Topics you will find in this deck

- Analyze Data set and Import to Tableau Public
- Slicing, Dicing, Splitting with Live examples
- Methods of combining data with Live examples
- Joins and Unions Live examples
- Cleaning Data duplicates, Nulls and more Theory & Live examples
- Enrich Data
- Implement Business calculations with Live examples
- Aggregation with Live examples
- Ranking by state, product
- Hierarchies With level examples
- What is cube?

You are free to

- Share — copy and redistribute the material in any medium or format
- Adapt — remix, transform, and build upon the material for any purpose, even commercially.

Under the following terms*

Attribution — You must give appropriate credit, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the Author endorses you or your use

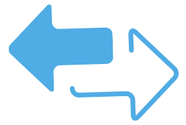
Please refer for the full terms: *https://creativecommons.org/licenses/by/4.0/*

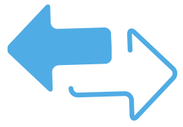
*The author does not provide any guarantees, warranties for the content and accuracy. The content may or may not be updated. By using this material you indemnify the author against any liabilities including copyrights and/or damages of any sort.

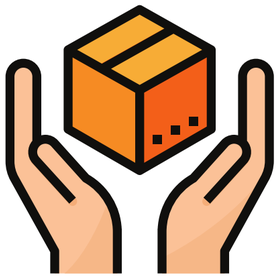
Analyze the business questions to be answered – Scenario

- Actual
- Forecast
- Actual vs Forecast

- **The total sales/profit**
- **Top 10 customers**
- **New customers**
- **Returns/attrition**
- **Forecasts vs actual**
- **Recovery rates**
- **Top issues**







Descriptive, diagnostic, predictive, prescriptive

In a typical project what happens is major functionalities are prioritized, captured and developed as use cases. These use cases are further divided into different requirements

Product Vision

BRD
Business Requirement Document

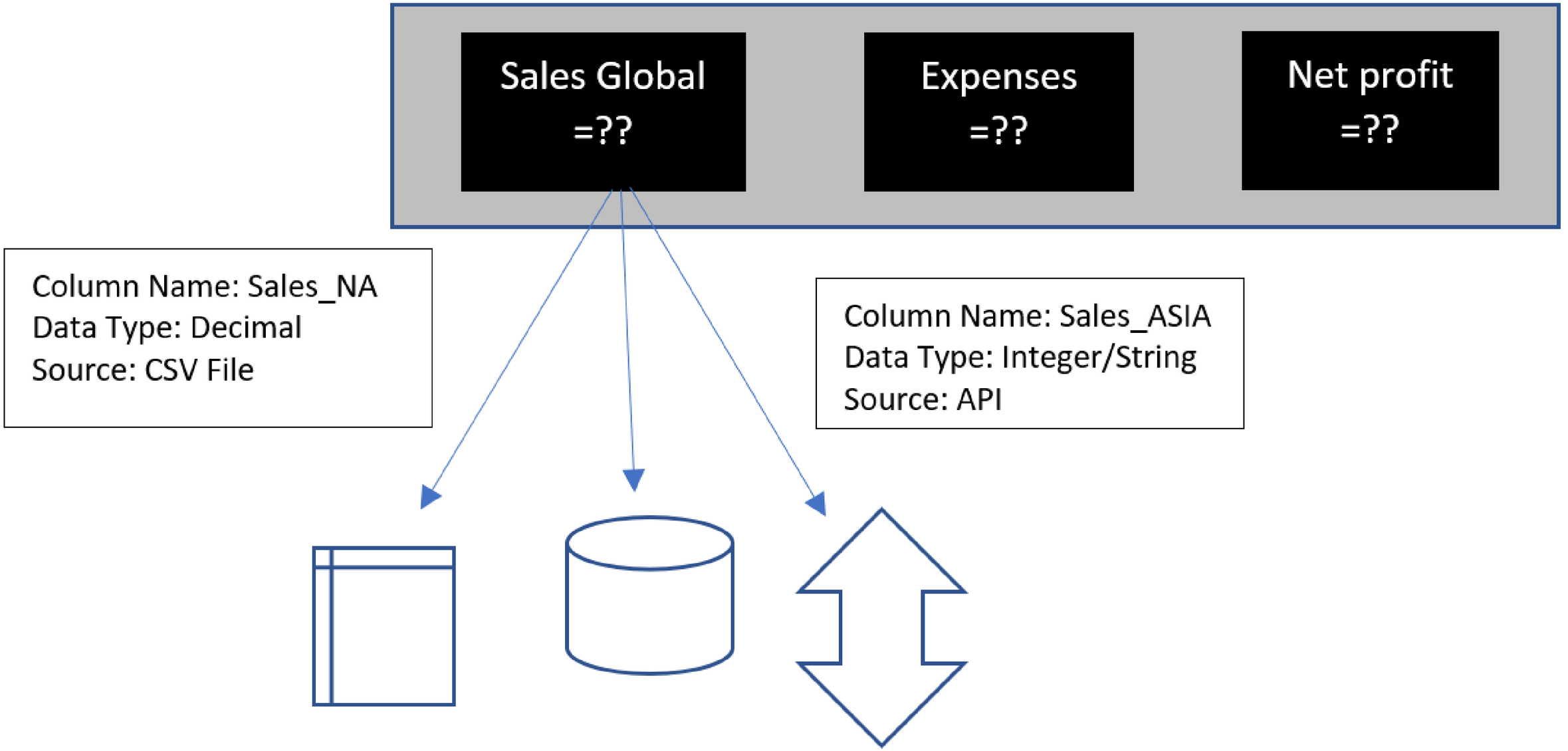
FRD
Functional Requirement Document

TRS
Technical Requirement Specification

UI design
Use cases
Test Cases

Data Mapping
Data Quality
Data Privacy/policy
Up /Down stream
Data Documents

Data Mapping



Identify data sources and data owners

from the data perspective not UI or others.

Data Analysis (Data Characteristics)

- Format
- Velocity
- Volume
- Method of extraction
- Residency
- Access Privileges
- Source system
- Temperature

Format

Structured

Unstructured

Velocity

How fast the data is captured?

The speed of transaction?

The speed of updates to dimensions?

Transactions per second?

Frequency

How often the data is captured and or transported?

Transactions happen every sub second or every hour?

Volume

Data is captured in MB, GB, TB ?

Volume will need a time dimension w.r.t the time dimension requirements of the report

Method of extraction

Simple import for files in csv or excel format?

Import via a JDBC driver connected to a Data base?

Import via ETL tools?

Import via Replication tools?

Import via API?

Import via a Cloud interface?

