SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

18CSE487T

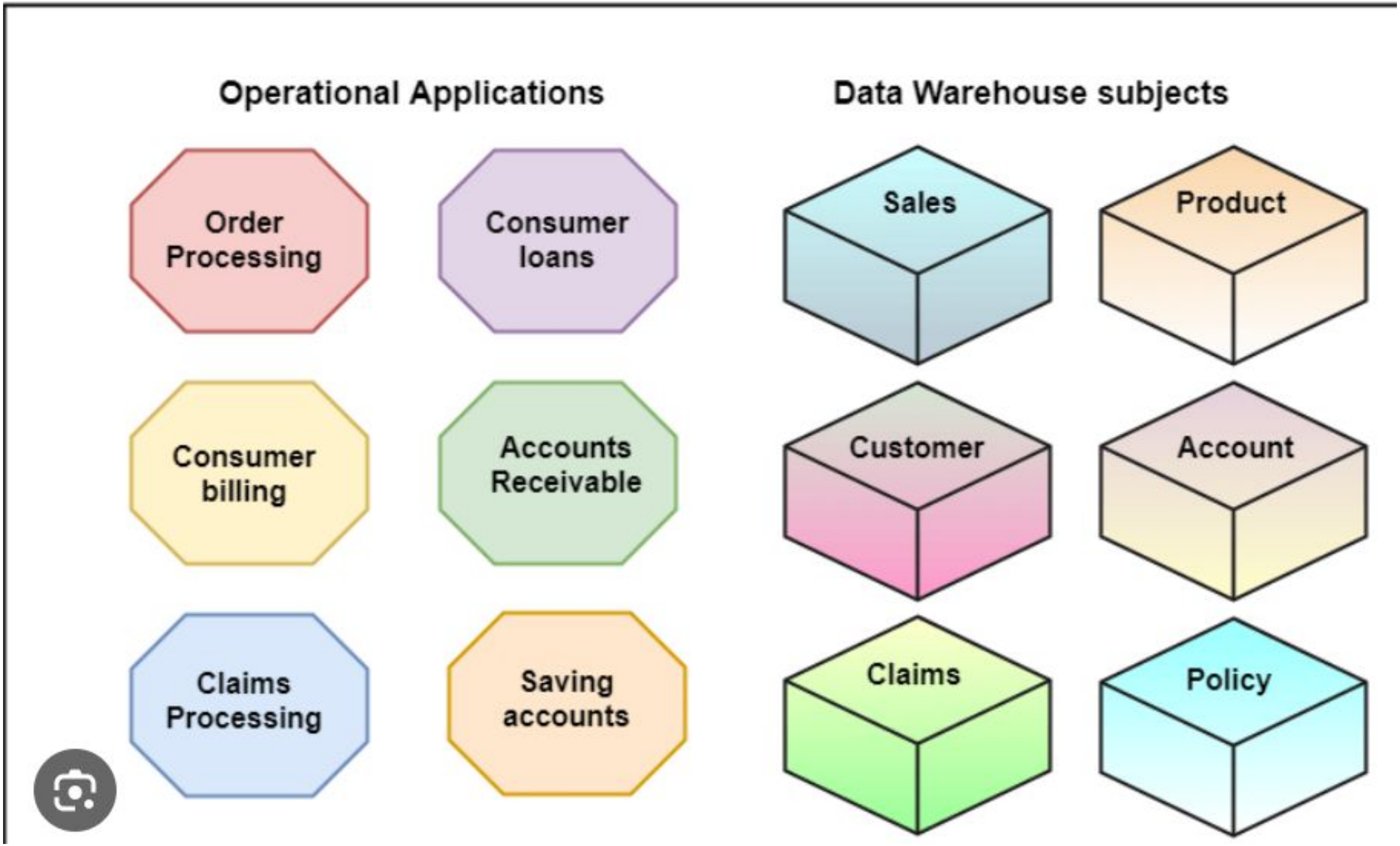
DATA WAREHOUSING AND ITS APPLICATIONS

UNIT-1

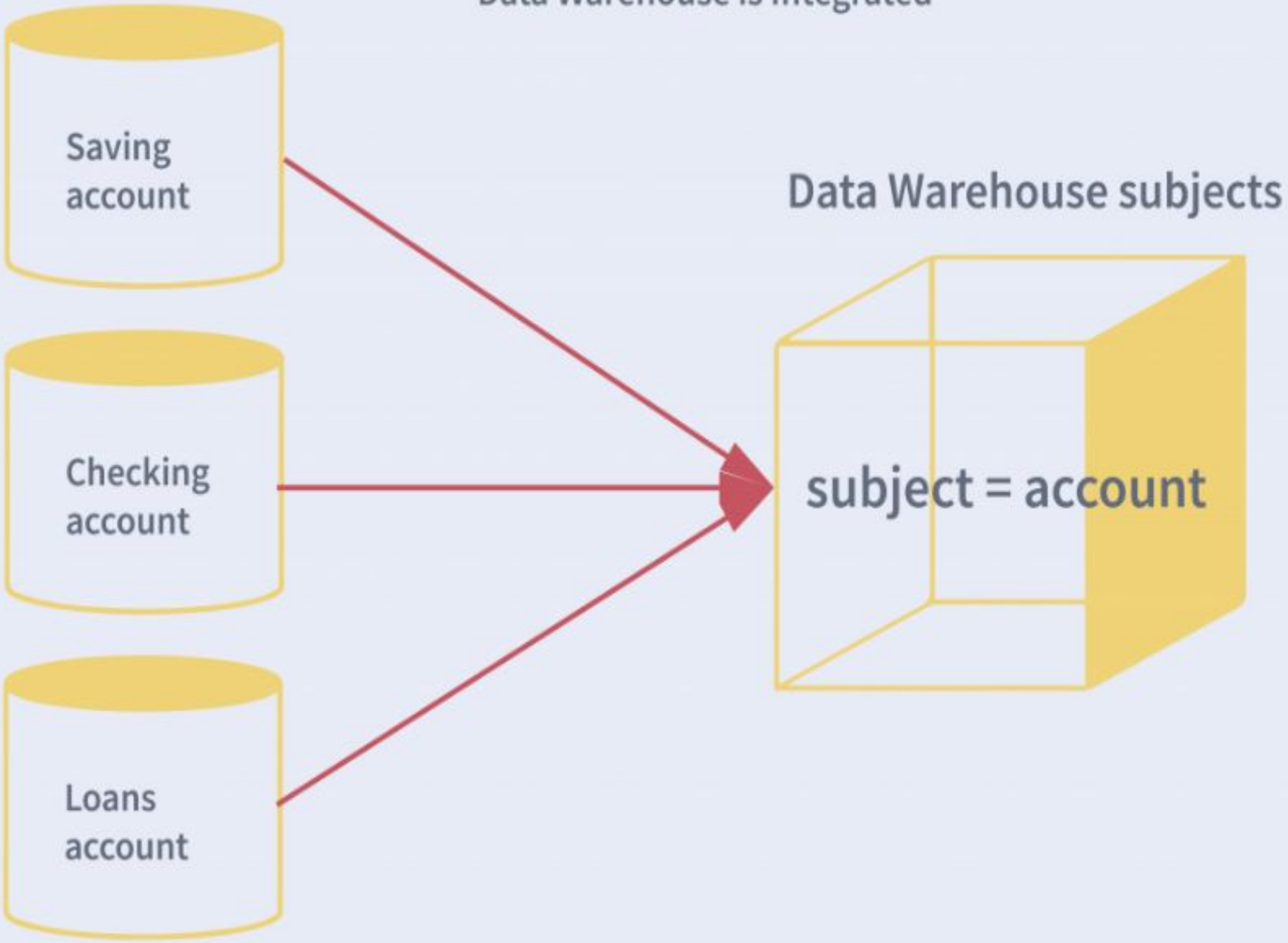
Data warehouse

- A subject-oriented, integrated, time-variant, non-updatable collection of data used in support of management decision- making processes
  - ***Subject-oriented:*** e.g. customers, patients, students, products
  - ***Integrated:*** consistent naming conventions, formats, encoding structures; from multiple data sources
  - ***Time-variant:*** can study trends and changes
  - ***Non-updatable:*** read-only, periodically refreshed
- Data Mart
  - A data warehouse that is limited in scope

