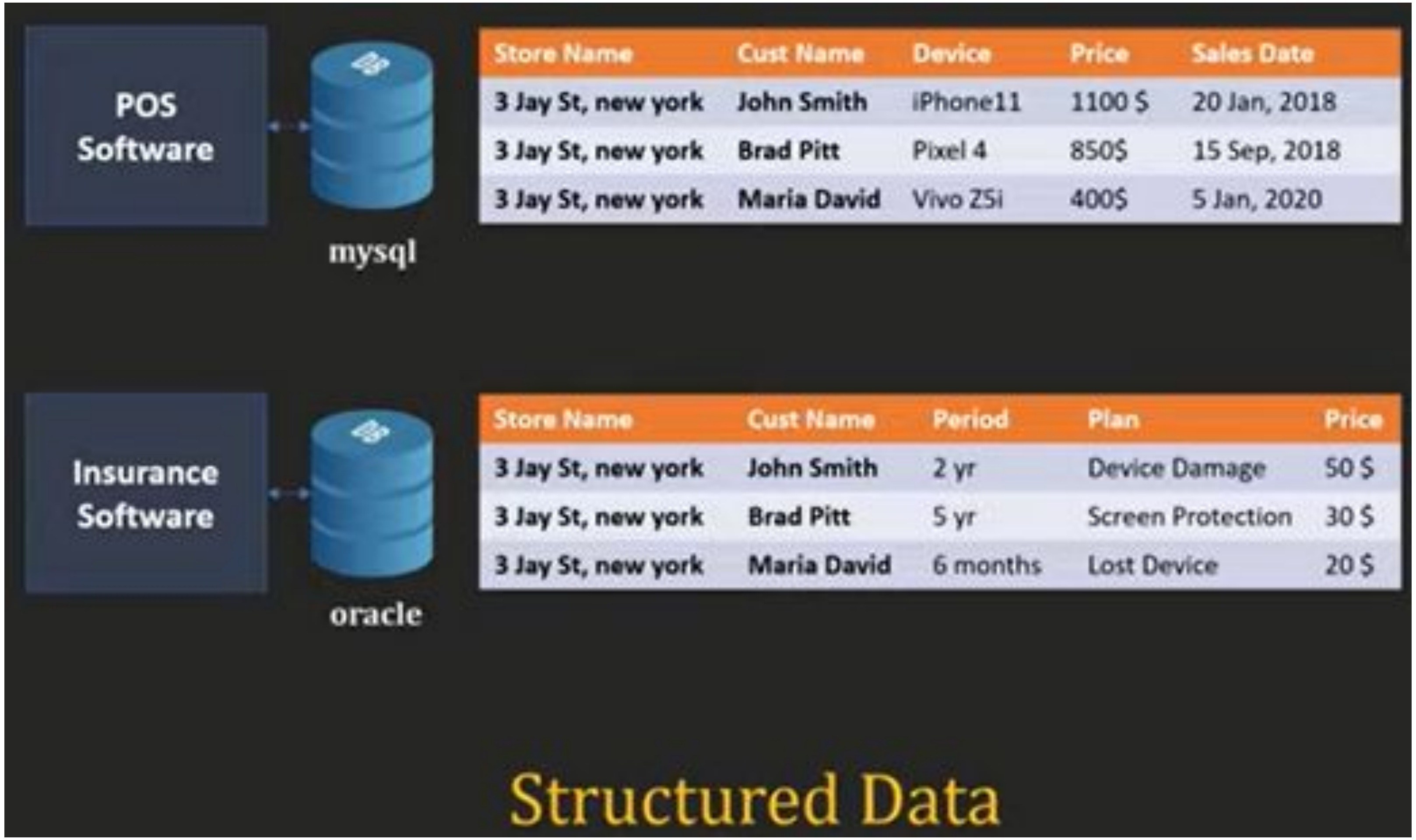
DATA WAREHOUSING & DATA MINING

By,

Mrs. Apeksha Waghmare
AP, IT Dept. TCET

STRUCTURED DATA

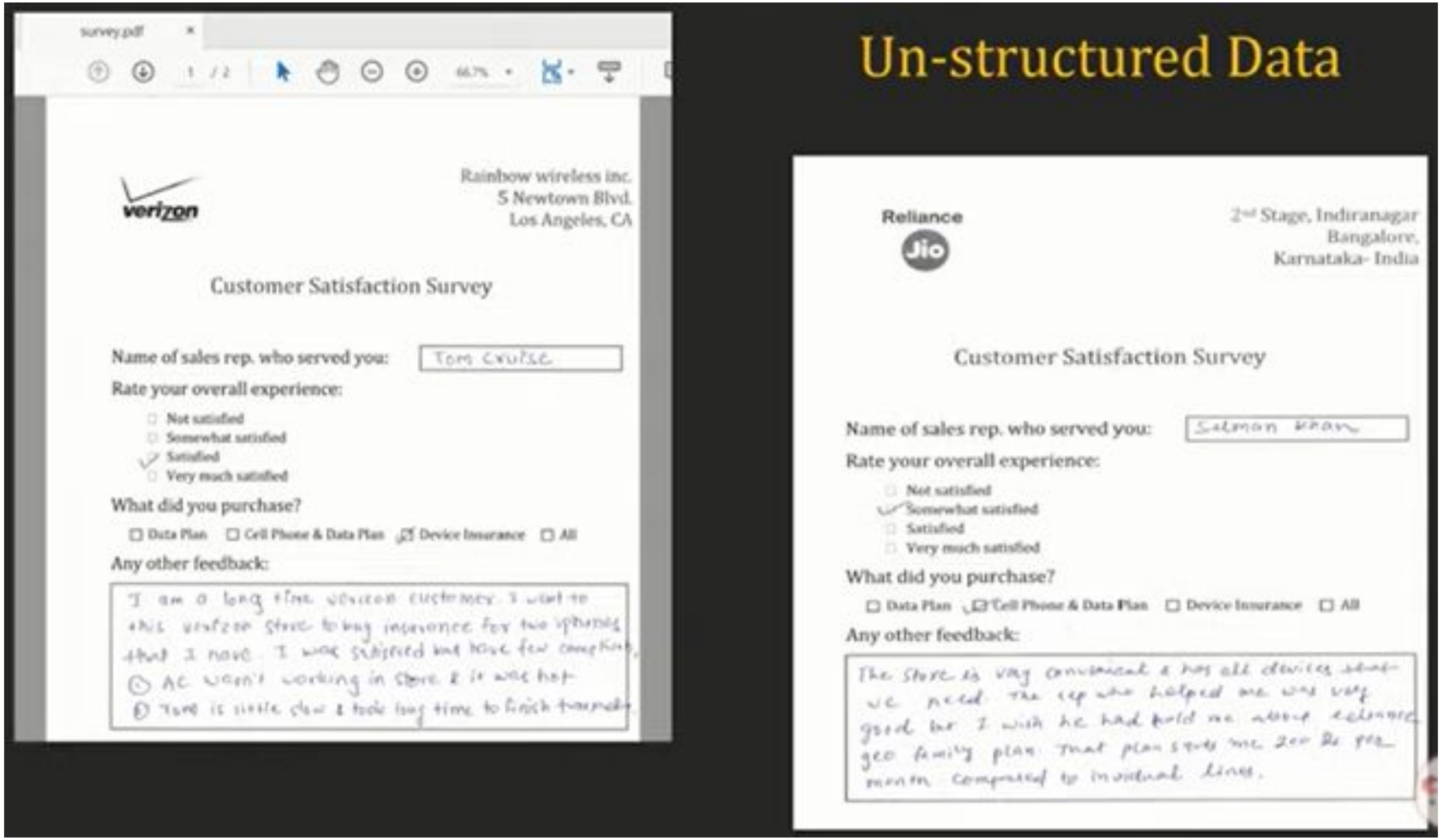


Store Name	Cust Name	Device	Price	Sales Date
3 Jay St, new york	John Smith	iPhone11	1100 $	20 Jan, 2018
3 Jay St, new york	Brad Pitt	Pixel 4	850$	15 Sep, 2018
3 Jay St, new york	Maria David	Vivo Z5i	400$	5 Jan, 2020

Store Name	Cust Name	Period	Plan	Price
3 Jay St, new york	John Smith	2 yr	Device Damage	50 $
3 Jay St, new york	Brad Pitt	5 yr	Screen Protection	30 $
3 Jay St, new york	Maria David	6 months	Lost Device	20 $

Structured Data

UN-STRUCTURED DATA



Un-structured Data

verizon

Rainbow wireless inc.
5 Newtown Blvd.
Los Angeles, CA

Customer Satisfaction Survey

Name of sales rep. who served you: Tom Cruise

Rate your overall experience:

- □ Not satisfied
- □ Somewhat satisfied
- ☑ Satisfied
- □ Very much satisfied

What did you purchase?

□ Data Plan □ Cell Phone & Data Plan ☑ Device Insurance □ All

Any other feedback:

I am a long time verizon customer. I went to this verizon store to buy insurance for two iphones that I have. I was satisfied but have few complaints.
① AC wasn't working in store & it was hot
② Rep is little slow & took long time to finish paperwork.

Reliance Jio

2nd Stage, Indiranagar
Bangalore,
Karnataka- India

Customer Satisfaction Survey

Name of sales rep. who served you: Salman khan

Rate your overall experience:

- □ Not satisfied
- ☑ Somewhat satisfied
- □ Satisfied
- □ Very much satisfied

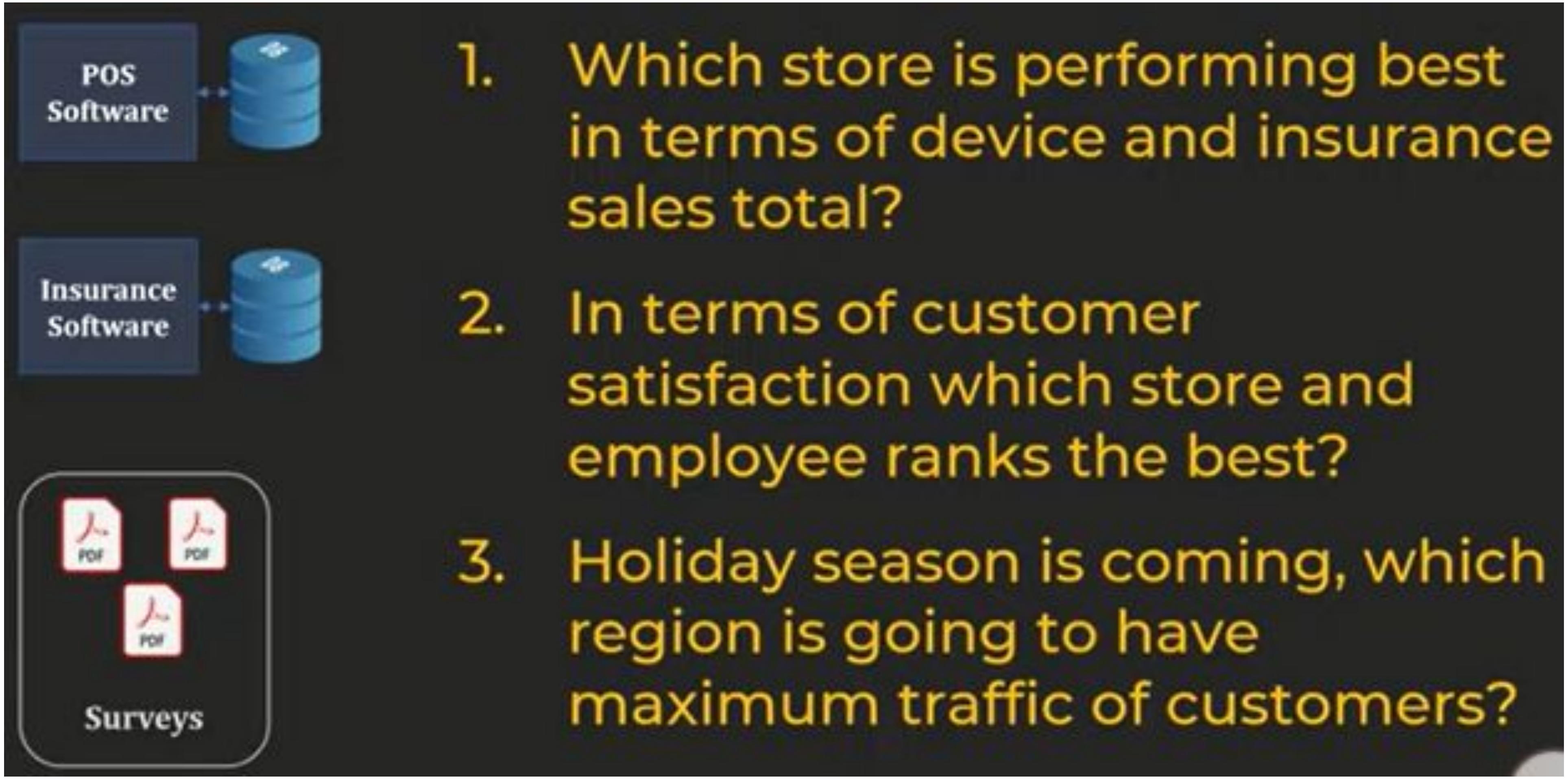
What did you purchase?