Residency

For how long do you have access to data before it is moved out

of the required systems or memory?

Data movement frequency between Transaction systems
and the layer you will using to build reports?

Access

one of the most important questions

Who can see and access what data is crucial

Authentication and authorization

RBAC, fine grained, coarse grained

Source systems

- CSV files
- Database
- No Sql Database
- Unstructured database
- API
- Cloud
- ERP
- Applications
- IoT devices
- Physical assets

Temperature

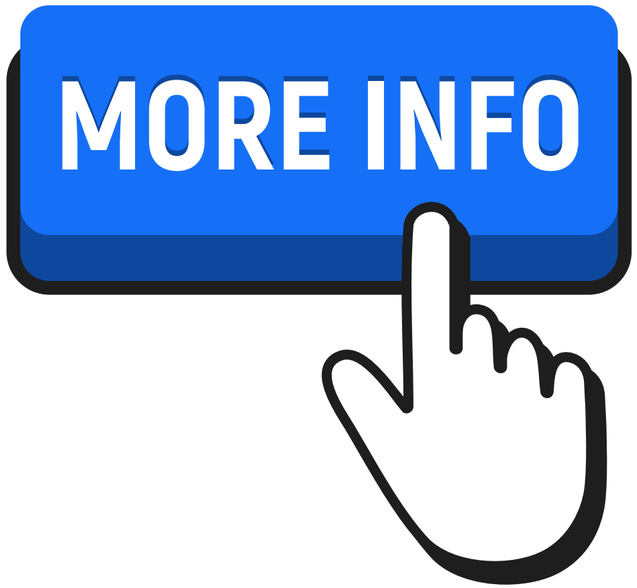
Hot, warm, cold ?

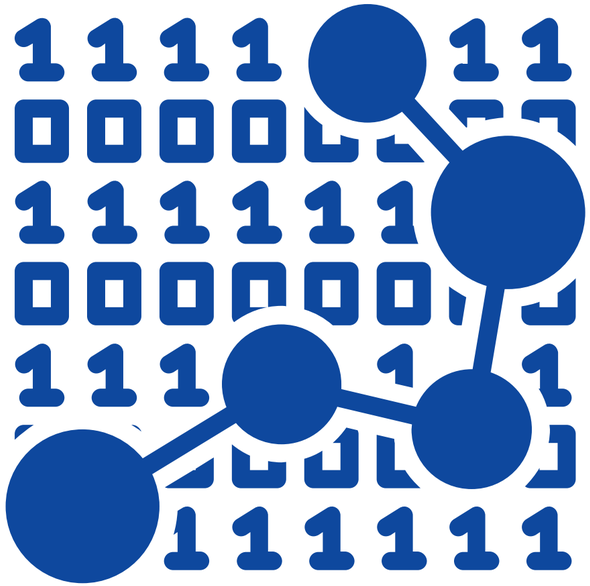
which and how do you differentiate between the most used data

and aged data?

Ask for more information

Sample data set





Gap Analysis

- Data quality
- Data completeness
- Business context
- Security, authentication and authorization
- Business logic required to be applied on raw columns
- Identify the time period required and get data only for that period
- Algorithms which can used to identify patterns and predictions

Next steps

Either this data is already sitting in your data warehouse or data store or data lake or you have to fetch the complete data or a part of the data directly from the source systems.

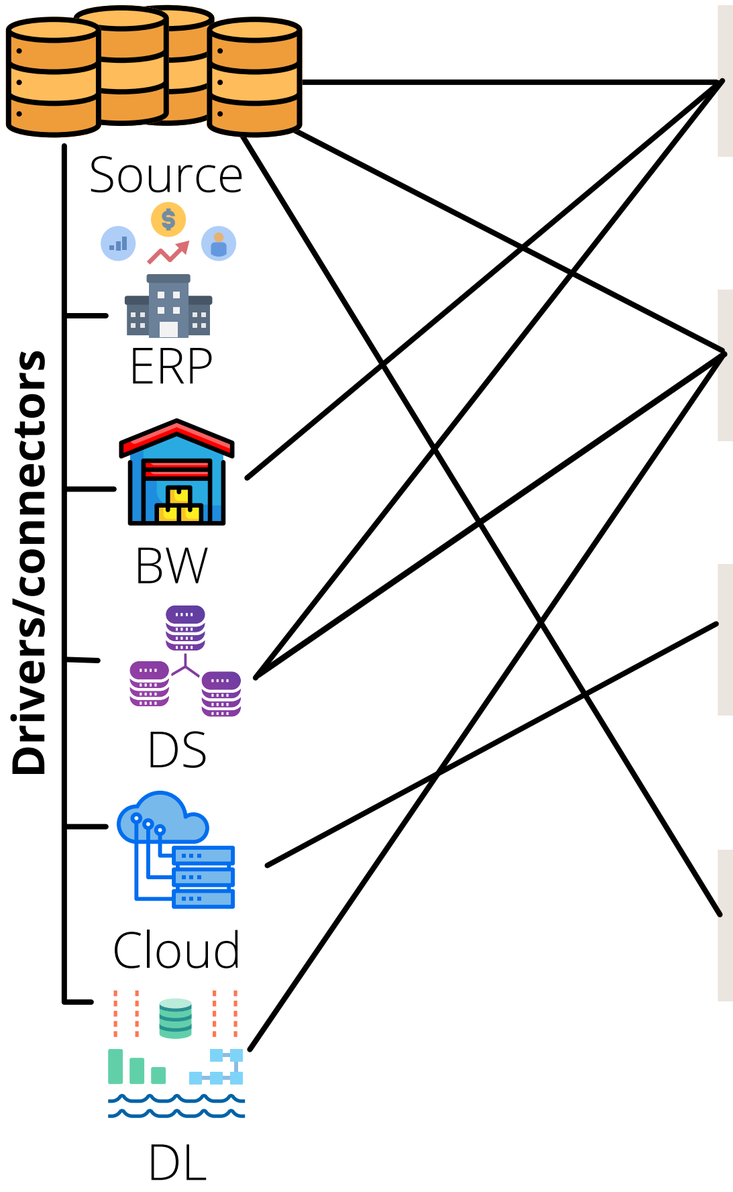
Now depending on the current architecture, tool and skills capability, the loading and transformation of data can be done by separate team or a consultant.

