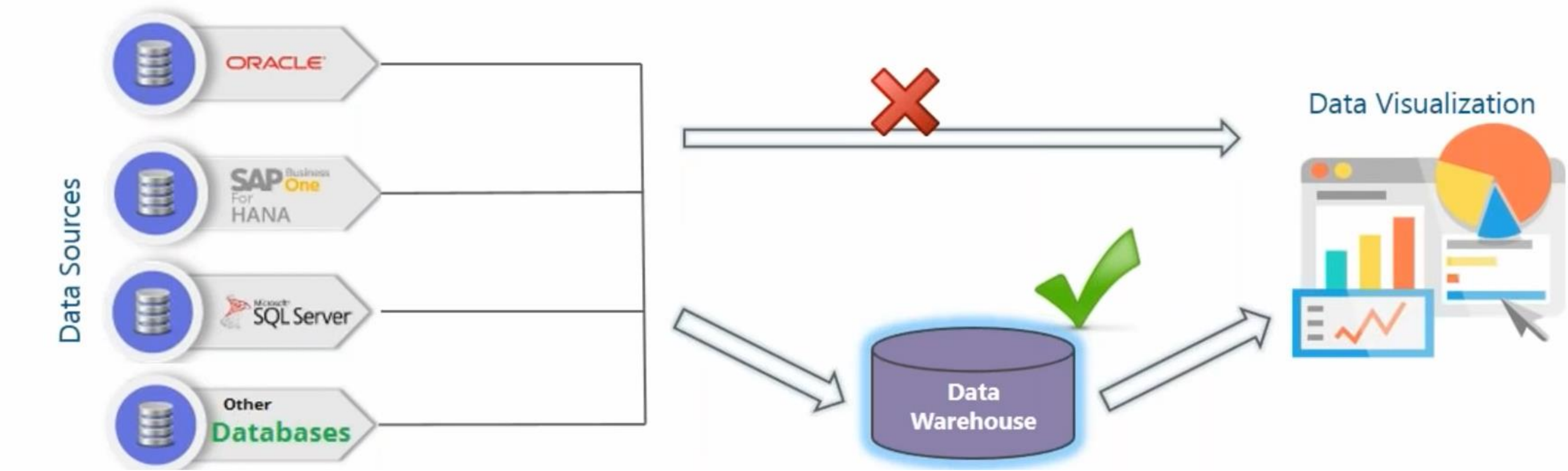
Data Mining for Business

Introduction to Data Warehouse

Dr. Shipra Maurya
Department of Management Studies
IIT (ISM) Dhanbad
Email: shipra@iitism.ac.in