□ Data Plan ☑ Cell Phone & Data Plan □ Device Insurance □ All

Any other feedback:

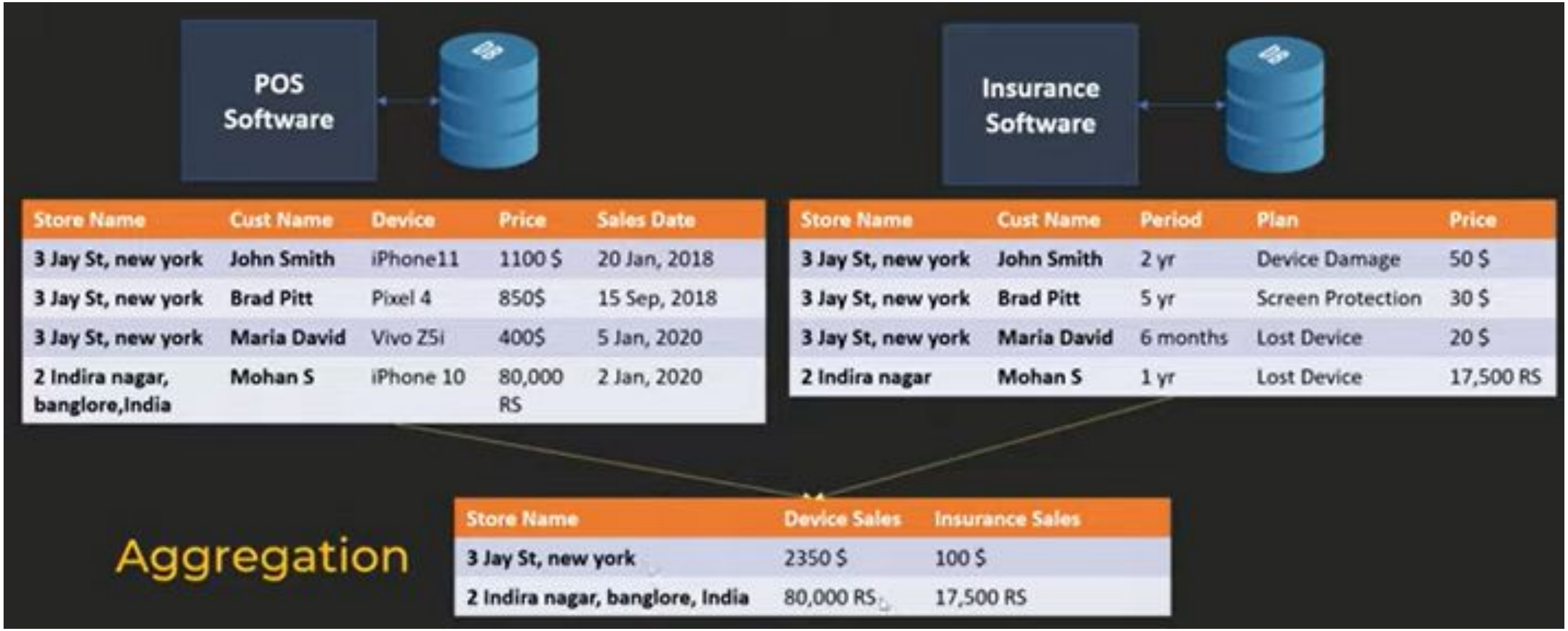
The store is very convenient & has all devices that we need. The rep who helped me was very good but I wish he had told me about reliance geo family plan. That plan saves me 200 Rs per month compared to individual lines.

QUESTIONS TO BE ANSWERED...



1. Which store is performing best in terms of device and insurance sales total?
2. In terms of customer satisfaction which store and employee ranks the best?
3. Holiday season is coming, which region is going to have maximum traffic of customers?

AGGREGATION



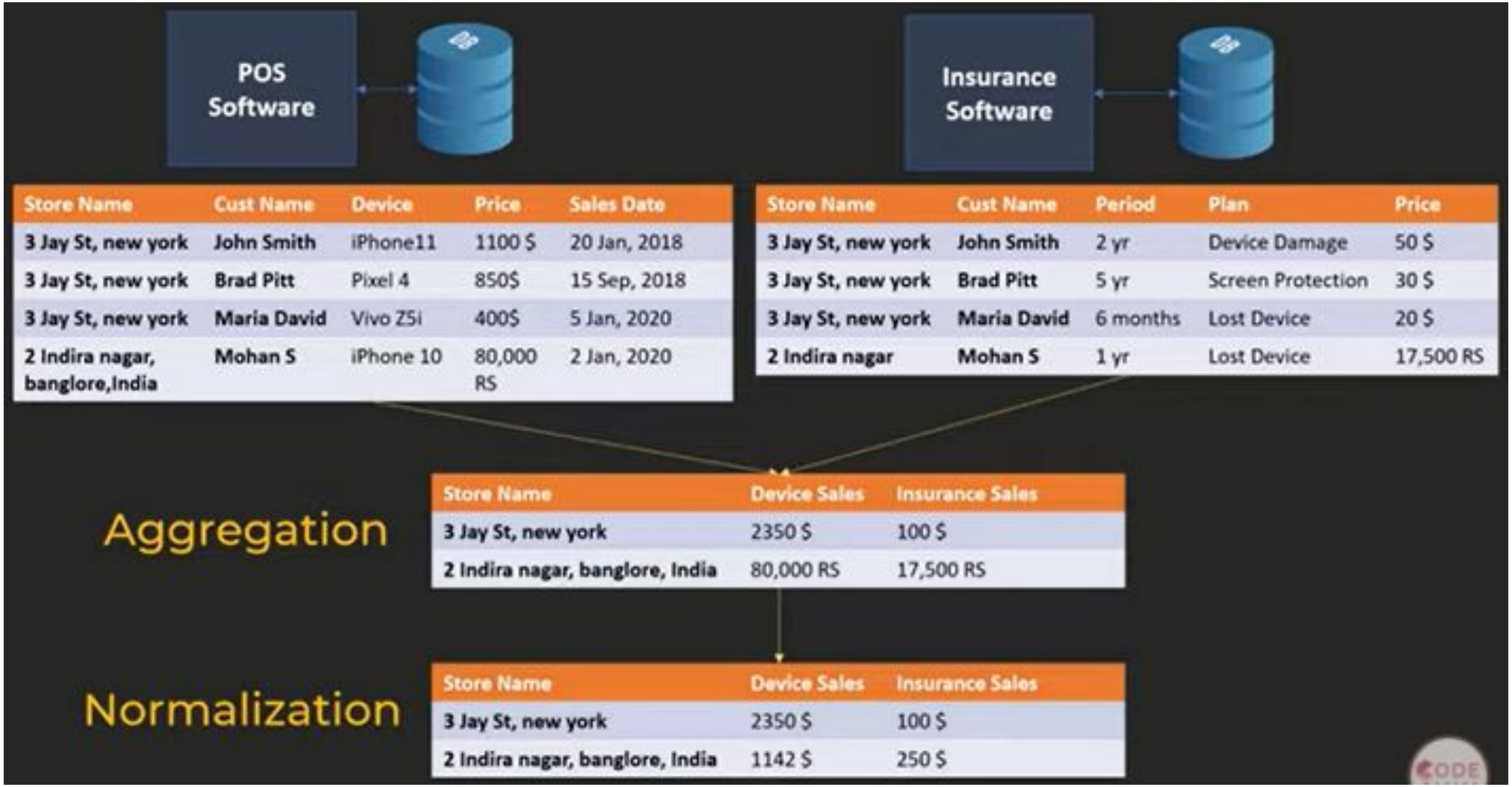
Store Name	Cust Name	Device	Price	Sales Date
3 Jay St, new york	John Smith	iPhone11	1100 $	20 Jan, 2018
3 Jay St, new york	Brad Pitt	Pixel 4	850$	15 Sep, 2018
3 Jay St, new york	Maria David	Vivo Z5i	400$	5 Jan, 2020
2 Indira nagar, banglore,India	Mohan S	iPhone 10	80,000 RS	2 Jan, 2020

Store Name	Cust Name	Period	Plan	Price
3 Jay St, new york	John Smith	2 yr	Device Damage	50 $
3 Jay St, new york	Brad Pitt	5 yr	Screen Protection	30 $
3 Jay St, new york	Maria David	6 months	Lost Device	20 $
2 Indira nagar	Mohan S	1 yr	Lost Device	17,500 RS

Aggregation

Store Name	Device Sales	Insurance Sales
3 Jay St, new york	2350 $	100 $
2 Indira nagar, banglore, India	80,000 RS	17,500 RS

NORMALIZATION



Store Name	Cust Name	Device	Price	Sales Date
3 Jay St, new york	John Smith	iPhone11	1100 $	20 Jan, 2018
3 Jay St, new york	Brad Pitt	Pixel 4	850$	15 Sep, 2018
3 Jay St, new york	Maria David	Vivo Z5i	400$	5 Jan, 2020
2 Indira nagar, banglore,India	Mohan S	iPhone 10	80,000 RS	2 Jan, 2020

Store Name	Cust Name	Period	Plan	Price
3 Jay St, new york	John Smith	2 yr	Device Damage	50 $
3 Jay St, new york	Brad Pitt	5 yr	Screen Protection	30 $
3 Jay St, new york	Maria David	6 months	Lost Device	20 $
2 Indira nagar	Mohan S	1 yr	Lost Device	17,500 RS

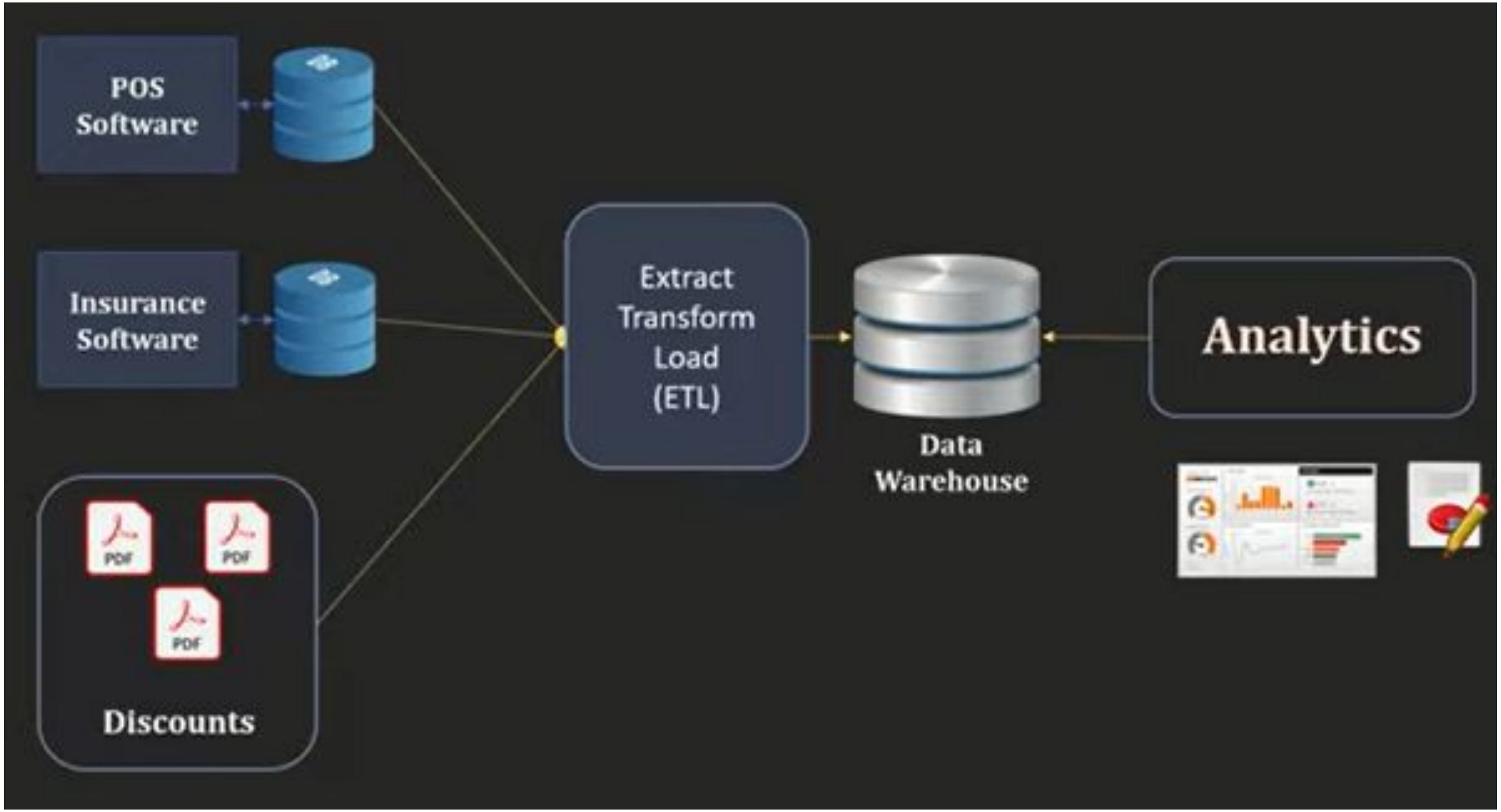
Aggregation

Store Name	Device Sales	Insurance Sales
3 Jay St, new york	2350 $	100 $
2 Indira nagar, banglore, India	80,000 RS	17,500 RS

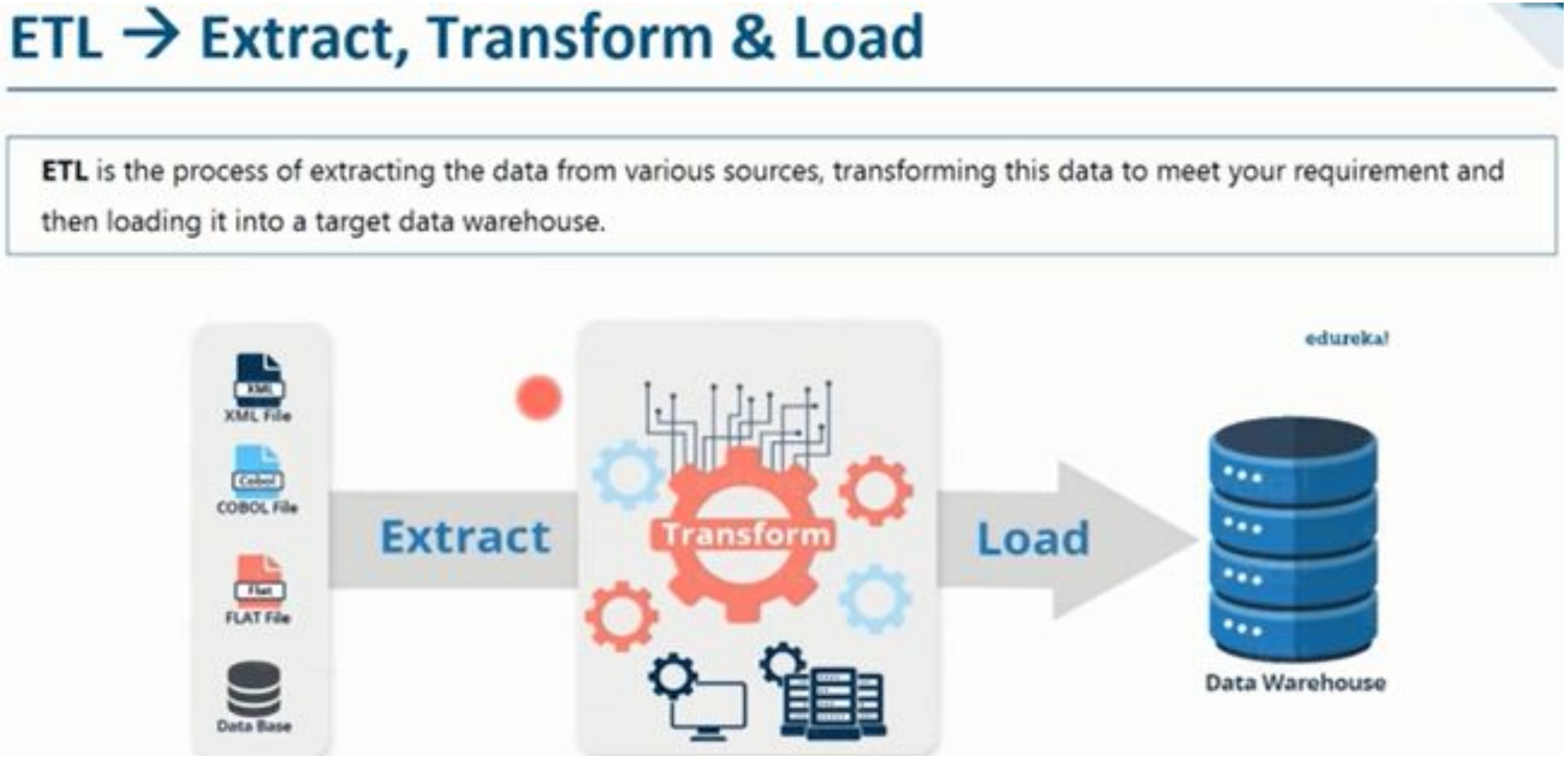
Normalization

Store Name	Device Sales	Insurance Sales
3 Jay St, new york	2350 $	100 $
2 Indira nagar, banglore, India	1142 $	250 $