Next steps

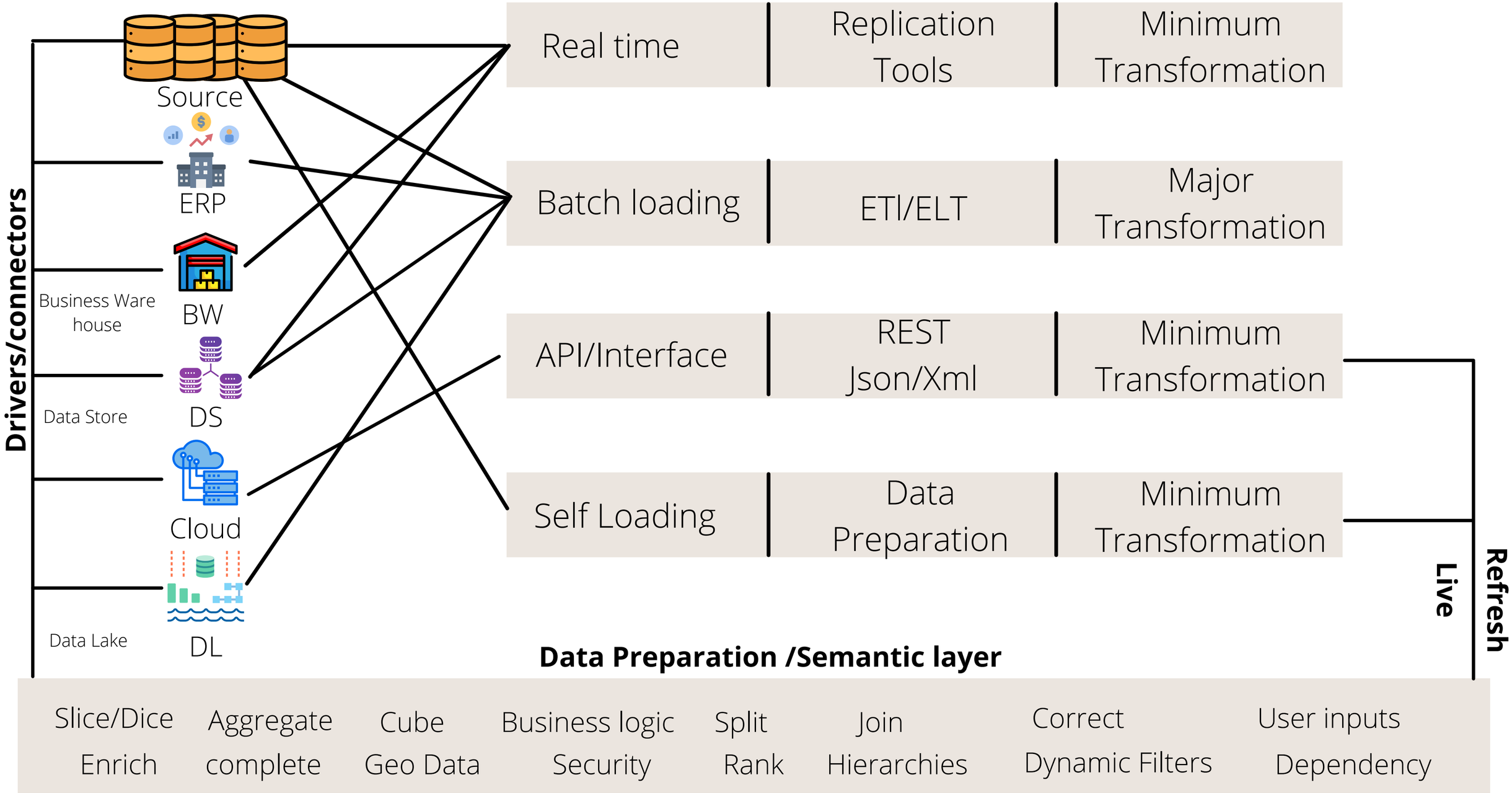
In most cases you will be able to get a majority of this data from a warehouse and the rest you will have to talk directly to the source systems, cloud or import csv, excel files.

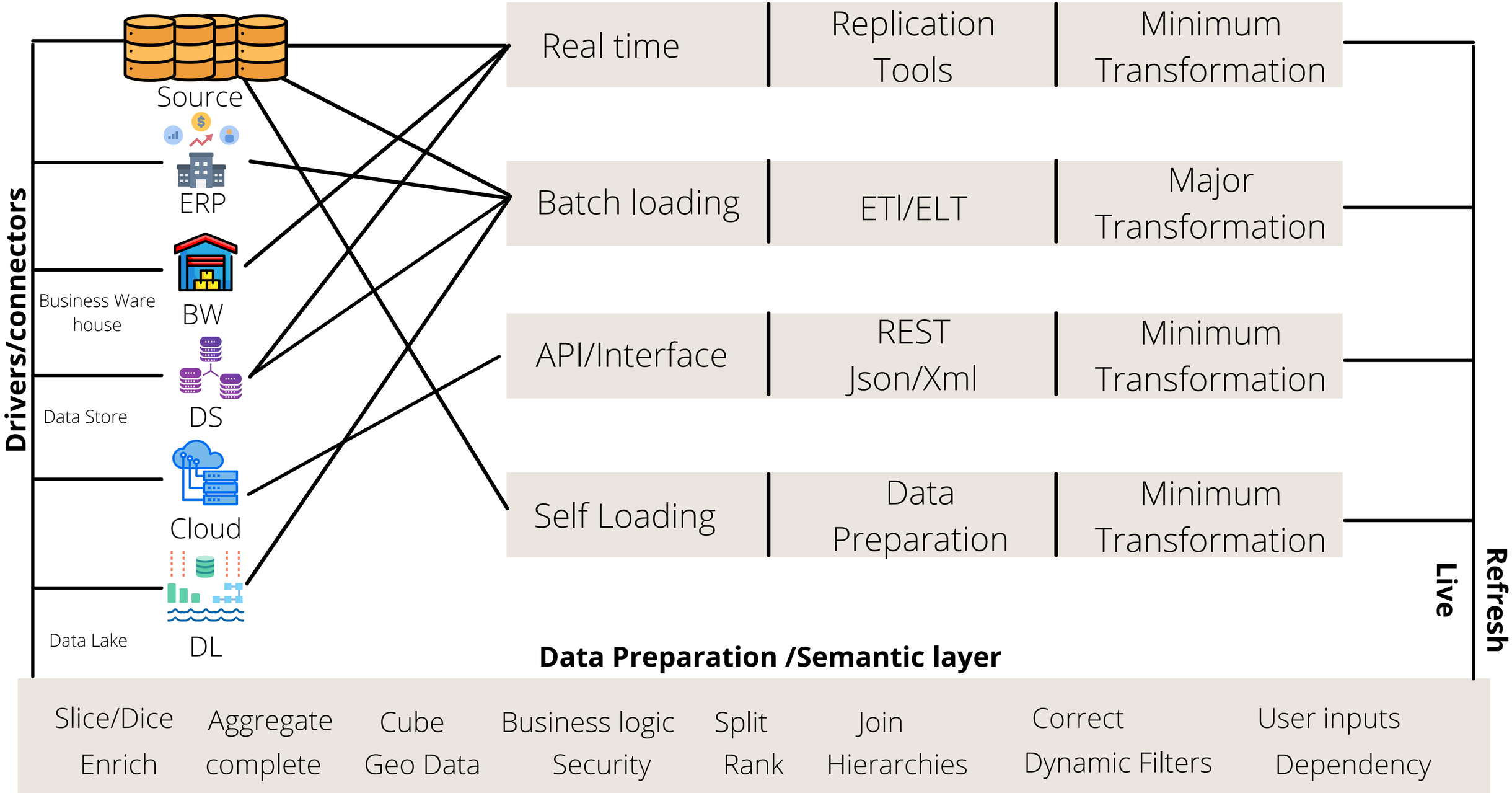
Depending on the level of DL and expertise you may involve yourself in simple importing of an csv file or more complicated tasks.

