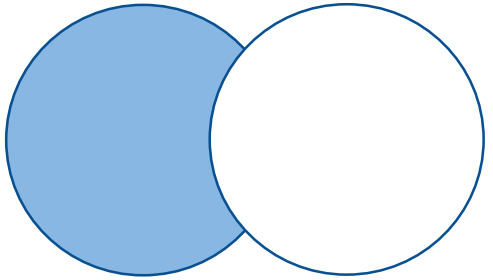
Sets can be used combine sets as per set theory



Union – All members in both sets



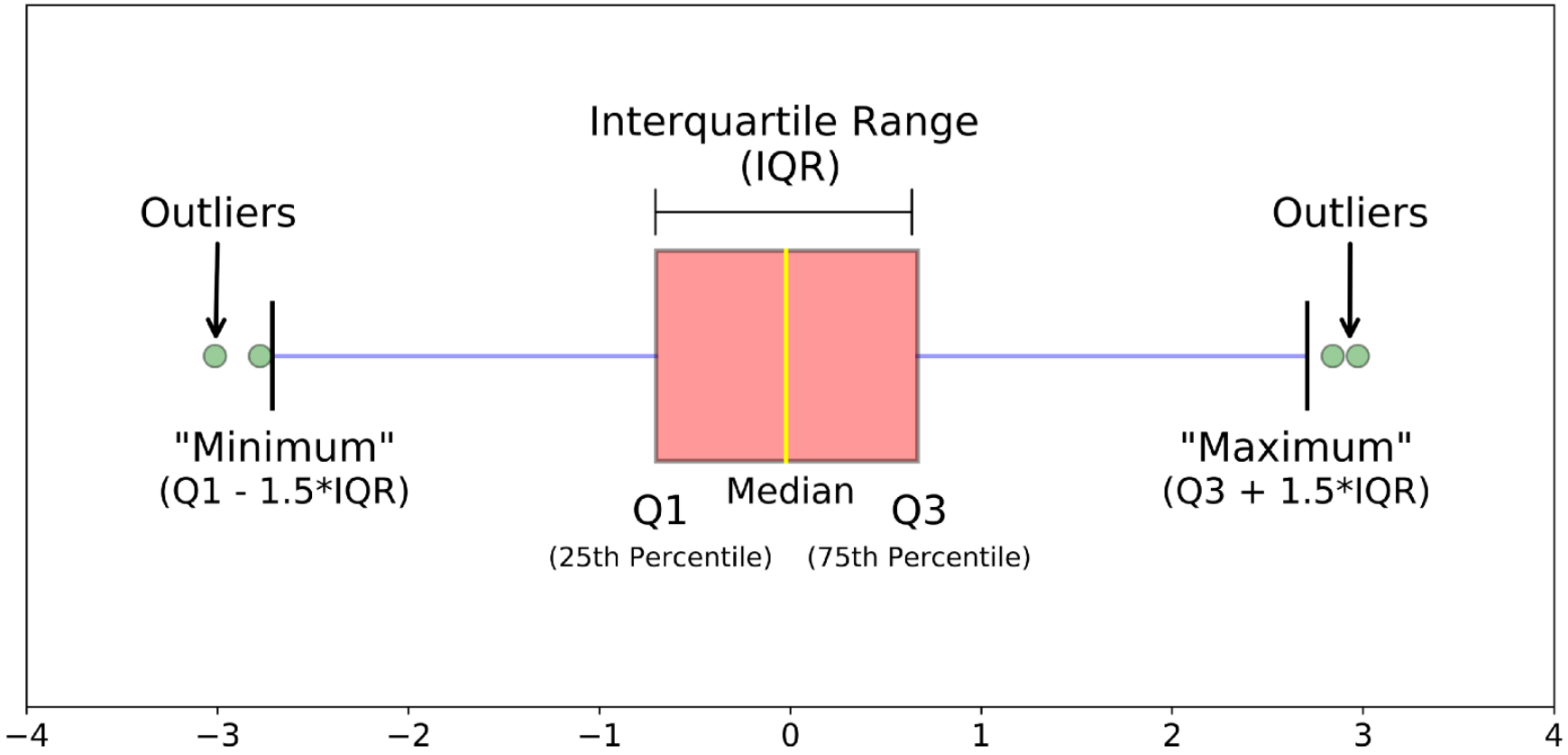
Intersect – Shared members in both sets



Except – members in one set except other set

Box plot

What



Level of Detail

What

Level of Detail is the granularity in data/ how fine is the information

Student	Exam Subject	Institute	Marks Scored
Student 1	Math	A	92
Student 2	Science	A	73
Student 3	English	A	86
Student 4	Math	A	66
Student 5	Science	A	52
Student 6	English	A	100
Student 7	Math	B	86
Student 8	Science	B	51
Student 9	English	B	99
Student 10	Math	B	67
Student 11	Science	B	81
Student 12	English	B	54