DATA WAREHOUSE



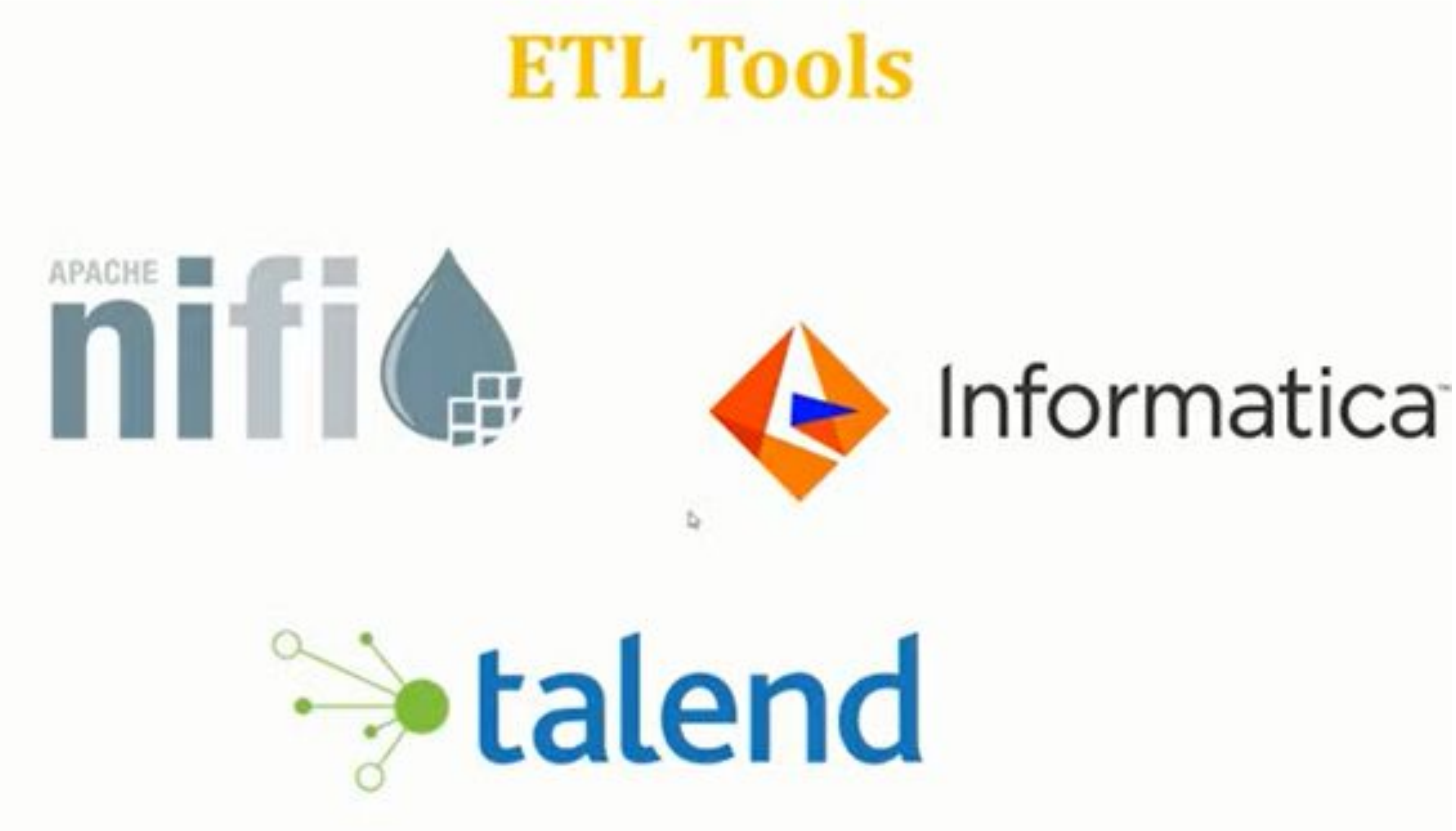
ETL PROCESS



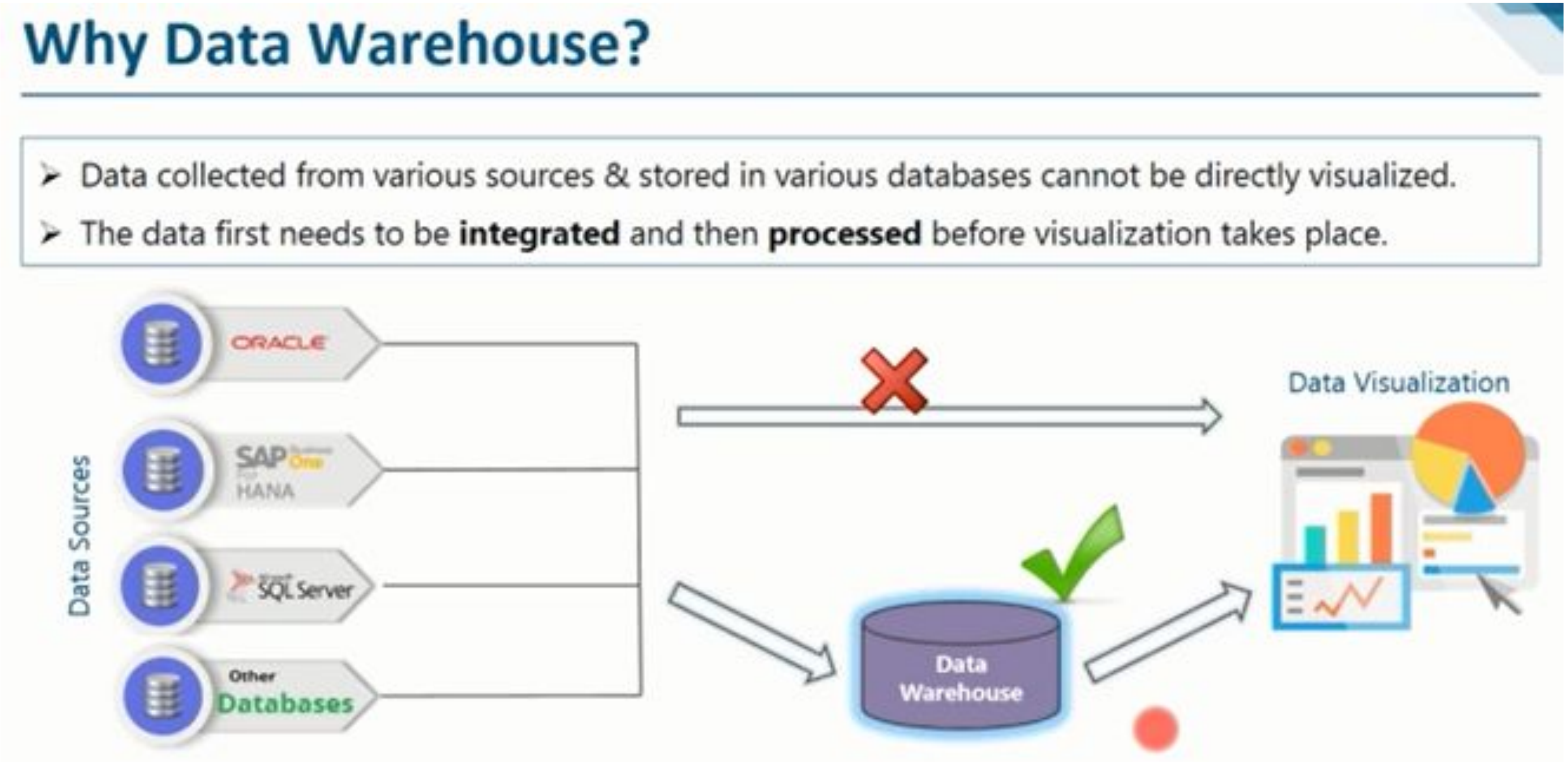
ETL → Extract, Transform & Load

ETL is the process of extracting the data from various sources, transforming this data to meet your requirement and then loading it into a target data warehouse.

ETL TOOLS



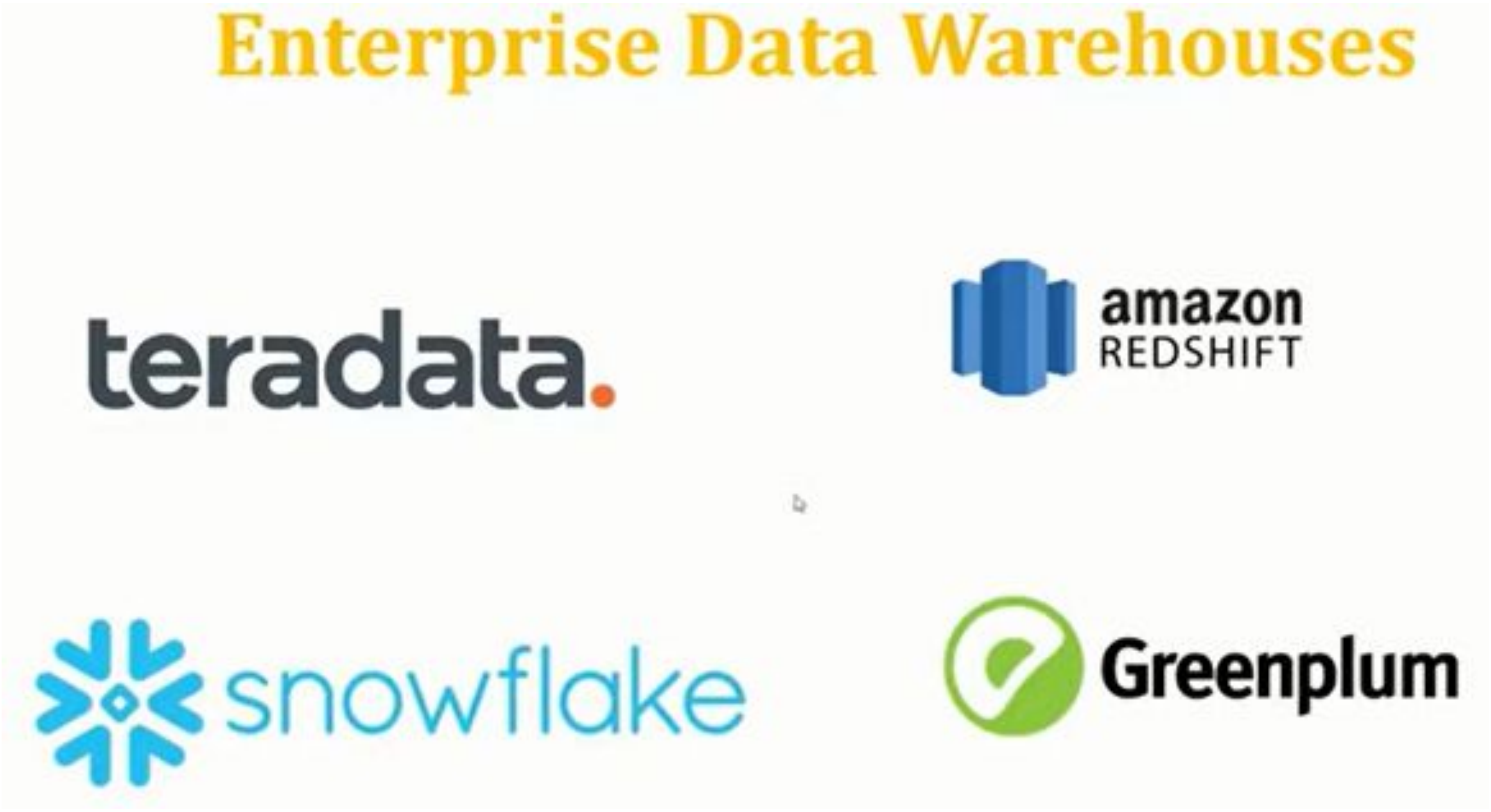
ETL Tools



Why Data Warehouse?