2



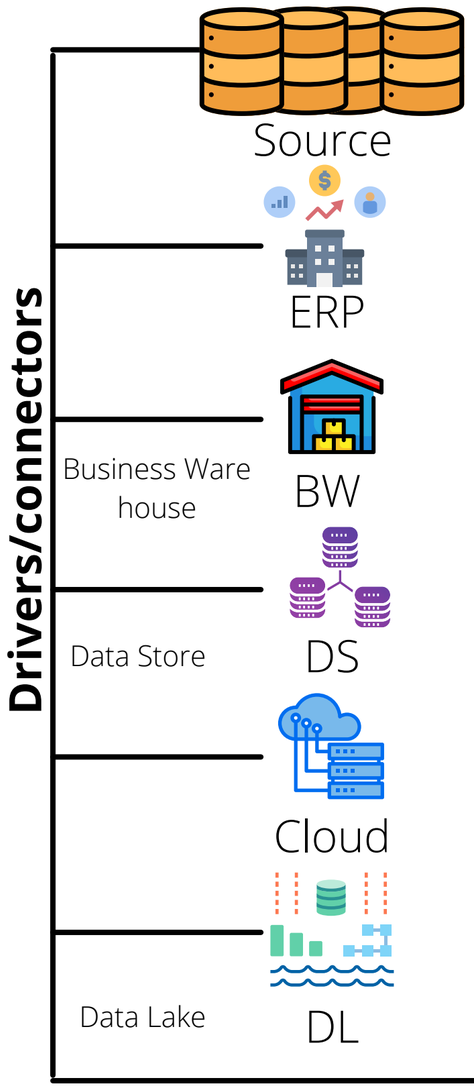
Real time	Replication Tools	Minimum Transformation
Batch loading	ETI/ELT	Major Transformation
API/Interface	REST Json/Xml	Minimum Transformation
Self Loading	Data Preparation	Minimum Transformation


Drivers/connectors
Source
ERP
Business Ware house
BW
Data Store
DS
Cloud
Data Lake
DL
Real time
Replication Tools
Minimum Transformation
Batch loading
ETl/ELT
Major Transformation
API/Interface
REST Json/Xml
Minimum Transformation
Self Loading
Data Preparation
Minimum Transformation
Live
Refresh
Data Preparation /Semantic layer
Slice/Dice
Enrich
Aggregate
complete
Cube
Geo Data
Business logic
Security
Split
Rank
Join
Hierarchies
Correct
Dynamic Filters
User inputs
Dependency


Drivers/connectors
Source
ERP
Business Ware house
BW
Data Store
DS
Cloud
Data Lake
DL
Real time
Replication Tools
Minimum Transformation
Batch loading
ETl/ELT
Major Transformation
API/Interface
REST Json/Xml
Minimum Transformation
Self Loading
Data Preparation
Minimum Transformation
Live
Refresh
Data Preparation /Semantic layer
Slice/Dice
Enrich
Aggregate
complete
Cube
Geo Data
Business logic
Security
Split
Rank
Join
Hierarchies
Correct
Dynamic Filters
User inputs
Dependency

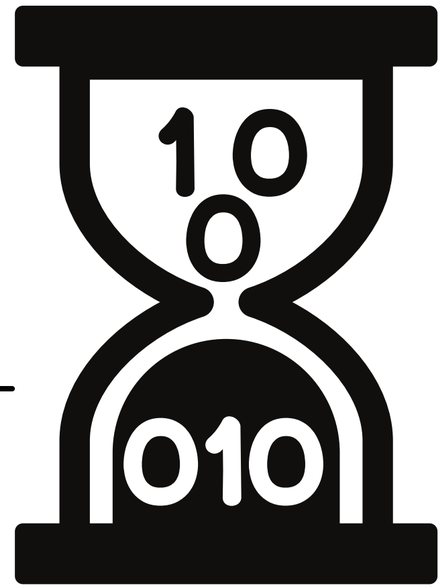
To make connection

- Driver
- Host
- Port
- Url
- Schema
- Cube
- Password
- Sid



- Driver
- Host
- Port
- Url
- Schema
- Cube
- Password
- Sid

Data Preparation /Semantic layer



3

- **Analyze the data**
- **Import data on Tableau Public**

Slice data

Slice data to match your requirement

Sales	Year	Product
1000000	2016	Electronics
2000000	2016	House hold
1500000	2017	Electronics
2100000	2017	House hold

Sales	Year	Product
1000000	2016	Electronics
2000000	2016	House hold
1500000	2017	Electronics
2100000	2017	House hold

Sales	Year	Product
1000000	2016	Electronics
2000000	2017	Electronics
1500000	2016	House hold
2100000	2017	House hold

Live example
Tableau

4

Dice data

Reduce/filter the dataset to answer your specific question

All well designed report which is easy to consume and it effective. No more than 7-10 pieces of information are presented