Need for Data Warehouse

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

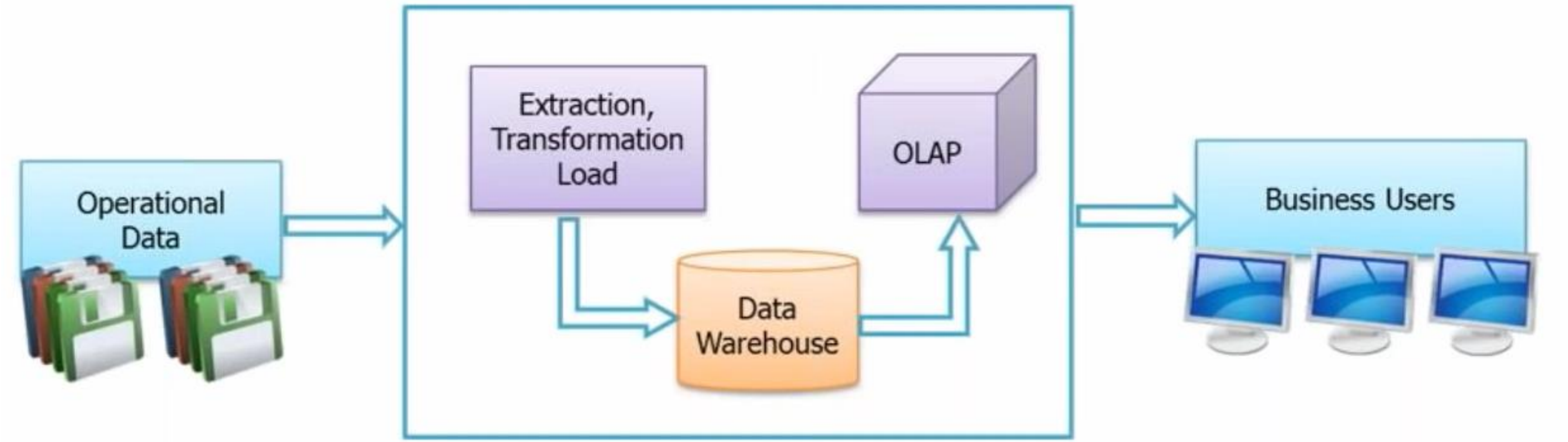


Note: It is not a product that a company can go and purchase, it needs to be designed based on company’s requirements

Image Source: Internet

Data Warehouse Introduction

- A data warehouse is a central repository where consolidated data from multiple databases is stored
- Data warehouse is maintained separately from an organization’s operational database
- End users access it whenever any information is needed
- Data warehouse is not loaded every time new data is added to database

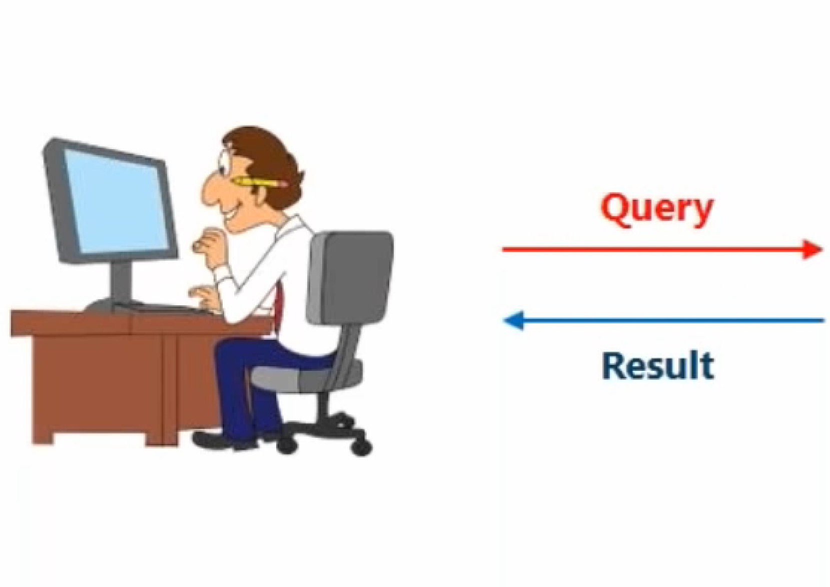


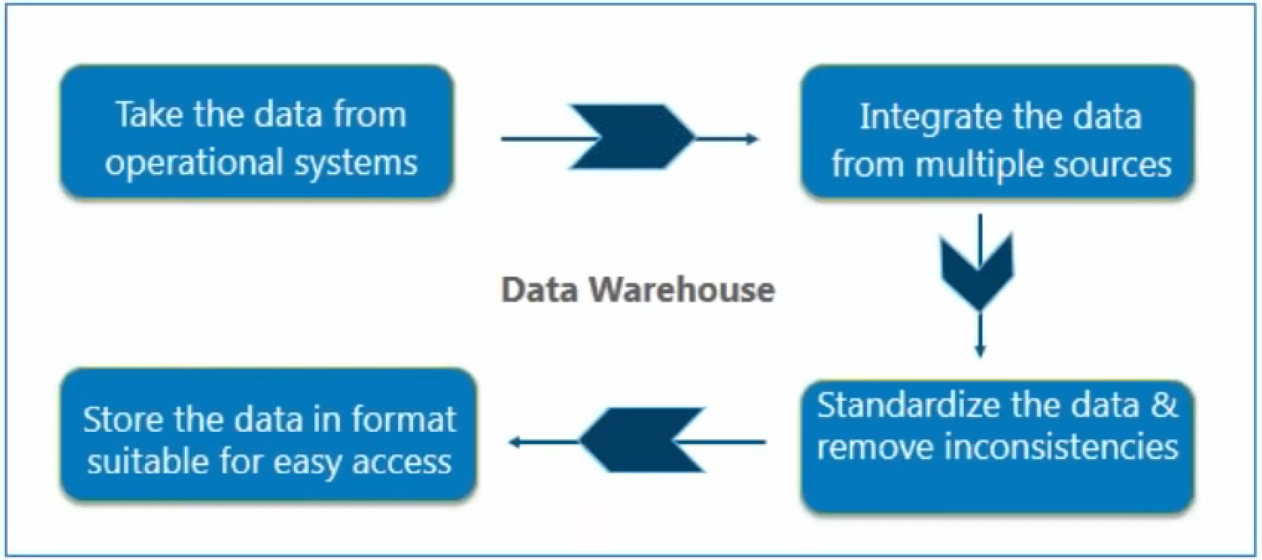
“Data Warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management’s decision-making process.” – Bill Inmon, Father of Data Warehousing