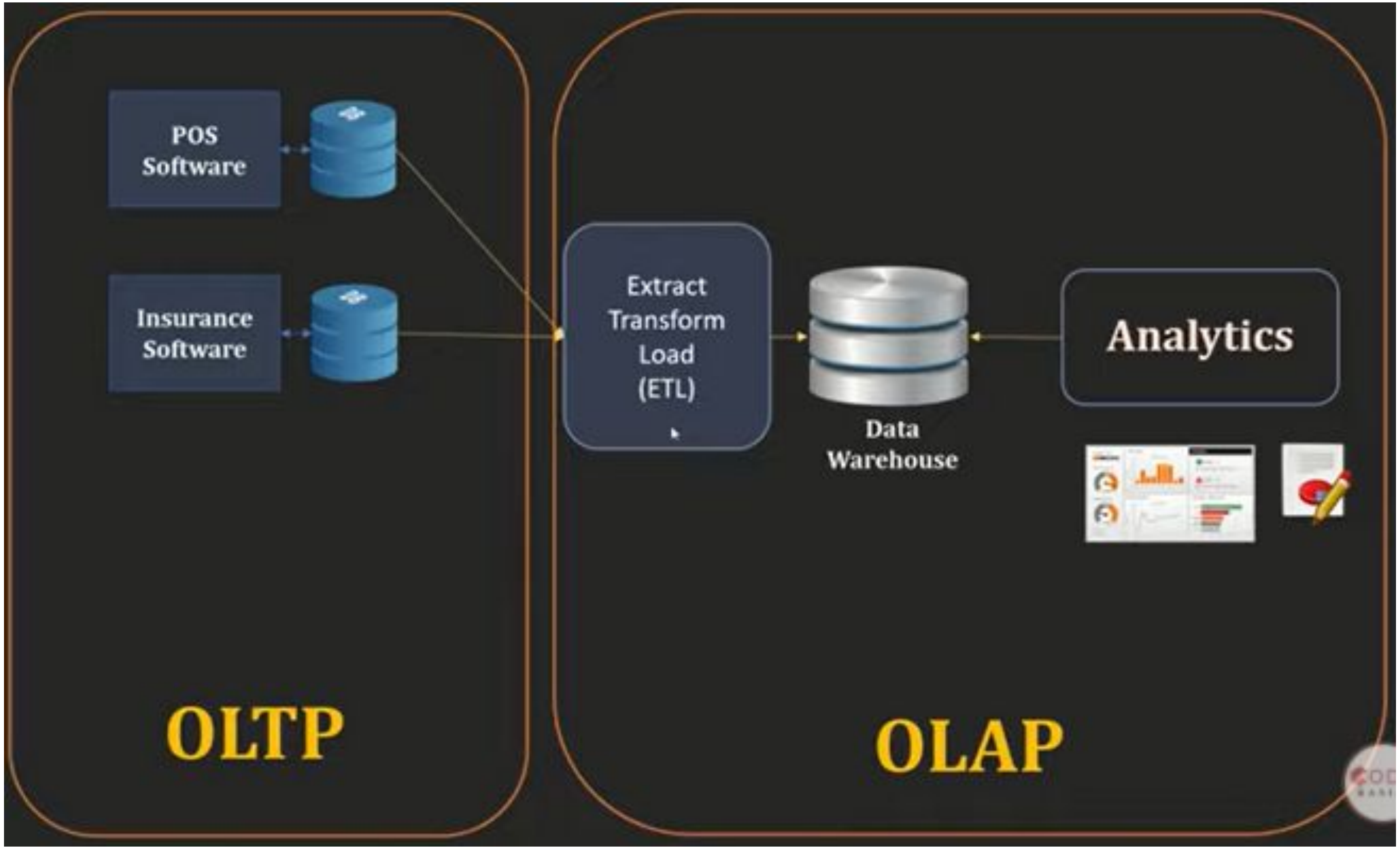
- Data collected from various sources & stored in various databases cannot be directly visualized.
- The data first needs to be **integrated** and then **processed** before visualization takes place.

DATA WAREHOUSES TOOLS



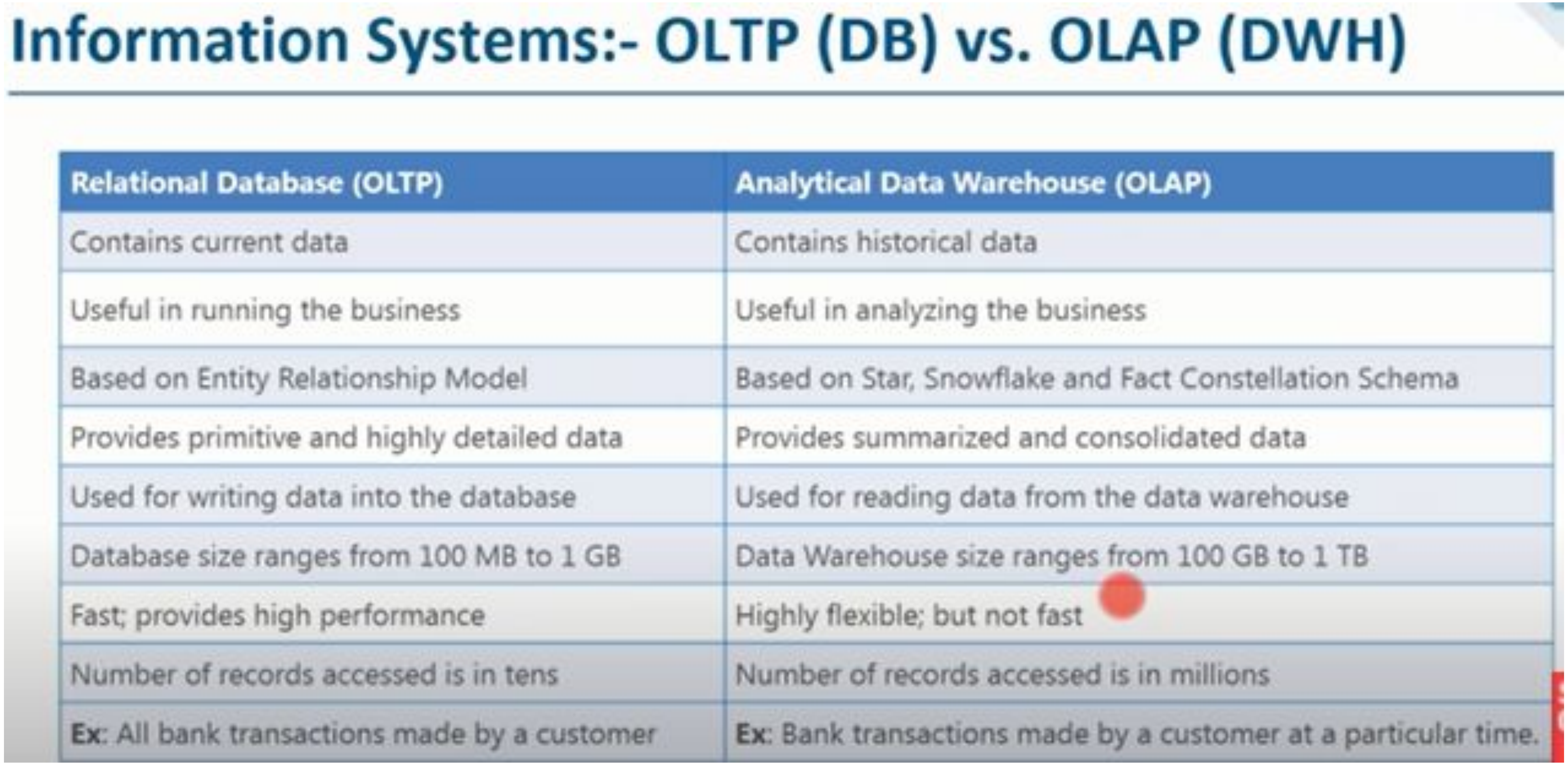
Enterprise Data Warehouses

OLTP/OLAP



POS Software
Insurance Software
Extract Transform Load (ETL)
Data Warehouse
Analytics
OLTP
OLAP

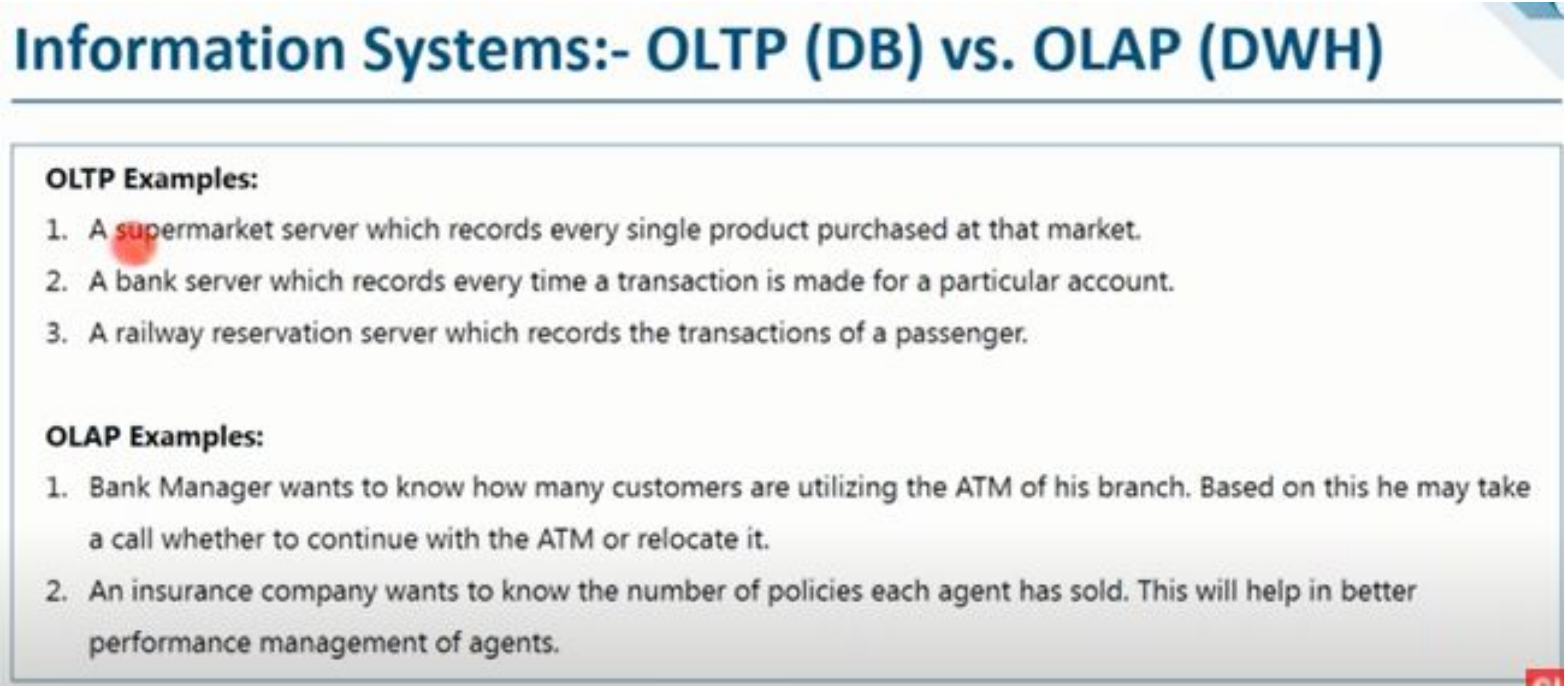
OLTP/OLAP



Information Systems:- OLTP (DB) vs. OLAP (DWH)

Relational Database (OLTP)	Analytical Data Warehouse (OLAP)
Contains current data	Contains historical data
Useful in running the business	Useful in analyzing the business
Based on Entity Relationship Model	Based on Star, Snowflake and Fact Constellation Schema
Provides primitive and highly detailed data	Provides summarized and consolidated data
Used for writing data into the database	Used for reading data from the data warehouse
Database size ranges from 100 MB to 1 GB	Data Warehouse size ranges from 100 GB to 1 TB
Fast; provides high performance	Highly flexible; but not fast
Number of records accessed is in tens	Number of records accessed is in millions
Ex: All bank transactions made by a customer	**Ex**: Bank transactions made by a customer at a particular time.

OLTP/OLAP



Information Systems:- OLTP (DB) vs. OLAP (DWH)

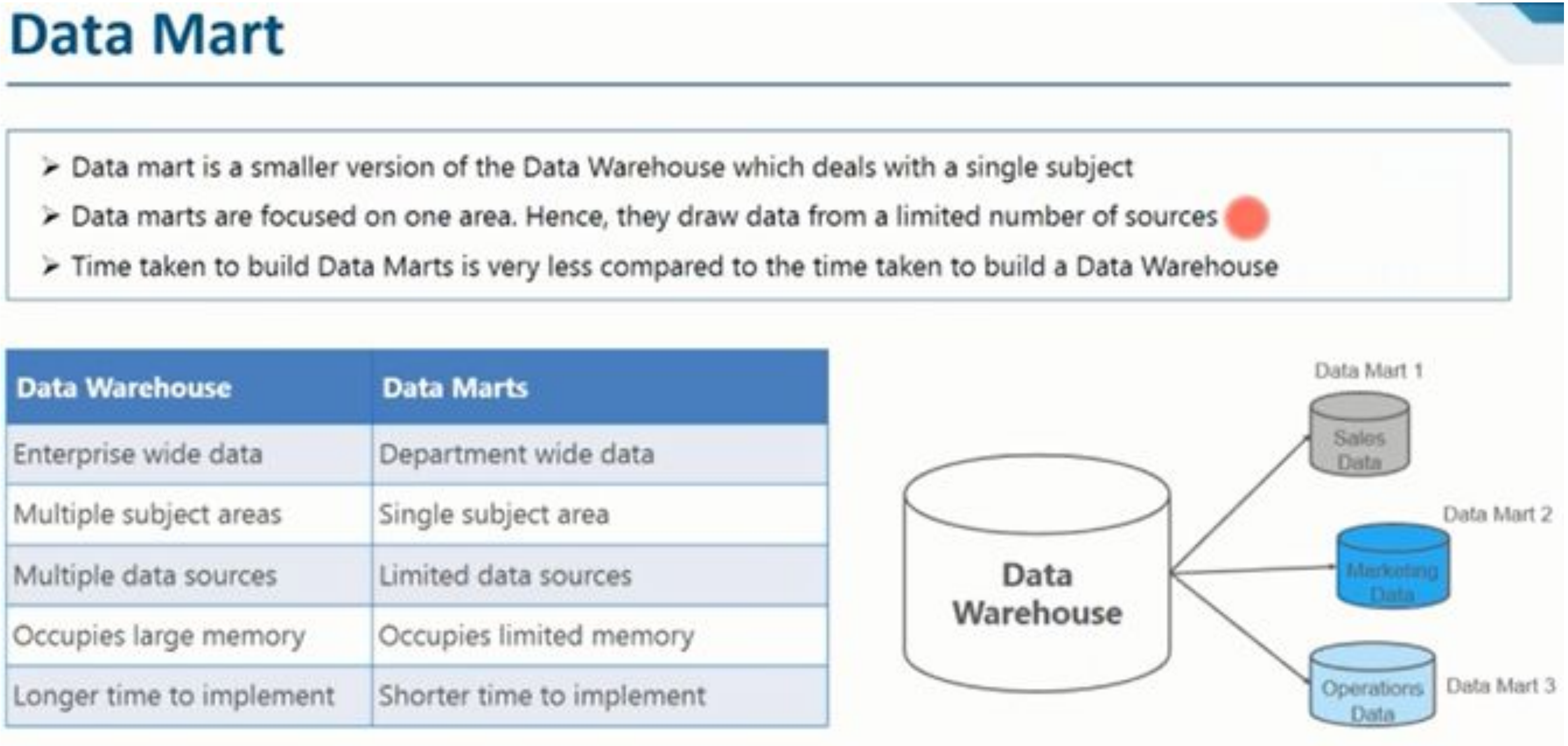
OLTP Examples:

1. A supermarket server which records every single product purchased at that market.
2. A bank server which records every time a transaction is made for a particular account.
3. A railway reservation server which records the transactions of a passenger.

OLAP Examples:

1. Bank Manager wants to know how many customers are utilizing the ATM of his branch. Based on this he may take a call whether to continue with the ATM or relocate it.
2. An insurance company wants to know the number of policies each agent has sold. This will help in better performance management of agents.