Market C	Close Pric	PE Ratio	1M Return vs	1W Retur	Daily Volu	RSI â€“ 1	RSI Exponential â	ADX Rating â€“ Trend	50D SMA	1D Return
29890.1	1203.1	102.518	-1.566183	2.22619	235454	47.7395	63.11521444	24.95740916	1173.37	-0.09135
29739	817.9	40.2259	10.142846	17.1441	2585800	73.105	80.86283612	30.96148721	703.318	7.25854
29035.2	160.05	10.0447	-5.3020993	3.09179	16465948	53.4301	60.98066594	25.06146848	153.56	-0.40448
29031.2	1047.3	45.7162	9.23239701	4.79288	579889	63.5256	72.84908802	46.71131923	939.054	2.081
28330.3	2571.25	24.3162	-9.3595438	-1.10577	13956	57.061	53.87526933	36.72157666	2569.97	0.77208
28044	45.15	58.0417	-15.249727	3.91254	12844249	51.2	56.24844186	51.13253158	47.0191	-1.09529
27708	2183.45	78.7539	-4.9590307	7.79807	120495	73.3224	73.40486499	44.38148998	2108.4	-0.02061
27621.3	6037.75	55.508	-3.0488428	5.3635	35232	67.9089	73.74378186	25.82144633	5810.35	0.04971
27595.2	24496.3	169.932	-6.8157433	6.6023	1442	56.4889	61.35504212	34.19806629	24071	-0.16628
27560.4	11	-7.89941	-19.018019	0.91743	58976619	33.3333	37.57571289	30.02707523	11.9676	-0.9009
27538.1	9307.5	41.9942	-4.5892214	1.68464	25577	61.5375	68.60941758	33.73033705	9098.5	0.96819
27495.9	3558.5	38.7425	-1.0535144	13.1083	62129	76.182	80.79532345	40.49037953	3221.47	1.67579
27473.2	27.35	8.49676	-1.8257115	2.43446	5534062	70.9091	61.7789002	43.44524716	26.3441	-1.26354
27221.9	8920.2	97.3045	14.2808834	14.5539	782161	85.744	88.15833851	37.76399182	7338.27	9.61843
27178.7	2139.6	27.7861	-5.9098747	-0.04205	73020	63.2699	61.26235288	31.37802533	2081.13	-0.69388
26646.4	599.45	58.6021	-1.1174804	-1.43867	131833	62.5156	56.79058462	28.73338282	573.862	-1.02369
26344.5	3531.7	58.455	6.23557235	6.26928	304948	90.9679	90.1545516	36.9956849	3164.58	2.62542
26335	159.8	16.0534	-4.6149433	-4.45441	400601	60.7692	51.83883093	39.09965581	157.238	-0.59098

Market C	Close Pric	PE Ratio	1M Return vs	1W Retur	Daily Volu	l â€“ 1	RSI Exponential â	ADX Rating â€“ Trend	50D SMA	1D Return
29890.1	1203.1	102.518	-1.566183	2.22619	235454	7.7395	63.11521444	24.95740916	1173.37	-0.09135
29739	817.9	40.2259	10.142846	17.1441	2585800	73.105	80.86283612	30.96148721	703.318	7.25854
29035.2	160.05	10.0447	-5.3020993	3.09179	16465948	3.4301	60.98066594	25.06146848	153.56	-0.40448
29031.2	1047.3	45.7162	9.23239701	4.79288	579889	3.5256	72.84908802	46.71131923	939.054	2.081
28330.3	2571.25	24.3162	-9.3595438	-1.10577	13956	57.061	53.87526933	36.72157666	2569.97	0.77208
28044	45.15	58.0417	-15.249727	3.91254	12844249	51.2	56.24844186	51.13253158	47.0191	-1.09529
27708	2183.45	78.7539	-4.9590307	7.79807	120495	3.3224	73.40486499	44.38148998	2108.4	-0.02061
27621.3	6037.75	55.508	-3.0488428	5.3635	35232	7.9089	73.74378186	25.82144633	5810.35	0.04971
27595.2	24496.3	169.932	-6.8157433	6.6023	1442	6.4889	61.35504212	34.19806629	24071	-0.16628
27560.4	11	-7.89941	-19.018019	0.91743	58976619	3.3333	37.57571289	30.02707523	11.9676	-0.9009
27538.1	9307.5	41.9942	-4.5892214	1.68464	25577	1.5375	68.60941758	33.73033705	9098.5	0.96819
27495.9	3558.5	38.7425	-1.0535144	13.1083	62129	76.182	80.79532345	40.49037953	3221.47	1.67579
27473.2	27.35	8.49676	-1.8257115	2.43446	5534062	0.9091	61.7789002	43.44524716	26.3441	-1.26354
27221.9	8920.2	97.3045	14.2808834	14.5539	782161	35.744	88.15833851	37.76399182	7338.27	9.61843
27178.7	2139.6	27.7861	-5.9098747	-0.04205	73020	3.2699	61.26235288	31.37802533	2081.13	-0.69388
26646.4	599.45	58.6021	-1.1174804	-1.43867	131833	2.5156	56.79058462	28.73338282	573.862	-1.02369
26344.5	3531.7	58.455	6.23557235	6.26928	304948	0.9679	90.1545516	36.9956849	3164.58	2.62542
26335	159.8	16.0534	-4.6149433	-4.45441	400601	0.7692	51.83883093	39.09965581	157.238	-0.59098

So when you import data you may have to import all the data in the specific table or cube.

Because it is very cumbersome and actually not practical to cater exactly to all the reporting requirements.

That happens in the preparation layer.

So at the most from on one table you may need around 10-12 columns.

Most tables have more than that and many have more than 100 columns.

So by dicing the data you are chopping vertically and removing all the unnecessary columns.

Many reasons you don’t need them and if by chance you leave it to the adhoc report builder they may unknowingly bring all the columns putting burden on the servers.

Live example.

5

Split

Split: This happens when you want separate pieces of the data set for different purposes on the report.

The goals is efficiency, loading times and sometime just user requirements.

You could split the useful data vertically or horizontally.

slicing of information.
The entire company name could be in one column whereas you only need the main name and move the remaining information to another column.

Combine

Combine

This happens when the data comes from

- 2 different systems,
- or sources,
- or universes,
- or domains.

Joins - Extend the data with more columns . Generally the data sets will have a common unique column.

Union - Sandwich 2 different data sets with the same structure.