Image Source: Internet

Properties/Benefits of Data Warehouse

- **Subject-oriented-** Enable analysis of data about a particular subject or functional area (such as sales). Strategic questions can be answered
- **Integrated-** creates consistency among different data types from disparate sources.
- **Nonvolatile-** Once data is in a data warehouse, it is not updated or deleted.
- **Time-variant** – Data is stored as a series of snapshots, each representing a period
- Faster and more accurate





Take the data from operational systems

Integrate the data from multiple sources

Standardize the data & remove inconsistencies

Store the data in format suitable for easy access

Data Warehouse

Image Source: Internet

Key terminologies related to Data Warehouse

- OLTP vs OLAP
- ETL
- Data Mart
- Metadata

What is OLTP?

OLTP – Online Transaction Processing – supports transaction-oriented applications. It administers day-to-day transactions of an organization. Primary objective is data processing.

Example –

- A railway reservation server which records the transactions of a customer
- A bank server which records every time a transaction is made for a particular account
- A supermarket server which records every single product purchased at that market

What is OLAP?

OLAP – Online Analytical Processing

- supports analysis of data from different database systems at one time for business decisions. Primary objective is data analysis.
- OLAP databases are divided into one or more cubes called OLAP cubes stored on OLAP server. The cubes are designed in such a way that creating and viewing reports become easy
- Analytical Operations of OLAP

Example –

- Amazon analyzed purchases by its customers to come up with a personalized homepage with products which likely interest to their customer
- Bank manager wants to know how many customers are using bank ATM of his branch. This will help him deciding whether to continue the ATM or relocate it.

OLTP vs OLAP

OLTP (Database)	OLAP (Data Warehouse)
Contains current data	Contains historical data
Useful in running the business	Useful in analysing the business
Based on the Entity-Relationship model	Based on Star, Snowflake and Galaxy (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 (such as 100 MB to 1 GB)	Data warehouse size ranges from (such as 100 GB to 1 TB)
Used by data critical users like clerk & Database professionals	Used by data knowledge users like workers, managers and CEO
Allows thousands of users	Allows only hundreds of users