DATA MART

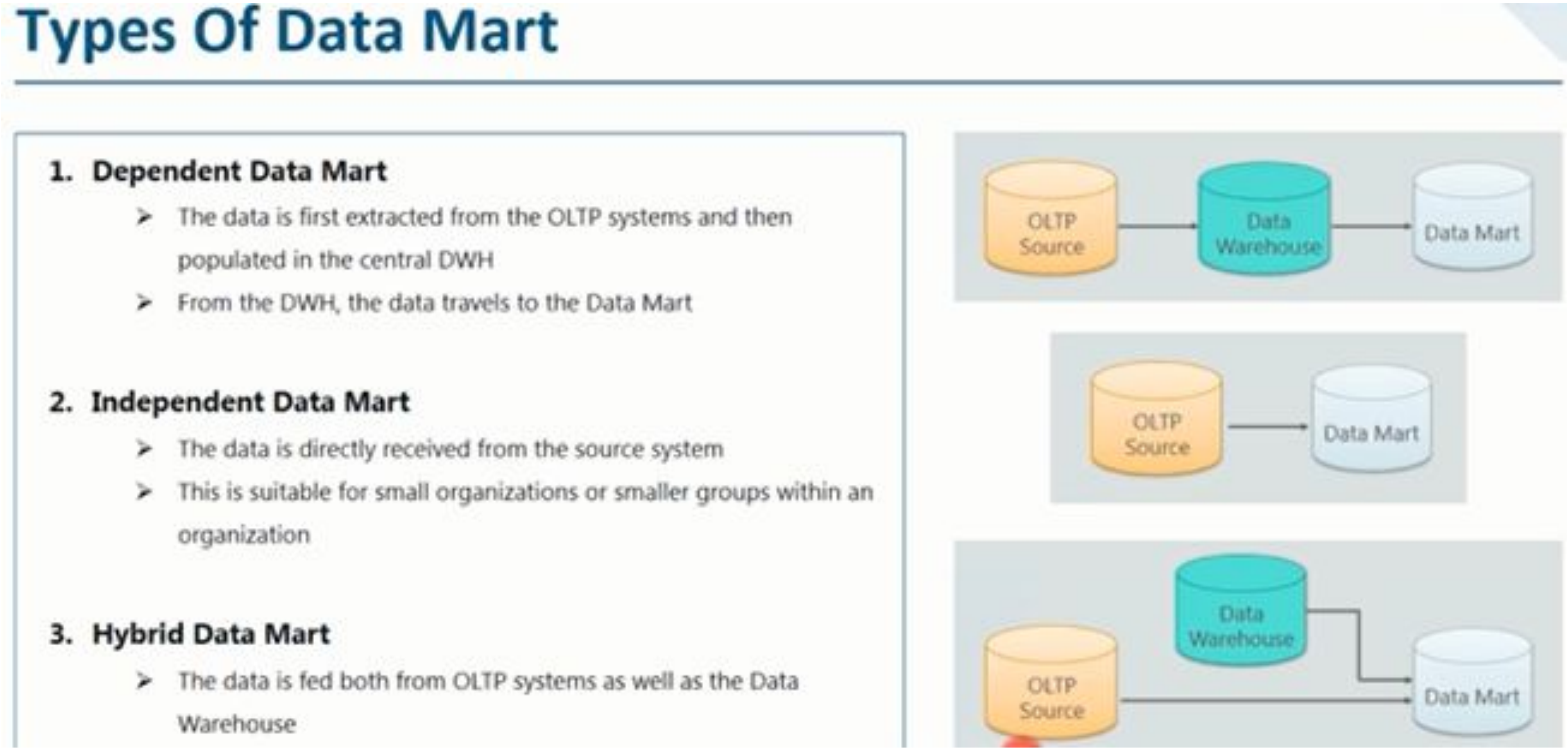


Data Mart

- ➢ Data mart is a smaller version of the Data Warehouse which deals with a single subject
- ➢ Data marts are focused on one area. Hence, they draw data from a limited number of sources
- ➢ Time taken to build Data Marts is very less compared to the time taken to build a Data Warehouse

Data Warehouse	Data Marts
Enterprise wide data	Department wide data
Multiple subject areas	Single subject area
Multiple data sources	Limited data sources
Occupies large memory	Occupies limited memory
Longer time to implement	Shorter time to implement

DATA MART



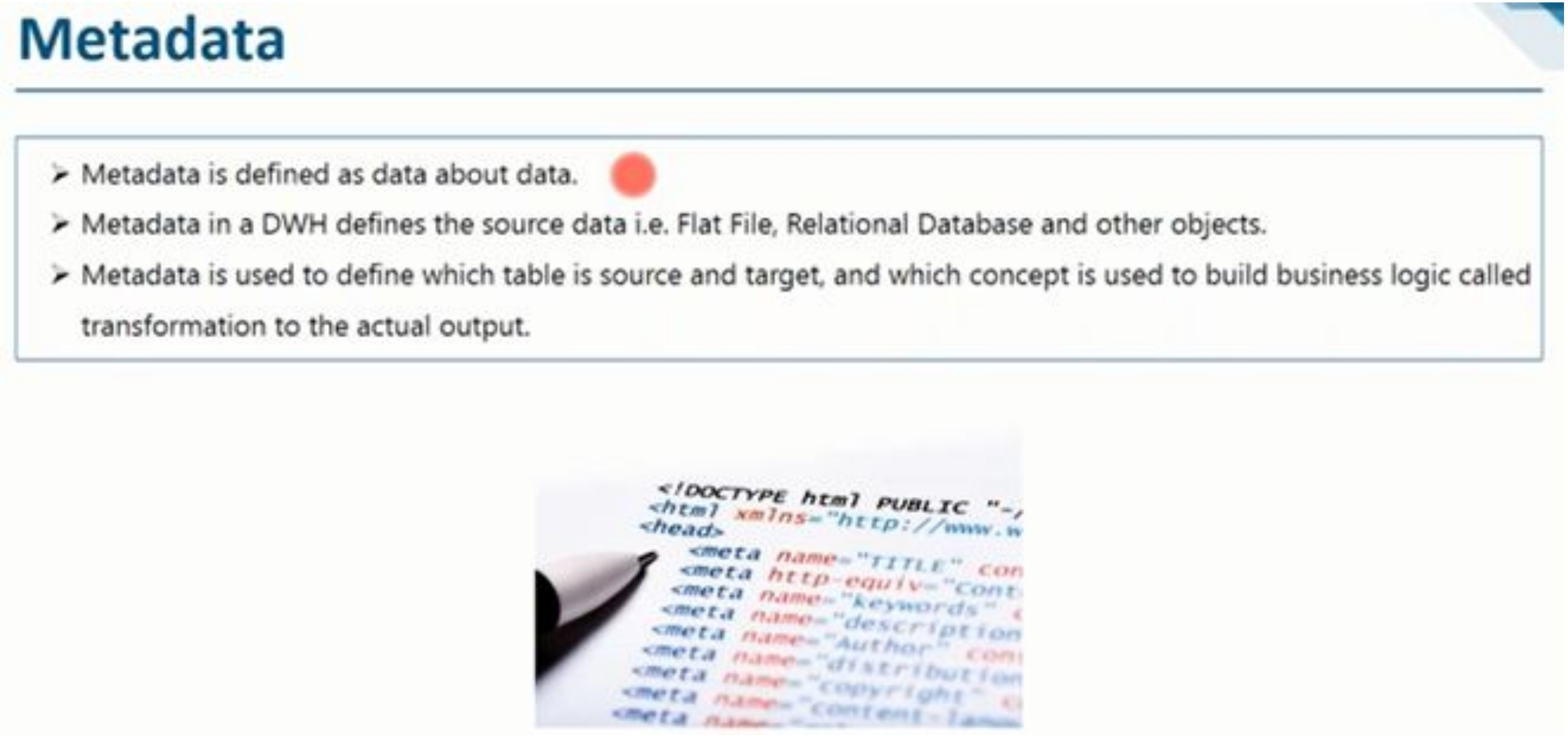
Types Of Data Mart

1. **Dependent Data Mart**
   - The data is first extracted from the OLTP systems and then populated in the central DWH
   - From the DWH, the data travels to the Data Mart

2. **Independent Data Mart**
   - The data is directly received from the source system
   - This is suitable for small organizations or smaller groups within an organization

3. **Hybrid Data Mart**
   - The data is fed both from OLTP systems as well as the Data Warehouse

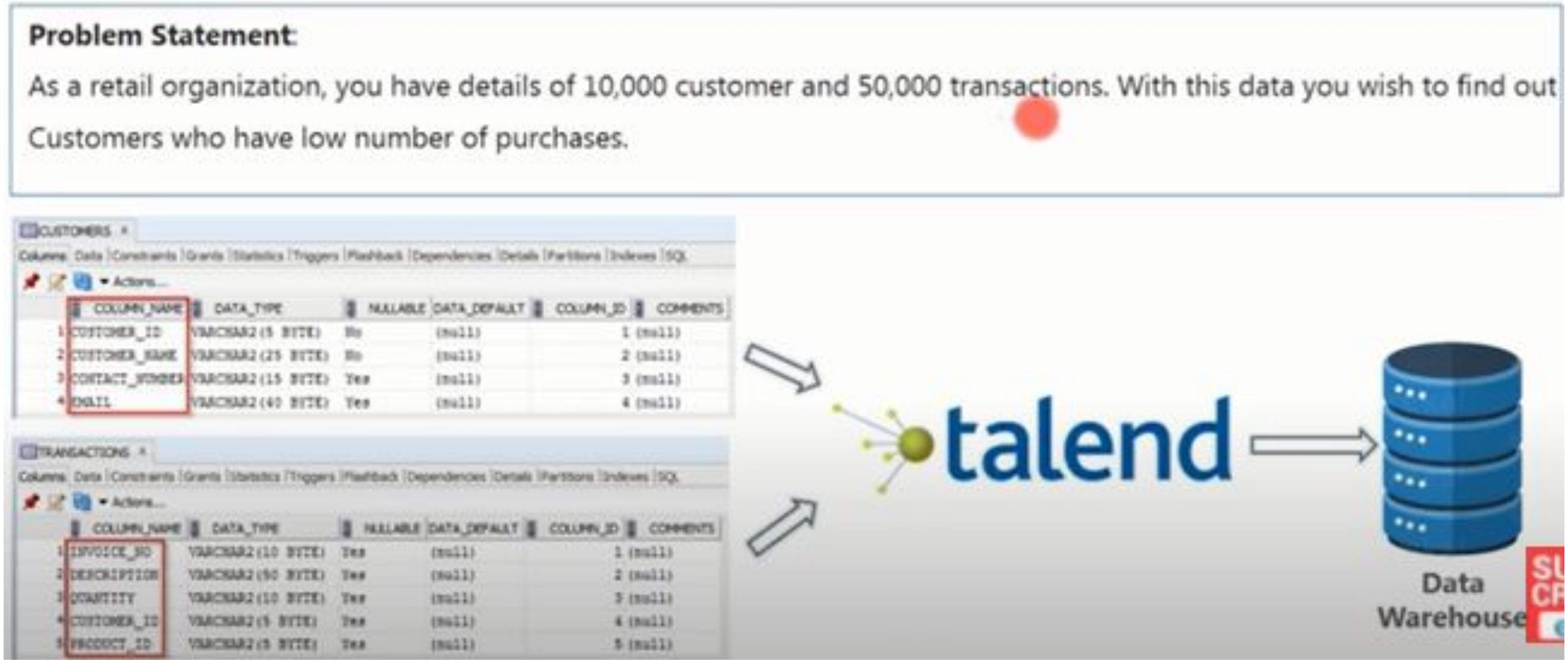
METADATA



Metadata

- ➢ Metadata is defined as data about data.
- ➢ Metadata in a DWH defines the source data i.e. Flat File, Relational Database and other objects.
- ➢ Metadata is used to define which table is source and target, and which concept is used to build business logic called transformation to the actual output.

TALEND ETL TOOL



Problem Statement:

As a retail organization, you have details of 10,000 customer and 50,000 transactions. With this data you wish to find out Customers who have low number of purchases.

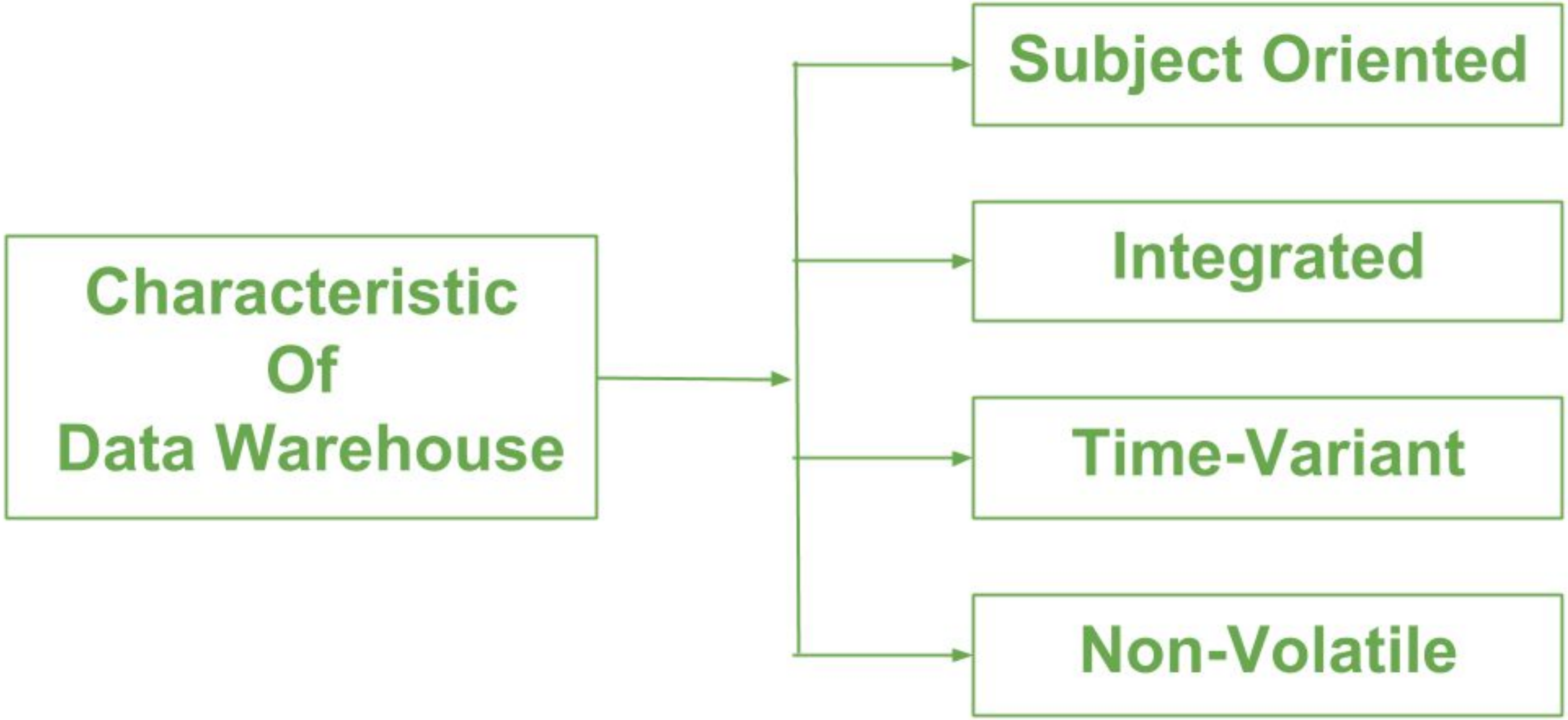
CUSTOMERS

	COLUMN_NAME	DATA_TYPE	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
1	CUSTOMER_ID	VARCHAR2(5 BYTE)	No	(null)	1	(null)
2	CUSTOMER_NAME	VARCHAR2(25 BYTE)	No	(null)	2	(null)
3	CONTACT_NUMBER	VARCHAR2(15 BYTE)	Yes	(null)	3	(null)
4	EMAIL	VARCHAR2(40 BYTE)	Yes	(null)	4	(null)