Normalized data model

Dimension- Product
Primary key
Name
Description
Brand
Channel

Dimension

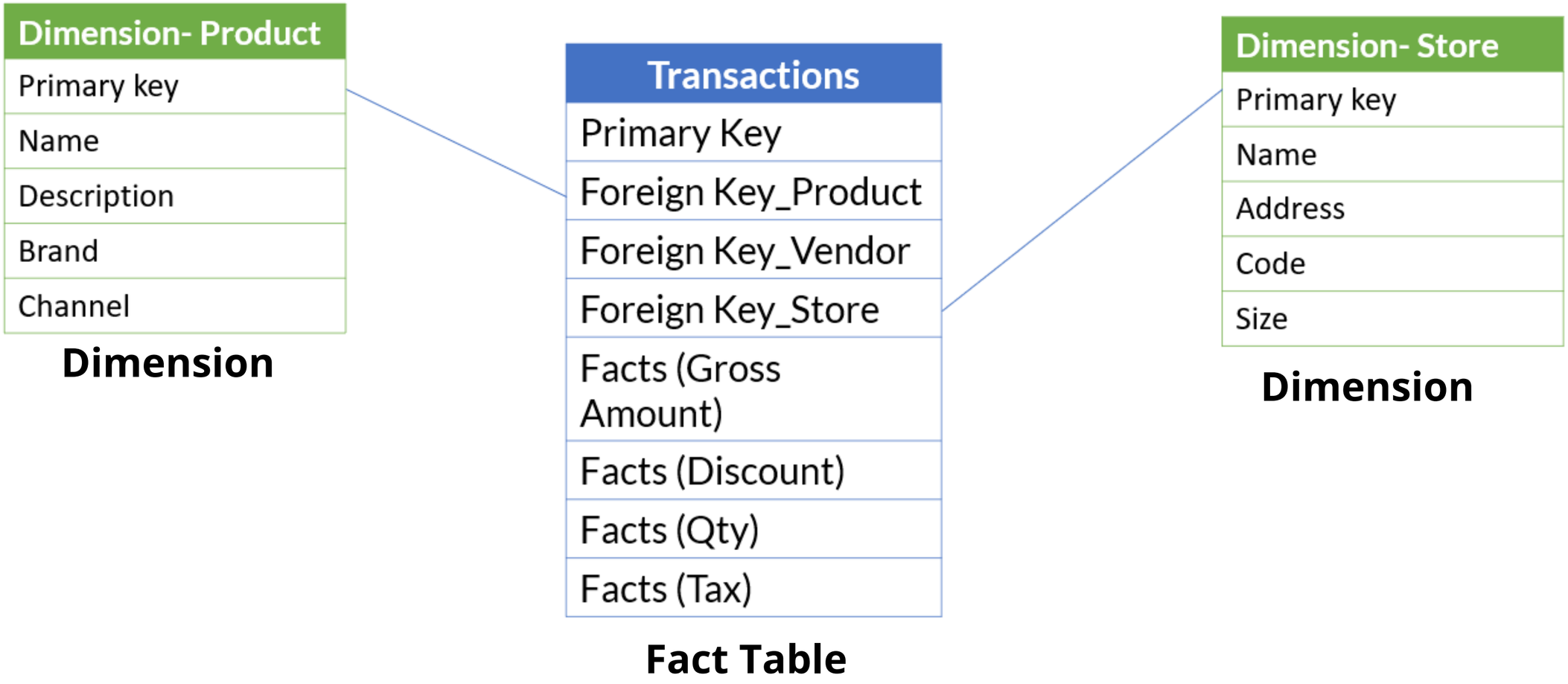
Transactions
Primary Key
Foreign Key_Product
Foreign Key_Vendor
Foreign Key_Store
Facts (Gross Amount)
Facts (Discount)
Facts (Qty)
Facts (Tax)

Fact Table

Dimension- Store
Primary key
Name
Address
Code
Size

Dimension

joins


Dimension- Product
Primary key
Name
Description
Brand
Channel
Dimension
Transactions
Primary Key
Foreign Key_Product
Foreign Key_Vendor
Foreign Key_Store
Facts (Gross Amount)
Facts (Discount)
Facts (Qty)
Facts (Tax)
Fact Table
Dimension- Store
Primary key
Name
Address
Code
Size
Dimension

Union

Sandwich 2 different data sets with the same structure

Another example for a combine is sandwich of 2 similar data sets belonging to two different domains or times(Realtime+ historic) using unions

You combine the data and create a single point of access for the consumption

Sales for 2014

SALESORDERID	HISTORY_(	HISTORY_CREATEDAT	HISTORY_(	HISTORY_CHANGEDAT	PARTNER_PARTNERID	LIFECYCLESTATUS	BILLINGSTATUS
500001000	5	06-09-2014	15	20-09-2014	100000041	X	I
500001000	5	06-09-2014	15	20-09-2014	100000041	X	I
500001001	24	07-09-2014	19	21-09-2014	100000037	X	I
500001001	24	07-09-2014	19	21-09-2014	100000037	X	I
500001002	5	08-09-2014	27	22-09-2014	100000043	X	I
500001002	5	08-09-2014	27	22-09-2014	100000043	X	I
500001003	24	09-09-2014	33	23-09-2014	100000044	X	I
500001003	24	09-09-2014	33	23-09-2014	100000044	X	I
500001004	2	10-09-2014	19	24-09-2014	100000036	N	I
500001004	2	10-09-2014	19	24-09-2014	100000036	N	I

Sales for 2015

SALESORDERID	HISTORY_(	HISTORY_CREATEDAT	HISTORY_(	HISTORY_CHANGEDAT	PARTNER_PARTNERID	LIFECYCLE	BILLINGST
500000000	24	01-01-2015	1	15-01-2015	100000000	X	I
500000000	24	01-01-2015	1	15-01-2015	100000000	X	I
500000001	32	02-01-2015	20	16-01-2015	100000002	N	I
500000001	32	02-01-2015	20	16-01-2015	100000002	N	I
500000002	5	03-01-2015	6	17-01-2015	100000005	P	I
500000002	5	03-01-2015	6	17-01-2015	100000005	P	I
500000003	33	04-01-2015	19	18-01-2015	100000006	N	I
500000003	33	04-01-2015	19	18-01-2015	100000006	N	I
500000004	22	05-01-2015	28	19-01-2015	100000006	C	P

Sales for 2014

SALESORDERID	HISTORY_(	HISTORY_CREATEDAT	HISTORY_(	HISTORY_CHANGEDAT	PARTNER_PARTNERID	LIFECYCLESTATUS	BILLINGSTATUS
500001000	5	06-09-2014	15	20-09-2014	100000041	X	I
500001000	5	06-09-2014	15	20-09-2014	100000041	X	I
500001001	24	07-09-2014	19	21-09-2014	100000037	X	I
500001001	24	07-09-2014	19	21-09-2014	100000037	X	I
500001002	5	08-09-2014	27	22-09-2014	100000043	X	I
500001002	5	08-09-2014	27	22-09-2014	100000043	X	I
500001003	24	09-09-2014	33	23-09-2014	100000044	X	I
500001003	24	09-09-2014	33	23-09-2014	100000044	X	I
500001004	2	10-09-2014	19	24-09-2014	100000036	N	I
500001004	2	10-09-2014	19	24-09-2014	100000036	N	I

No joins!