ETL - Extract, Transform and Load

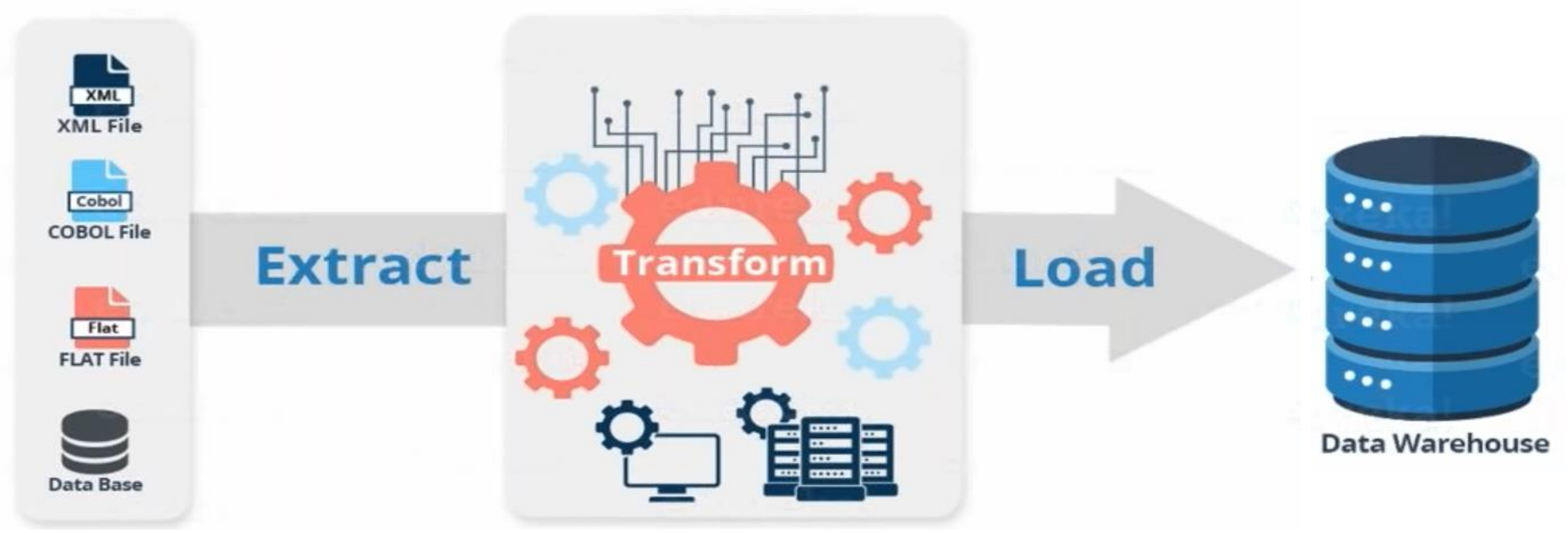
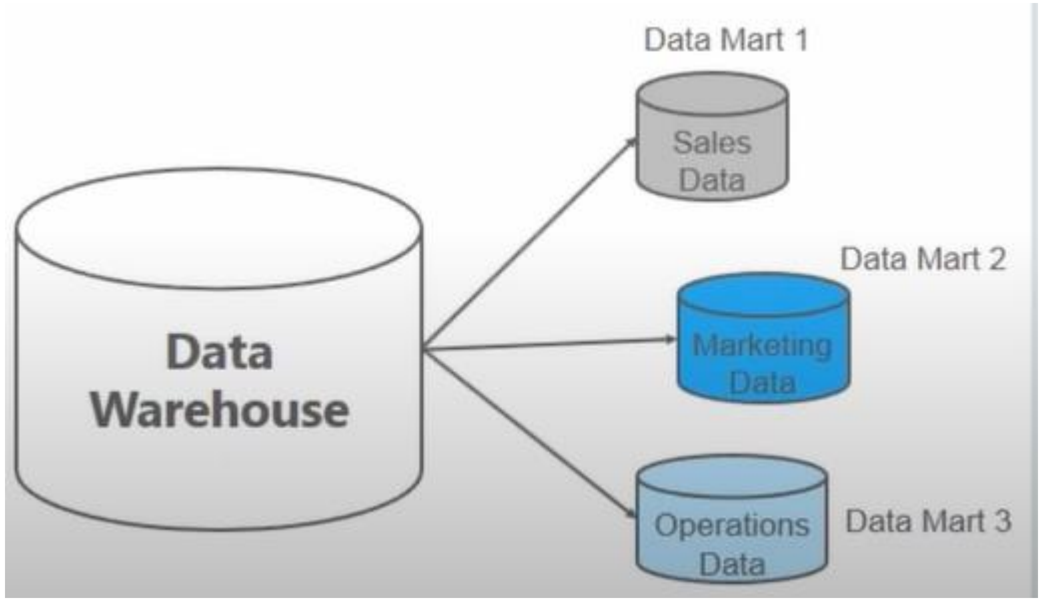


Image Source: Internet

Data Mart

- Data mart is a smaller version of the Data Warehouse which deals with a single subject

- Example – Sales data mart, Purchasing data mart etc.