TRANSACTIONS

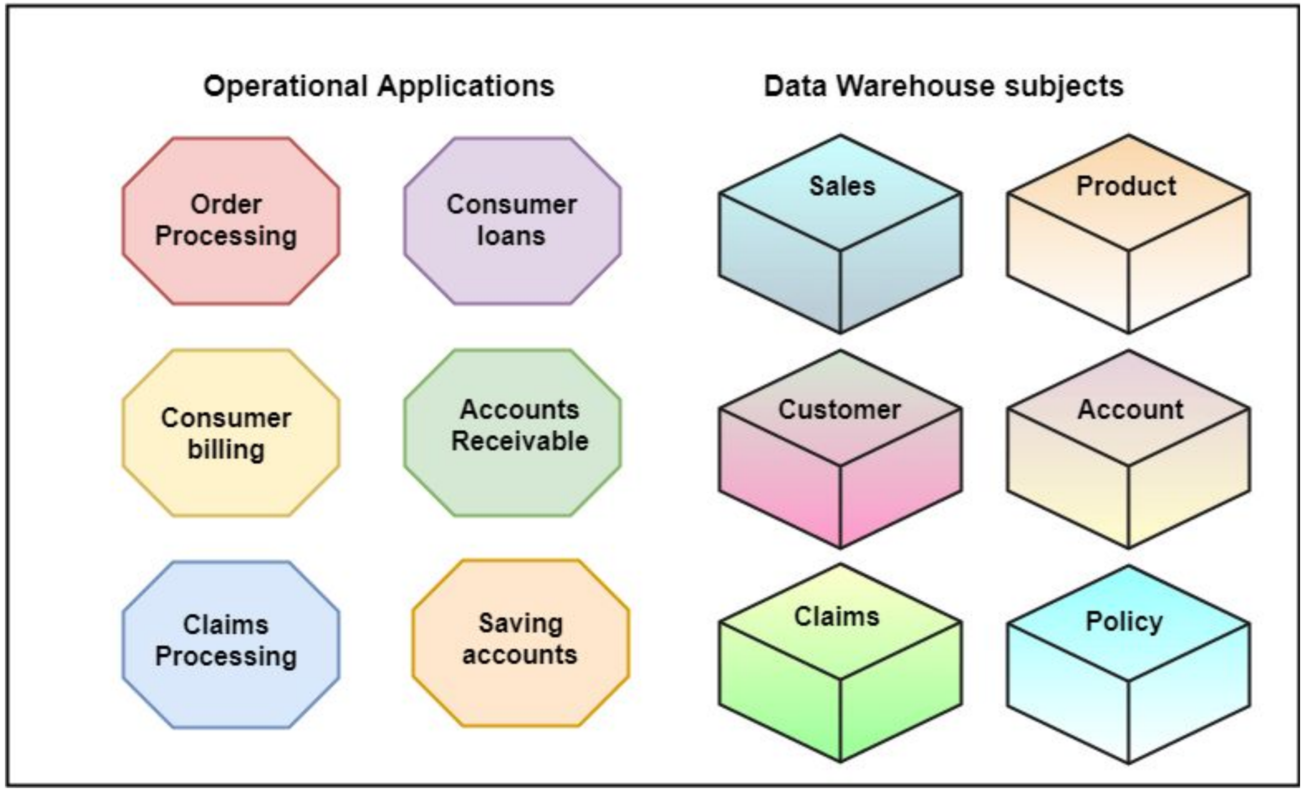
	COLUMN_NAME	DATA_TYPE	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
1	INVOICE_NO	VARCHAR2(10 BYTE)	Yes	(null)	1	(null)
2	DESCRIPTION	VARCHAR2(50 BYTE)	Yes	(null)	2	(null)
3	QUANTITY	VARCHAR2(10 BYTE)	Yes	(null)	3	(null)
4	CUSTOMER_ID	VARCHAR2(5 BYTE)	Yes	(null)	4	(null)
5	PRODUCT_ID	VARCHAR2(5 BYTE)	Yes	(null)	5	(null)

CHARACTERISTICS OF DATA WAREHOUSE



CHARACTERISTICS OF DATA WAREHOUSE

Data Warehouse is Subject-Oriented

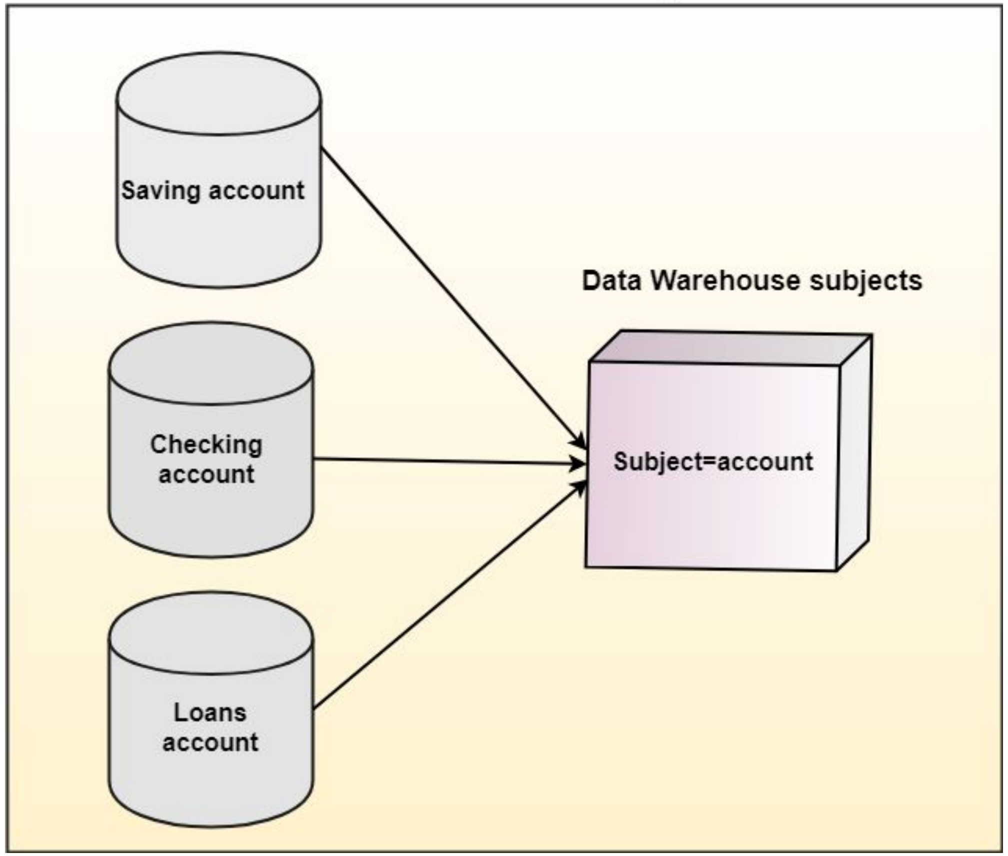


1. Subject-oriented –

- A data warehouse is always a subject oriented as it delivers information about a theme instead of organization's current operations.
- These themes can be sales, distributions, marketing etc.

CHARACTERISTICS OF DATA WAREHOUSE

Data Warehouse is Integrated



2. Integrated

- A data warehouse integrates various heterogeneous data sources like RDBMS, flat files, and online transaction records.
- It requires performing data cleaning and integration during data warehousing to ensure consistency in naming conventions, attributes types, etc., among different data sources.

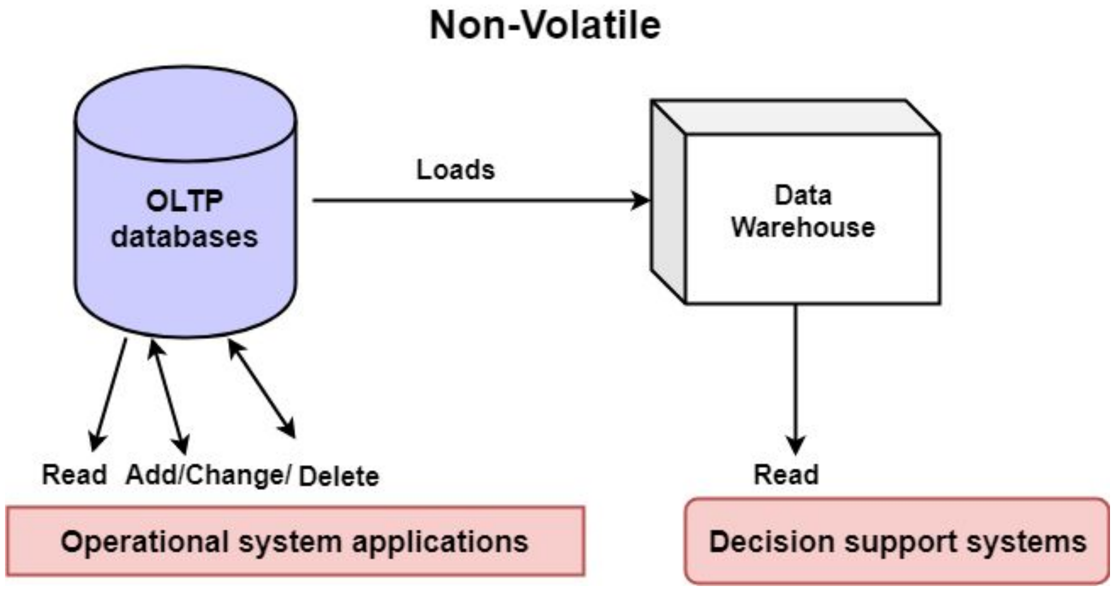
CHARACTERISTICS OF DATA WAREHOUSE

3. Time-Variant

Historical information is kept in a data warehouse. For example, one can retrieve files from 3 months, 6 months, 12 months, or even previous data from a data warehouse.

4. Non-Volatile

Non-Volatile defines that once data entered into the warehouse, and data should not change.



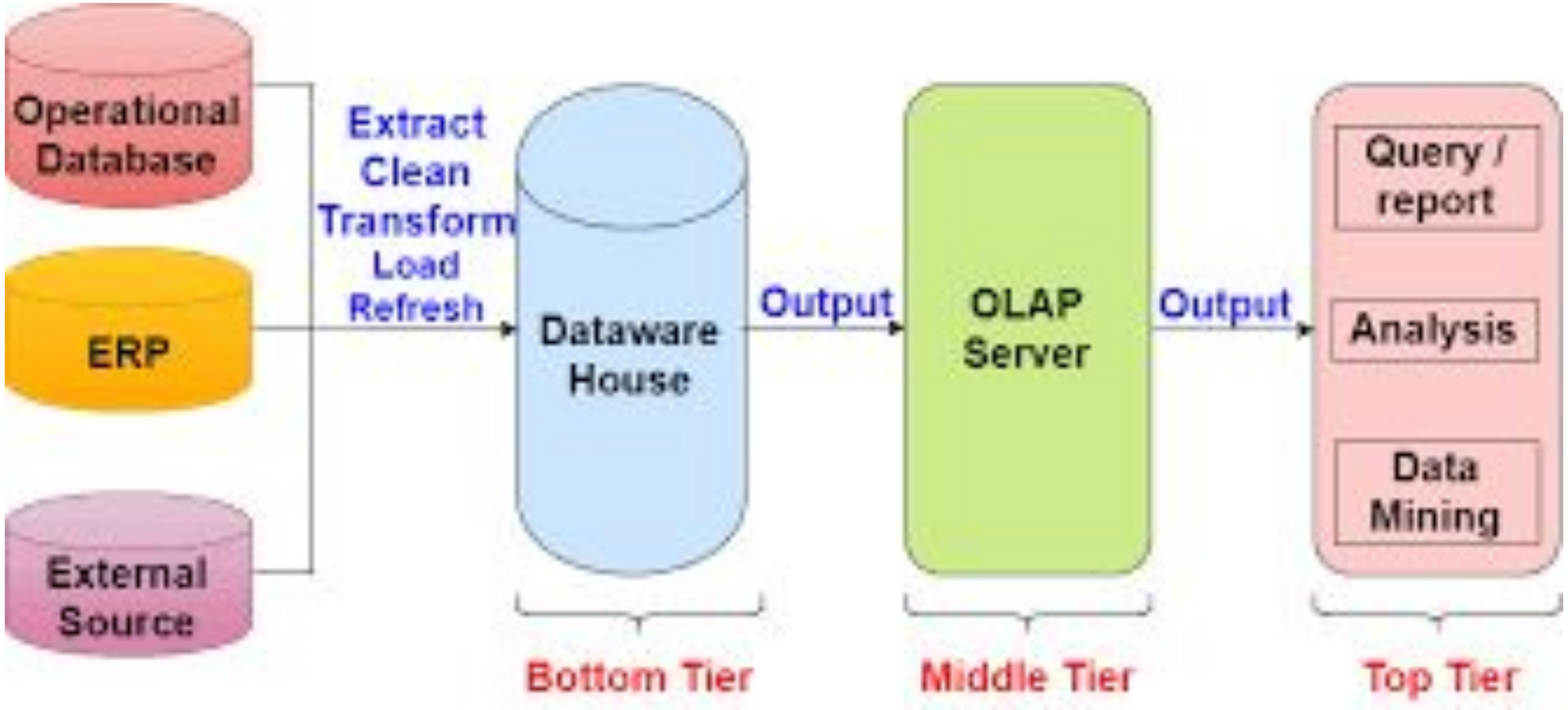
THREE-TIER DATA WAREHOUSE ARCHITECTURE

Bottom Tier: The database of the Data warehouse servers as the bottom tier. It is usually a relational database system. Data is cleansed, transformed, and loaded into this layer using back-end tools.