SALESORDERID	HISTORY_(	HISTORY_CREATEDAT	HISTORY_(	HISTORY_CHANGEDAT	PARTNER_PARTNERID	LIFECYCLE	BILLINGST
500000000	24	01-01-2015	1	15-01-2015	100000000	X	I
500000000	24	01-01-2015	1	15-01-2015	100000000	X	I
500000001	32	02-01-2015	20	16-01-2015	100000002	N	I
500000001	32	02-01-2015	20	16-01-2015	100000002	N	I
500000002	5	03-01-2015	6	17-01-2015	100000005	P	I
500000002	5	03-01-2015	6	17-01-2015	100000005	P	I
500000003	33	04-01-2015	19	18-01-2015	100000006	N	I
500000003	33	04-01-2015	19	18-01-2015	100000006	N	I
500000004	22	05-01-2015	28	19-01-2015	100000006	C	P

Sales for 2015

Live example.

6

Clean/Correct/Complete

Now source data seldom comes clean and complete.

Even though it may go through various value addition layers it still may not be in a format you need it to be.

Data values carry a Null Value

Data has un-standard values for the same information

Data has Duplicates

Data has records have incomplete information

Ways to Correct it

- Use tool in built Functions
- Distinct
- Replace
- Filter
- Conditional Logic
- Join/Extend
- ETL Transforms
- Address Transforms
- Data Quality Transfroms

Method to solve it.

- You might have to go back on layer down and ask them to combine this information. They mostly will resort to making a join.

- You can do this in the preparation layer yourself. Some advanced features of text analysis and address standardization may not be out of the box.

- You can always ask the vendor or your colleagues!

Live example.
Efashion

Enrich

This is done to add more business meaning and increase the operability on the data set.

The idea is to add additional information to the data set which will help the business understand this data better and also be able to perform advanced visualizations like geo maps.

For example, if you have a column with only company abbreviations then you can add an additional column from your underlying data sources which have the column which contains the fully qualified name of the company.

There are various formats based on the geographic location of the company.

Is for example a column contains a time stamp. It data and time combined in one column.

Each of these columns will hold specific information like. Date , year, qtr, week, hours along with the original information.
Which will allow to filter and use logic on these columns effectively.

Business Logic
Calculated Column

Depending on your DL level you could do it in the data preparation layer on even in the underlying layers or even in the reporting tools directly

- Calculation
- Strings Operation
- Concatenation
- Conditional Logic
- Other complex functions

- Profit
- EBITA
- Variance
- Breakeven
- Cost of goods
- ROI
- Wages
- Taxes
- Annuities
- Marketing Metrics
- Mortgage, interest Rates
- Exchange Rates
- Term of payment
- Insurance, Healthcare, education, Technology, Electronics

Total Sales = Unit Price*quantity

Aggregation

Store	Type	Qty	Price/Piece	Total sale Price
NYC	T-Shirt	3	20	60
SFO	Pants	1	25	25
NYC	Pants	2	30	60
SFO	Scarf	2	15	30
SFO	Jacket	1	60	60
NYC	Overcoat	1	250	250
NYC	Shoes	2	125	300
		SUM = 12	AVG :75	SUM = 785

End users want to see the summary information

So for example when we day we want to see total sale by the quarter then you have to group all the transaction in the qtr and summarize them.

When we say summarize mathematically it is just an addition of all the sales in that quarter

This is called aggregation or a group by function.

The higher the grouping the denser the information but less in number.

Group by year – summarized to just one tuple/cell value
The lower the grouping the less dense the information and more in number.

Group by week – There 52 weeks in a year. So, 52 cell values.

Aggregation Types

- Sum
- Min
- Max
- Count
- Standard Deviation
- Variance
- Average

Ranking

You need to rank your customers, vendors, problems, risks etc

- Top 10 customers
- Top 5 countries/counties
- Top issues
- Top 5-time windows.

Ranking can be done in a coarse and granular way .

There are also various situation which you need to handle to satisfy customer requirements.

After decades of customers bugging the vendors.

Most enterprise tools come with ranking features and they also have inbuilt features to handle subtle configurations and changes which are required by most customers.

Sometimes even then some requirements will be un serviceable.
For this you will have to come up with a creative solution.
Again, most tool sets allow different degrees to express your creativity.

Some tools allow you to write code on some components. Some tools allow you to bring in external plugins.

So, this is where most of the time will be spent servicing the unserviceable. When the tools are made more sophisticated.

Then the end users ask more sophisticated questions!

Hierarchies

Grouping can be done at different Levels

- Year, Quarter, Month, Week
- Geographies(global, country, state, city)
- Product
- Division
- Line of Business
- Operating Center
- Business Unit

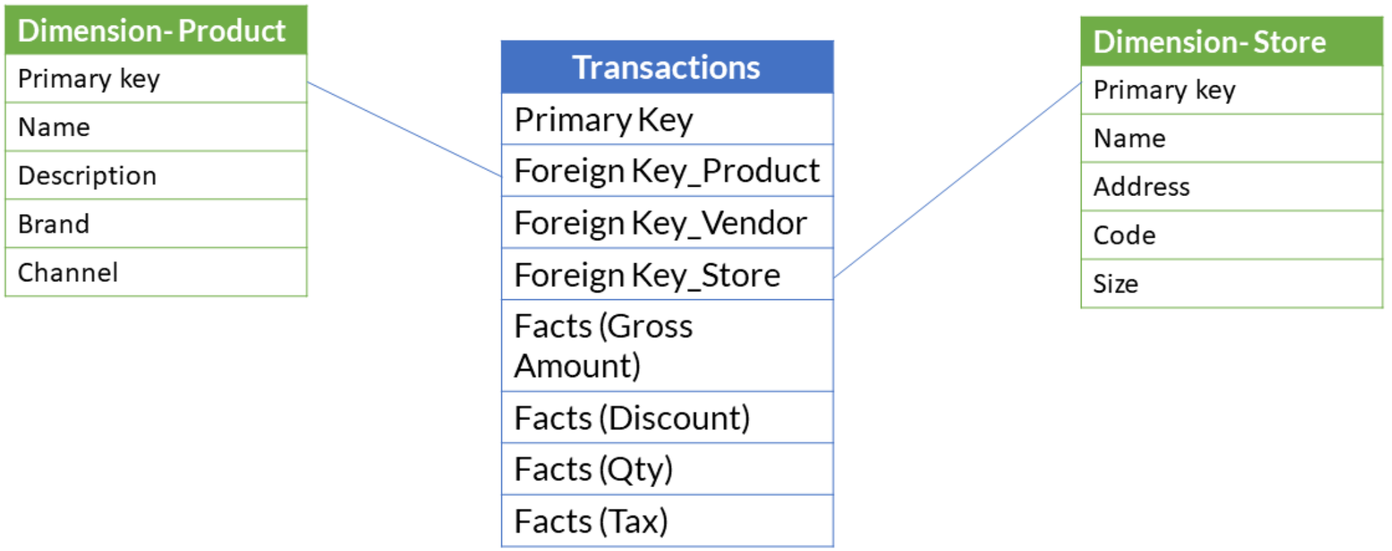
Most tools give this functionality
There many ways the data can be presented at different levels.
Do you want summaries at all levels?