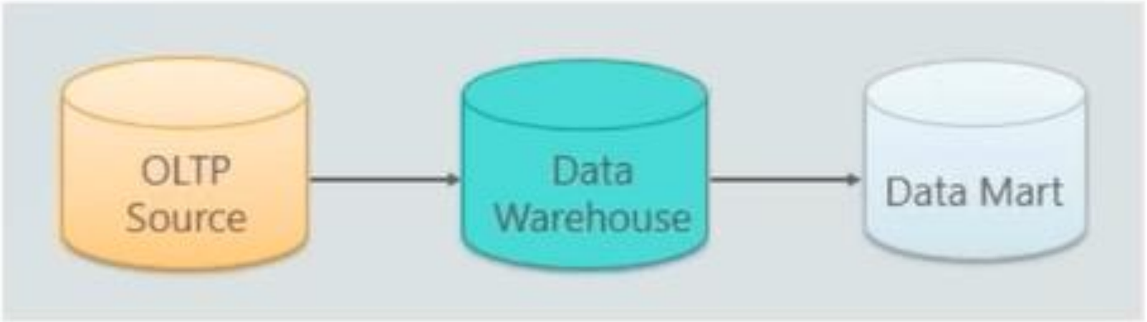


Data Marts	Data Warehouse
Department-wide data	Enterprise-wide data
Single subject area	Multiple subject areas
Limited data sources	Multiple data sources
Occupies limited memory	Occupies large memory
Shorter time to implement	Longer time to implement

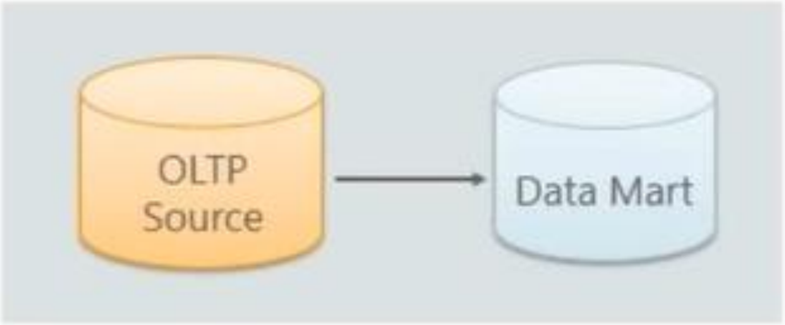
Image Source: Internet

Types of Data Mart

Dependent Data Mart



Independent Data Mart – suitable for small organizations or smaller groups within organization



Hybrid Data Mart – useful when you need ad hoc integration, such as after a new group or product is added to the organization

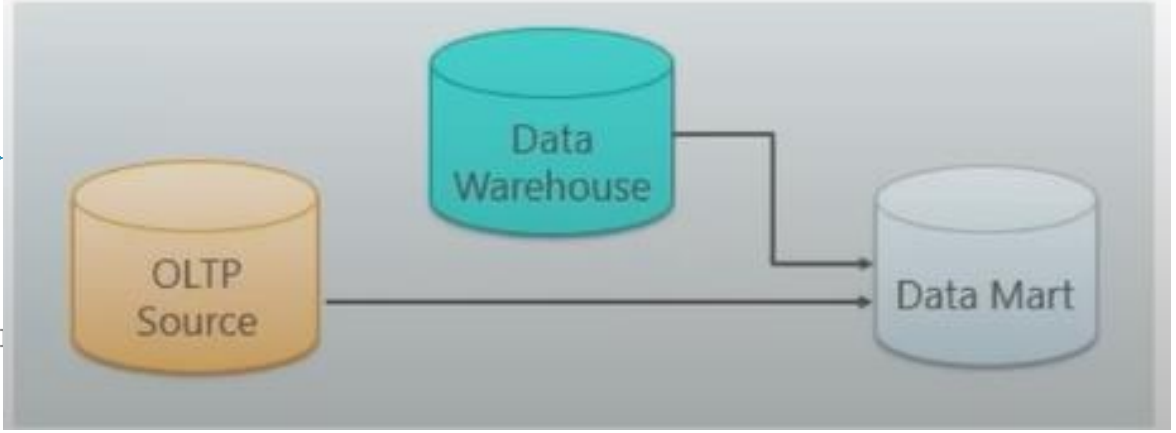


Image Source: Internet

Metadata

- Metadata is data about data. A library catalog and an index in the book are examples of metadata.
- Metadata is used for building, maintaining, managing, and using the data warehouses.
- Metadata includes the following:
  - The location and descriptions of warehouse systems and components.
  - Names, definitions, structures, and content of data-warehouse and end-users views.
  - Identification of authoritative data sources.
  - Integration and transformation rules used to populate data.
  - Integration and transformation rules used to deliver information to end-user analytical tools.
  - Metrics used to analyze warehouses usage and performance.
  - Security authorizations, access control list, etc.