Middle Tier: The middle tier in Data warehouse is an OLAP server which is implemented using either ROLAP or MOLAP model. For a user, this application tier presents an abstracted view of the database. This layer also acts as a mediator between the end-user and the database.

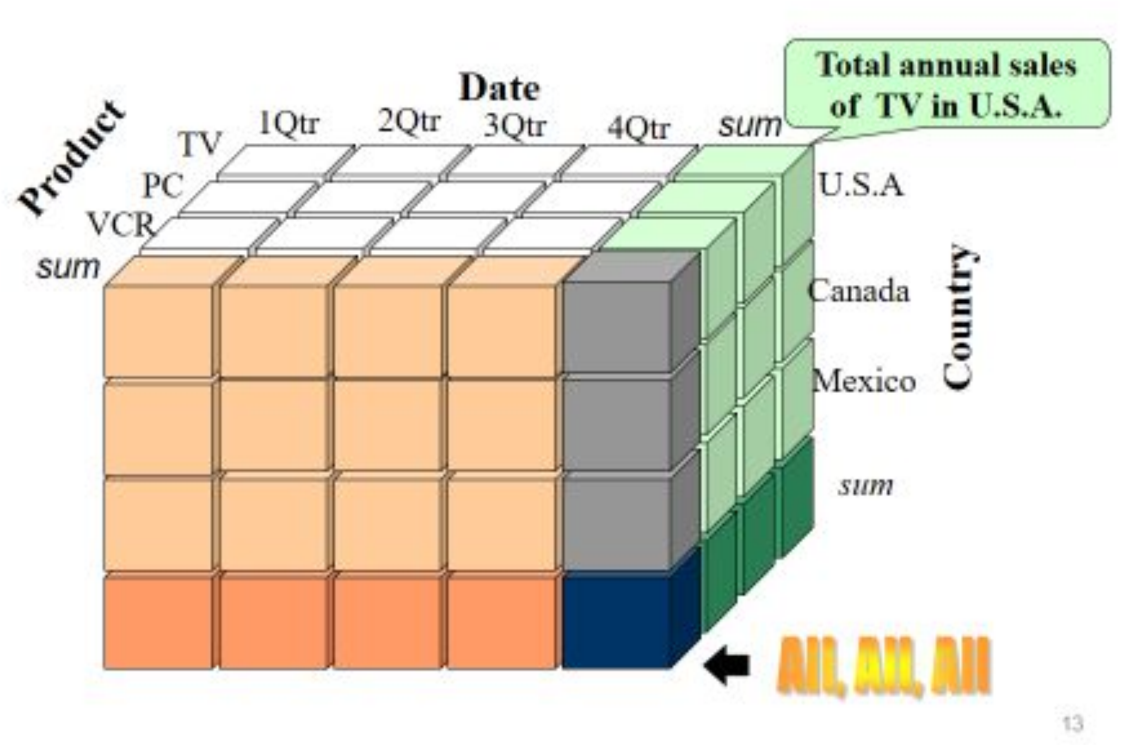
Top-Tier: The top tier is a front-end client layer. Top tier is the tools and API that you connect and get data out from the data warehouse. It could be Query tools, reporting tools, managed query tools, Analysis tools and Data mining tools.

THREE-TIER DATA WAREHOUSE ARCHITECTURE



TYPES OF OLAP SERVERS

A Sample Data Cube



1. **Relational OLAP (ROLAP) – Star Schema based –**
   - In ROLAP data is stored in a relational database

2. Multidimensional OLAP (MOLAP) – Cube based –

- MOLAP cubes are fast data retrieval, optimal for slicing and dicing and they can perform complex calculation.

3. Hybrid OLAP (HOLAP)

- HOLAP is a combination of ROLAP and MOLAP
- Cubes are smaller than MOLAP since detail data is kept in the relational database.

TYPICAL OLAP OPERATIONS

1. Roll up (drill-up): summarize data

□by climbing up hierarchy or by dimension reduction

2. Drill down (roll down): reverse of roll-up

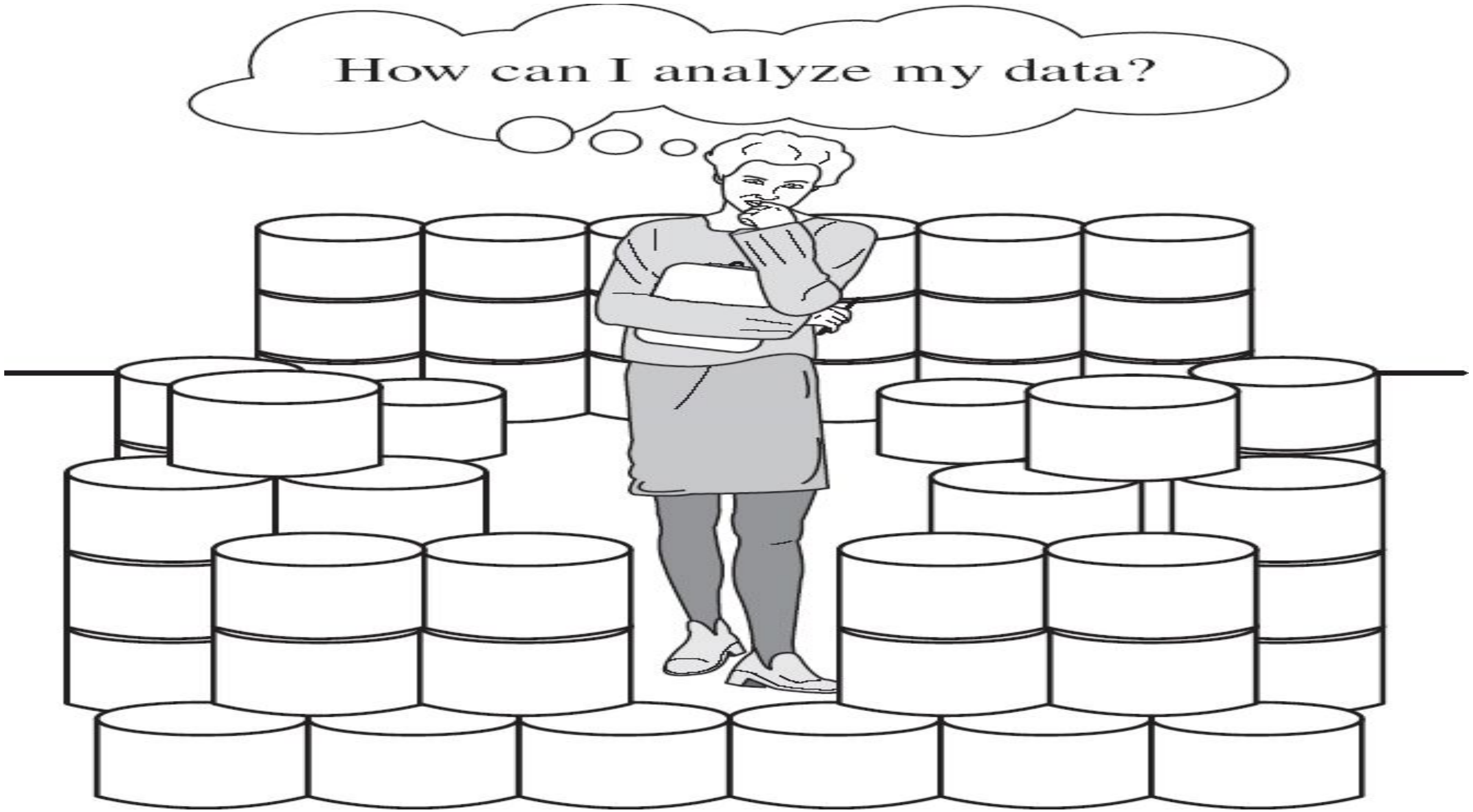
□from higher level summary to lower level summary or detailed data

3. Slice : selection on one dimension of the given cube, resulting in a sub cube

4. Dice: define a sub cube by performing a selection on two or more dimensions

5. Pivot (rotate): rotates the data axes in order to provide an alternative presentation of the data.

Data Rich, Information Poor



WHAT IS DATA MINING?

Data mining :knowledge discovery from data(KDD)

- Extraction of interesting patterns or knowledge from huge amount of data
- It looks for hidden patterns within the data set and try to predict future behavior.
- This allows the business to take the data-driven decision

WHAT IS DATA MINING?

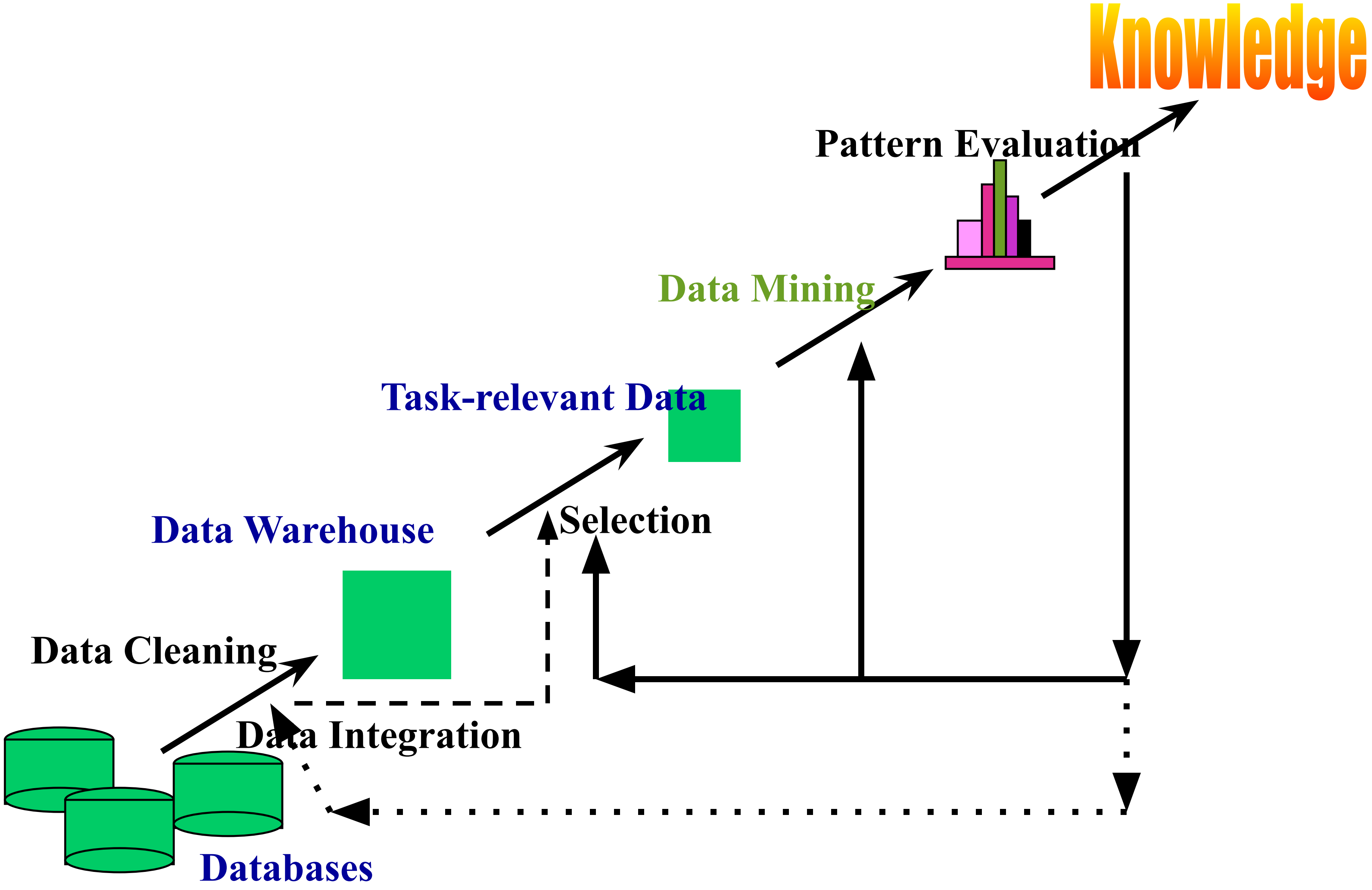
Data mining (knowledge discovery from data)

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

Other names

- Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Knowledge Discovery (KDD) Process



Knowledge

Pattern Evaluation

Data Mining

Task-relevant Data

Selection

Data Warehouse

Data Cleaning

Data Integration

Databases

KNOWLEDGE DISCOVERY FROM DATA

KDD process includes

1. Data cleaning (to remove noise and inconsistent data)

2. data integration (where multiple data sources may be combined)

3. data selection (where data relevant to the analysis task are retrieved from the database)

4. data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations)

KDD CONTINUED….