Do you want summaries at the current level?
So again each end user will have there own specifications.

Cube

Tecquarian
Technology in hands of the people

For Reporting and Intelligence



Normalized - Source systems Data

Dimension (Product) Data

Table 2 hundreds of entries

Dimension (Customer) Data

Table 3 thousands of entries

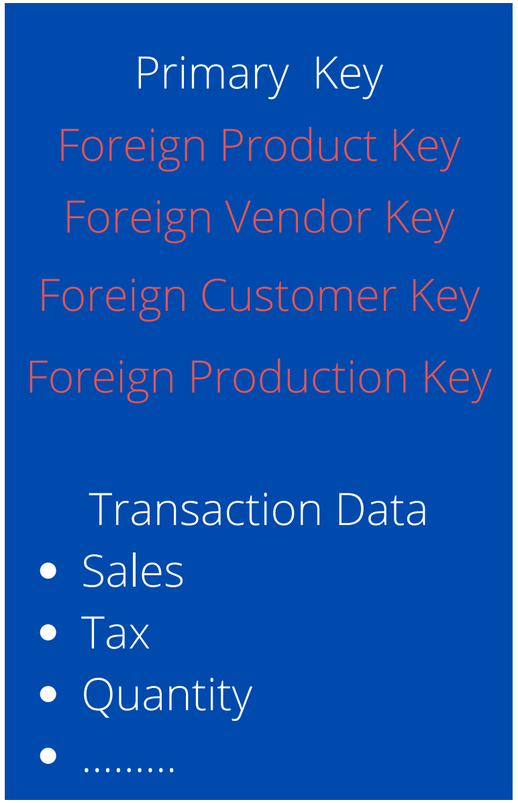
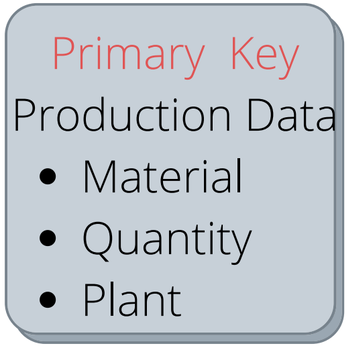


Table 1 - Millions of transactions

Dimension (Vendor) Data

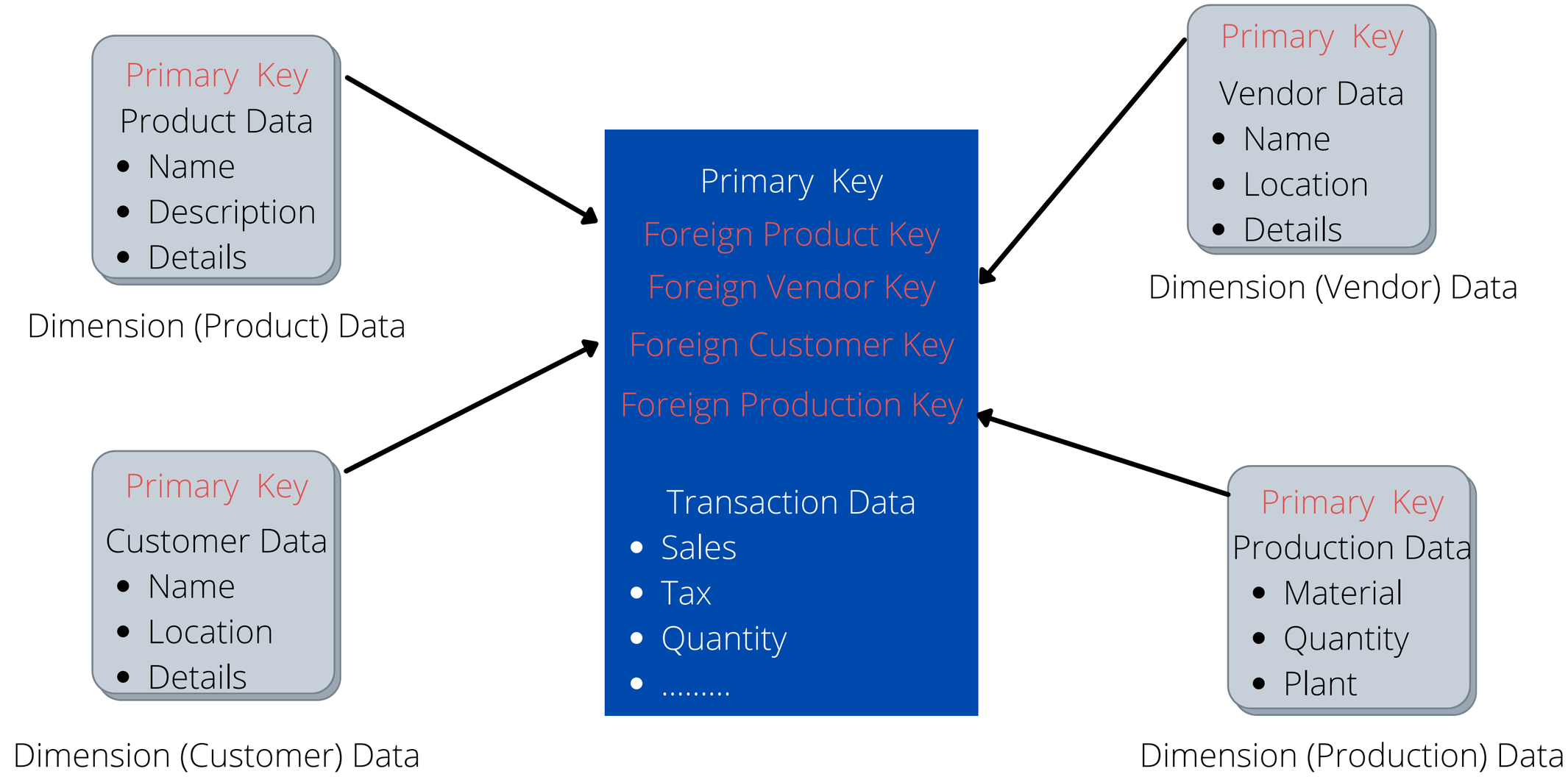
Table 5 hundres of entries



Dimension (Production) Data

Table 4 thousands of entries

Star Schema / Cube



Multi Dimensional Model

De-normalized