Types of Metadata

- **Operational Metadata** – contains all the information about the operational data sources
- **Extraction and Transformation Metadata** – contains extraction frequencies, extraction methods and rules
- **End-user Metadata** – navigational map of the data warehouse

How does a Data Warehouse Work?

- It may contain multiple databases
- Within each database, data is organized into tabular format
- Data type is defined for each attribute/column
- Tables can be organized inside of schemas
- When data is ingested, it is stored in various tables described by the schema.
- Query tools use the schema to determine which data tables to access and analyze.

Users of Data Warehouse - Examples

Data Warehouse, Data Lakes and Databases

- A **data warehouse** is specially designed for data analytics, which involves reading large amounts of data to understand relationships and trends across the data.
- A **database** is used to capture and store data, such as recording details of a transaction.
- Unlike a data warehouse, a **data lake** is a centralized repository for all data, including structured, semi-structured, and unstructured.

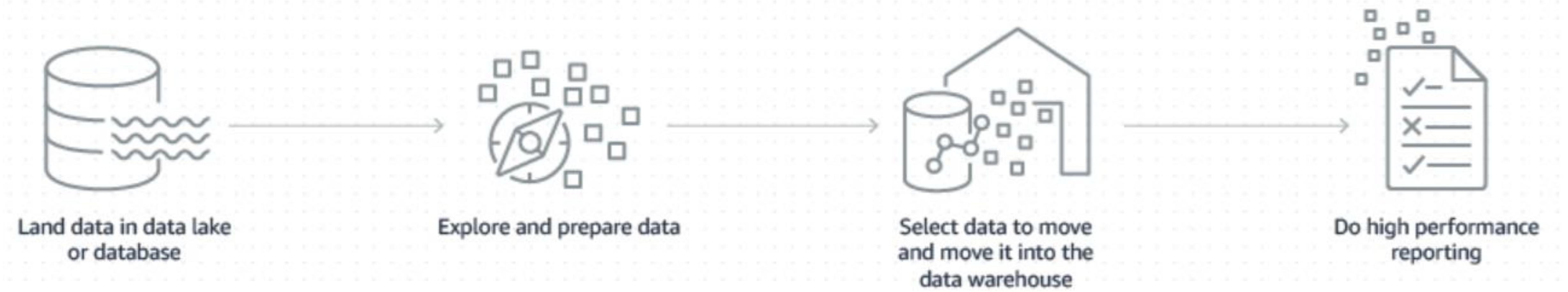

Land data in data lake or database
Explore and prepare data
Select data to move and move it into the data warehouse
Do high performance reporting