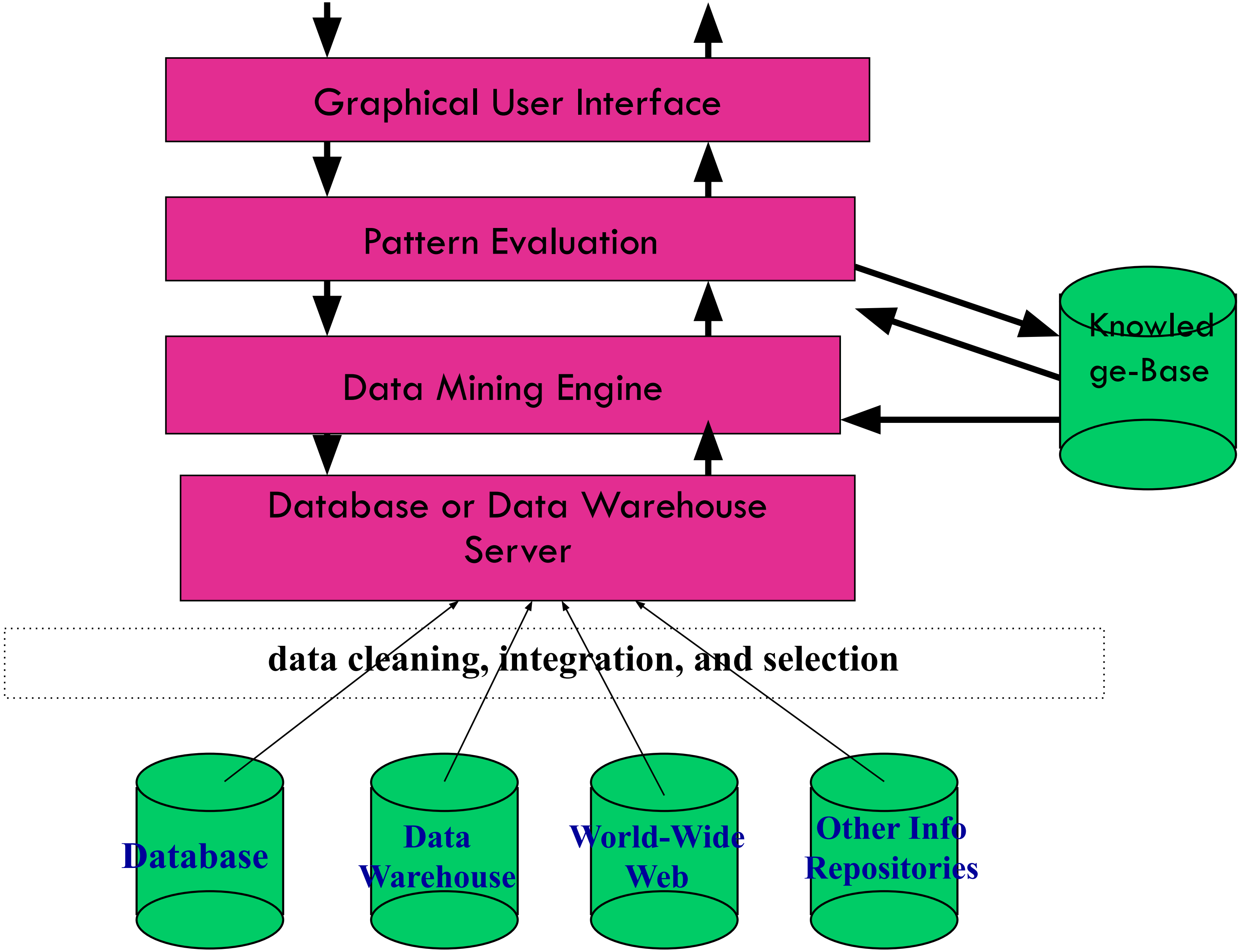
5. data mining (an essential process where intelligent methods are applied in order to extract data patterns.

6. pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)

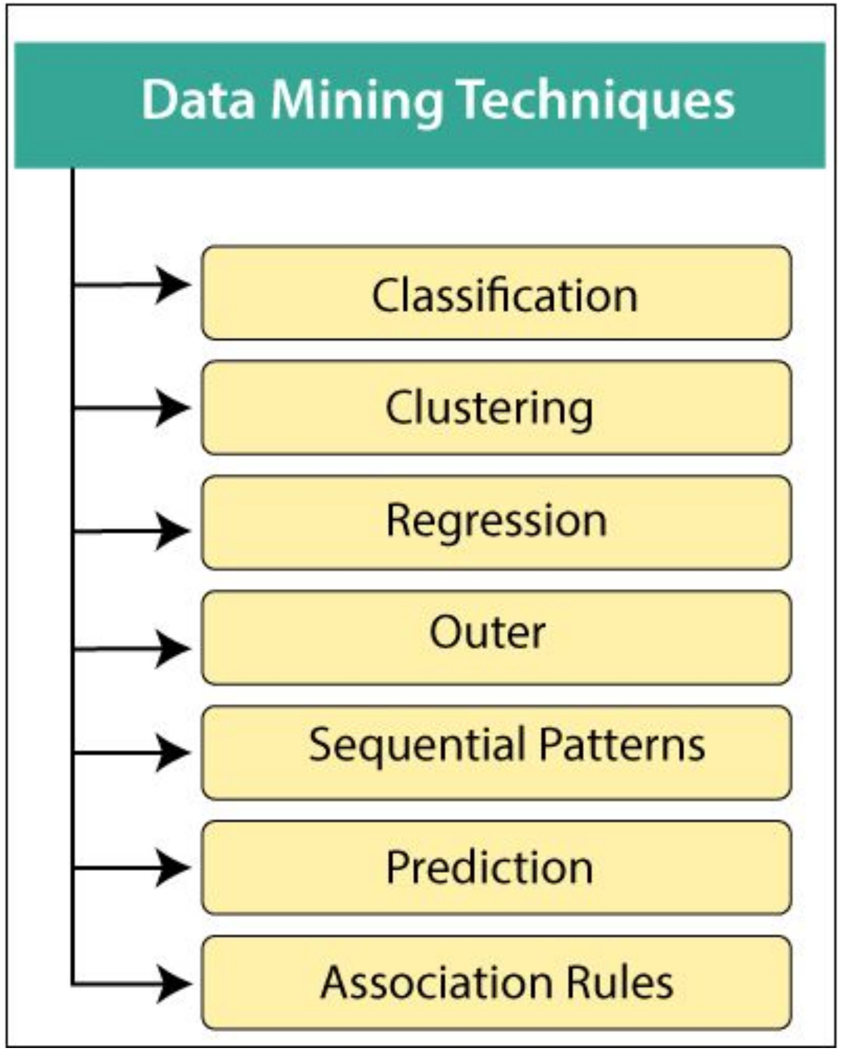
7. knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

Data mining is a core of knowledge discovery process

ARCHITECTURE: TYPICAL DATA MINING SYSTEM



WHAT KIND OF PATTERN CAN BE MINED? (DATA MINING FUNCTIONALITIES)



Data Mining Techniques

- Classification
- Clustering
- Regression
- Outer
- Sequential Patterns
- Prediction
- Association Rules

WHICH TECHNOLOGIES ARE USED?

Statistics

- Studies the collection , analysis, interpretation or explanation and presentation of data
- Statistical model are widely used to model data

Machine Learning

- Investigate how computers can learn based on data
- Computer is program to automatically learn to recognize complex pattern and make intelligent decision base on data
  - Supervised learning
  - Unsupervised learning
  - Semi-supervised learning

WHICH TECHNOLOGIES ARE USED?

Database systems and Data warehouse

- Focuses on the creation, maintenances and use of databases for organizations and end users
- Data warehouse consolidate data in multidimensional space

Information retrieval (IR)

- Science of searching a document or information which can be text or multimedia in a document which may reside on a web.

WHICH KIND OF APPLICATIONS ARE TARGETED?

Business Intelligence

- Provide historical, current, and predictive views of business operations

Web search engines

- It is a specialized computer server that searches for information on the web

MAJOR ISSUES IN DATA MINING

It is partition into five groups

1. Mining methodology
2. User interaction
3. Efficiency and scalability
4. Diversity of data types
5. Data mining and society

MAJOR ISSUES IN DATA MINING

1. Mining methodology issues

- Mining different kinds of knowledge in database
- Mining knowledge in multidimensional space
- Handling noisy or incomplete data
- Pattern evaluation

MAJOR ISSUES IN DATA MINING

2. User interaction issues

- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query languages and ad hoc data mining
- Presentation and visualization of data mining result

MAJOR ISSUES IN DATA MINING

3. Efficiency and scalability issues

- Efficiency and scalability of data mining algorithm
- Parallel, distributed and incremental mining algorithms

4. Diversity of database type issues

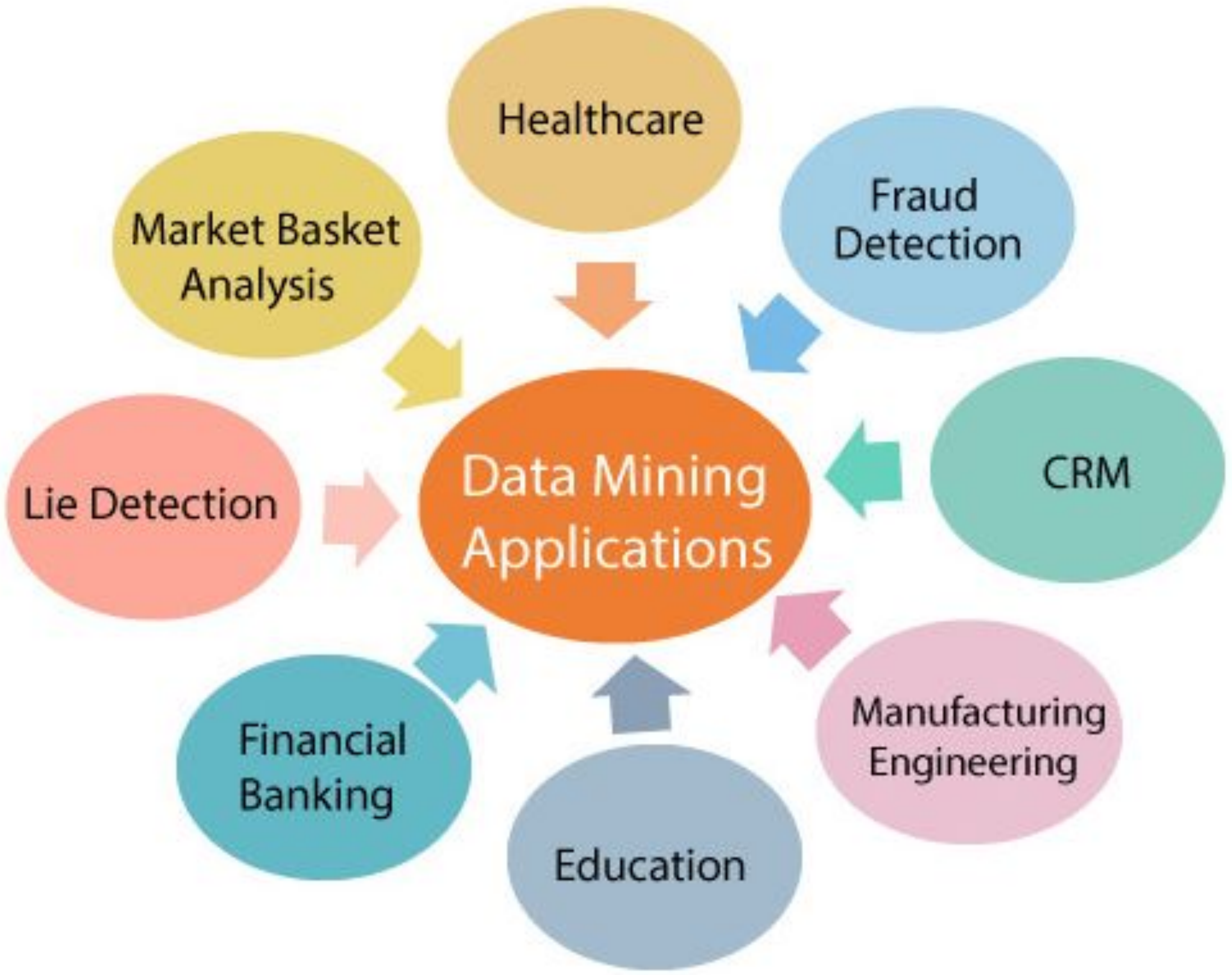
- Handling of relational and complex types of data
- Mining information from heterogeneous databases and global information system

MAJOR ISSUES IN DATA MINING

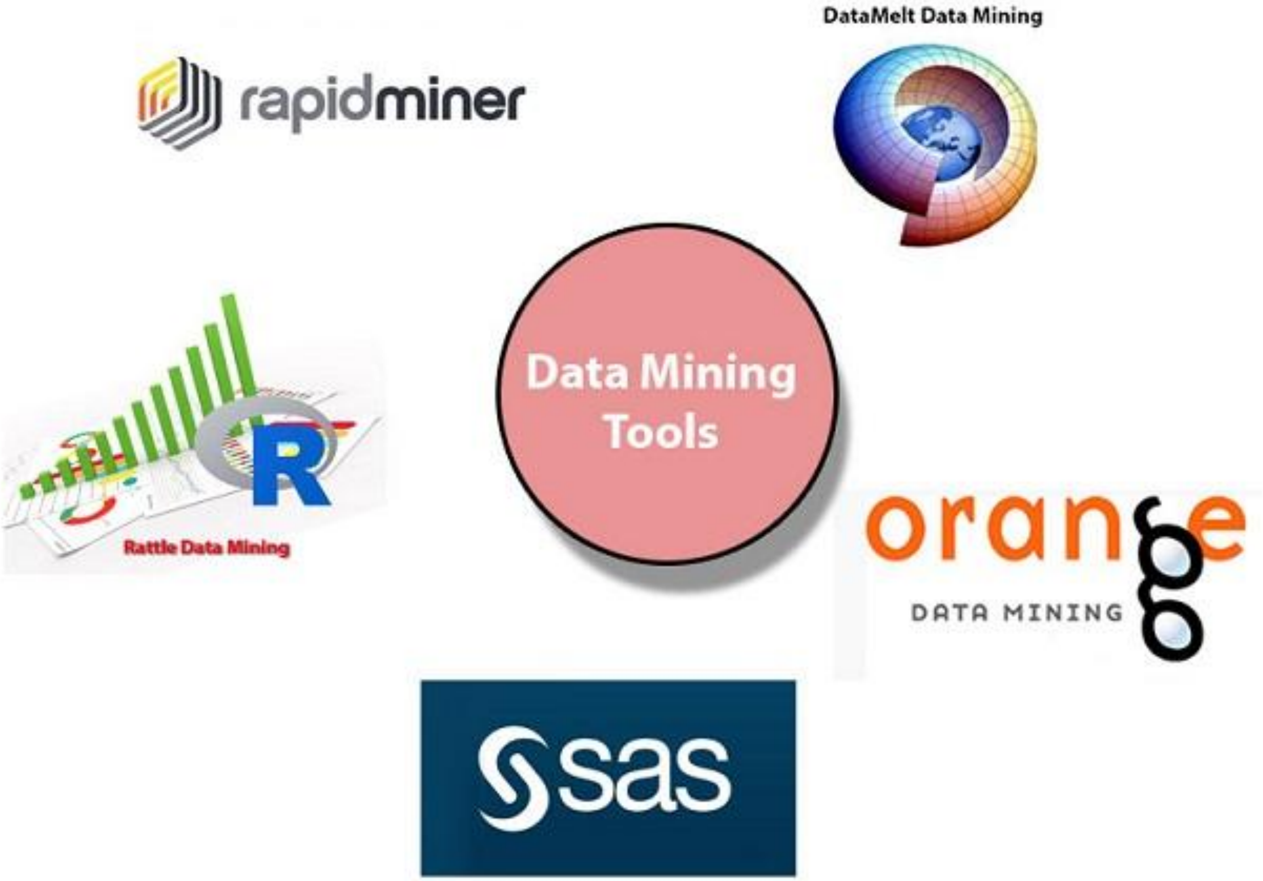
5. Data mining and Society

- Social impact of data mining
- Privacy preserving data mining
- Invisible data mining

DATA MINING APPLICATIONS



DATA MINING TOOLS



Data Mining Tools

Thank You!!!