Original table – has 3 * 12 = 36 marks
Very high level of detail

Subject	Institute A	Institute B
English	93.0	76.5
Math	79.0	76.5
Science	62.5	66.0

2 * 3 = 6 marks - less level of detail

Subject	Average Marks Scored
English	84.8
Math	77.8
Science	64.3

1 * 3 = 3 marks - lesser level of detail

Average of Marks Scored
75.6

1 mark - least level of detail

LOD expressions

What & Why

Level of Detail expressions help in specifying the level of detail for aggregation of a calculated field

Usually aggregation happens at visualization level

With LOD expressions, we can control the level of detail of the aggregation

Syntax – { LOD keyword Dimension(s) : Aggregate Calculation }

Example: { **FIXED [segment] : SUM([profit]) }**

There are three LOD keywords:

1. FIXED
2. INCLUDE
3. EXCLUDE

LOD expressions

Example

{ FIXED [Student]: SUM([Marks Scored]) }

Student	Exam Subject	Institute	Marks Scored
Student 1	Math	A	92
Student 1	Science	A	73
Student 1	English	A	86
Student 2	Math	A	66
Student 2	Science	A	52
Student 2	English	A	100
Student 3	Math	B	86
Student 3	Science	B	51
Student 3	English	B	99
Student 4	Math	B	67
Student 4	Science	B	81
Student 4	English	B	54

Students	Sum of Marks Scored
Student 1	251
Student 2	218
Student 3	236
Student 4	202

Student	Exam Subject	Institute	Marks Scored	FIXED LOD
Student 1	Math	A	92	251
Student 1	Science	A	73	251
Student 1	English	A	86	251
Student 2	Math	A	66	218
Student 2	Science	A	52	218
Student 2	English	A	100	218
Student 3	Math	B	86	236
Student 3	Science	B	51	236
Student 3	English	B	99	236
Student 4	Math	B	67	202
Student 4	Science	B	81	202
Student 4	English	B	54	202

Compare: AVG(marks scored) on institute dimension vs AVG(FIXED LOD) on institute dimension

LOD expressions

Example

{ Include [Student]: SUM([Marks Scored]) }

Student	Exam Subject	Institute	Marks Scored
Student 1	Math	A	92
Student 1	Science	A	73
Student 1	English	A	86
Student 2	Math	A	66
Student 2	Science	A	52
Student 2	English	A	100
Student 3	Math	B	86
Student 3	Science	B	51
Student 3	English	B	99
Student 4	Math	B	67
Student 4	Science	B	81
Student 4	English	B	54

Students	Sum of Marks Scored
Student 1	251
Student 2	218
Student 3	236
Student 4	202

Student	Exam Subject	Institute	Marks Scored	Include
Student 1	Math	A	92	251
Student 1	Science	A	73	251
Student 1	English	A	86	251
Student 2	Math	A	66	218
Student 2	Science	A	52	218
Student 2	English	A	100	218
Student 3	Math	B	86	236
Student 3	Science	B	51	236
Student 3	English	B	99	236
Student 4	Math	B	67	202
Student 4	Science	B	81	202
Student 4	English	B	54	202

Compare: AVG(marks scored) on institute dimension vs AVG(Include) on institute dimension
Compare: AVG(Fixed LOD) on Subject dimension vs AVG(Include) on subject dimension

LOD expressions

Example

{ Exclude [Student]: SUM([Marks Scored]) }

Student	Exam Subject	Institute	Marks Scored
Student 1	Math	A	92
Student 1	Science	A	73
Student 1	English	A	86
Student 2	Math	A	66
Student 2	Science	A	52
Student 2	English	A	100
Student 3	Math	B	86
Student 3	Science	B	51
Student 3	English	B	99
Student 4	Math	B	67
Student 4	Science	B	81
Student 4	English	B	54

Students	Sum of Marks Scored
Student 1	907
Student 2	907
Student 3	907
Student 4	907

Aggregation depends on Visualization LOD
Student field will be excluded while aggregating

Compare: AVG(marks scored) vs AVG(Include) on institute dimension
Compare: AVG(Fixed LOD) on Subject dimension vs AVG(Include) on subject dimension