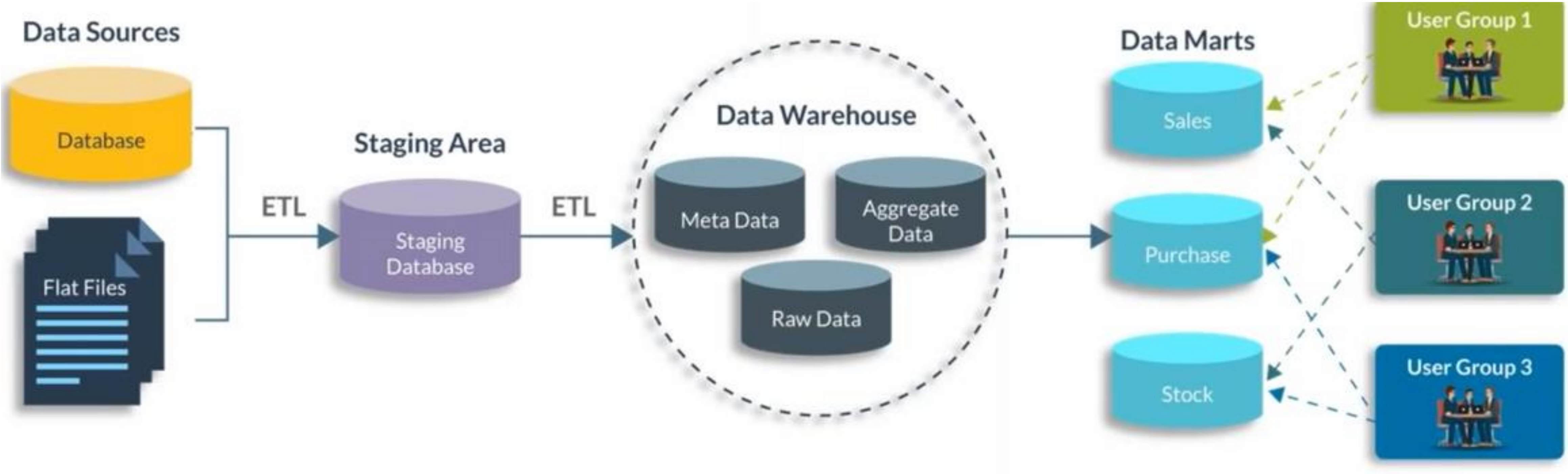
Image Source: Internet

Do Business need a Data Lake?

- Data Lake –
  - store an abundance of disparate, unfiltered data to be used later for a particular purpose
  - The structure, integrity, selection, and format of the various datasets is derived at the time of analysis by the person doing the analysis
  - Organizations choose data lake when they need low-cost storage for unformatted, unstructured data from multiple sources that they intend to use for some purpose in the future

- Data Warehouse –
  - specifically intended to analyze data
  - When organizations need advanced data analytics or analysis that draws on historical data from multiple sources across their enterprise, a data warehouse is likely the right choice

Data Warehouse 3-tier Architecture (1/2)



- Bottom Tier
- Middle Tier
- Top Tier

Image Source: Internet

Data Warehouse 3-tier Architecture (2/2)

- Data is stored in two different ways:
  - Data that is accessed frequently is stored in very fast storage like SSD drives
  - Data that is infrequently accessed is stored in a cheap object store, like Amazon S3.

- The data warehouse will automatically make sure that frequently accessed data is moved into the "fast" storage so query speed is optimized.

Cloud Data Warehouse (1/2)

- A cloud data warehouse uses the cloud to ingest and store data from disparate data sources.

- Originally, data warehouses were built with on-premises servers

On-premises data warehouses:

- Advantages: improved governance, security, data sovereignty
- Disadvantages: less elastic, managing is little complex

Cloud data warehouses:

- Advantages: elastic, ease of management, cost savings as pay-as-you-go model
- Disadvantages: less control

Cloud Data Warehouse (2/2)

Autonomous data warehouse:

- The most recent iteration of the data warehouse is the autonomous data warehouse
- An as-a-service autonomous data warehouse in the cloud requires no human-performed database administration, hardware configuration or management, or software installation
- The autonomous data warehouse removes complexity, speeds deployment, and frees up resources so organizations can focus on activities that add value to the business
- Eg. Oracle Autonomous Data Warehouse

Data Warehouse Tools (1/2)

Oracle –

- industry-leading database
- offers a wide range of choice of data warehouse solutions for both on-premises and in the cloud

Amazon RedShift –

- simple and cost-effective tool to analyze all types of data using standard SQL and existing BI tools

MarkLogic –

- makes data integration easier and faster using an array of enterprise features
- It can query different types of data like documents, relationships, etc.

Data Warehouse Tools (2/2)

- Microsoft Azure
- Google BigQuery
- Snowflake
- Micro Focus Vertica
- Amazon DynamoDB
- PostgreSQL
- Teradata
- Amazon RDS
- IBM DB2 Warehouse
- Cloudera

Thank you!

You can reach me on :

Email : shipra@iitism.ac.in

LinkedIn : https://www.linkedin.com/in/shipra-maurya1205/?originalSubdomain=in