Data Warehouse is Subject-Oriented



Data Warehouse is Integrated



Saving account

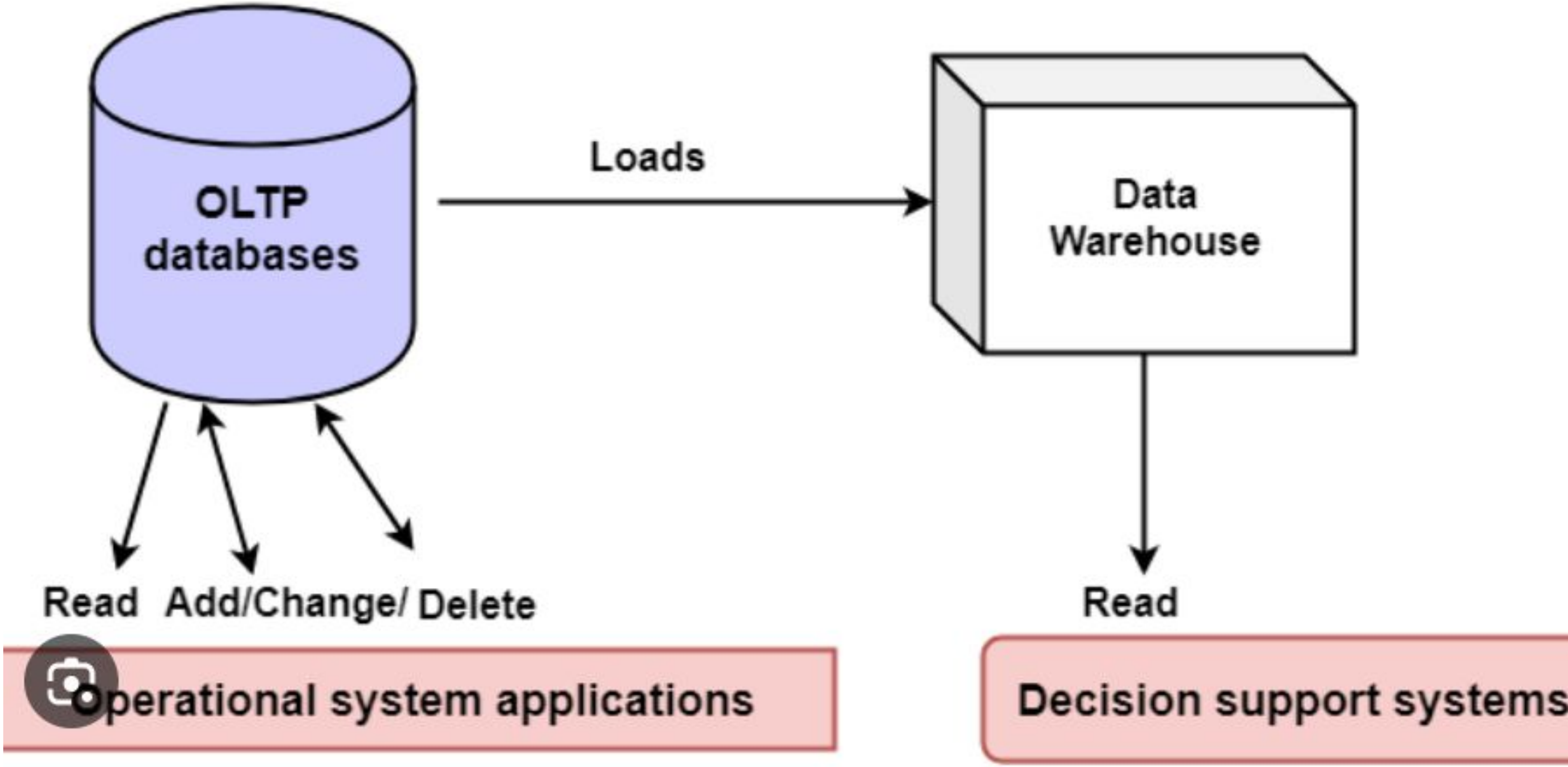
Checking account

Loans account

Data Warehouse subjects

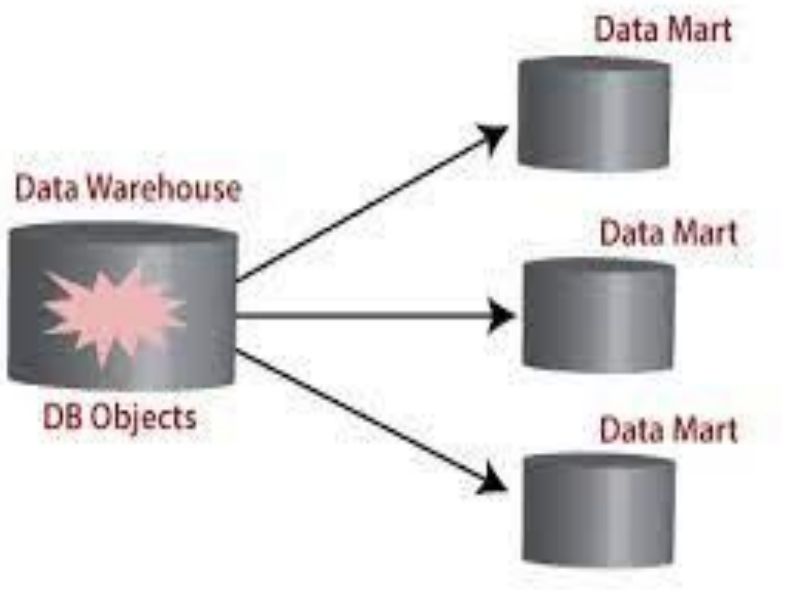
subject = account

Non-Volatile

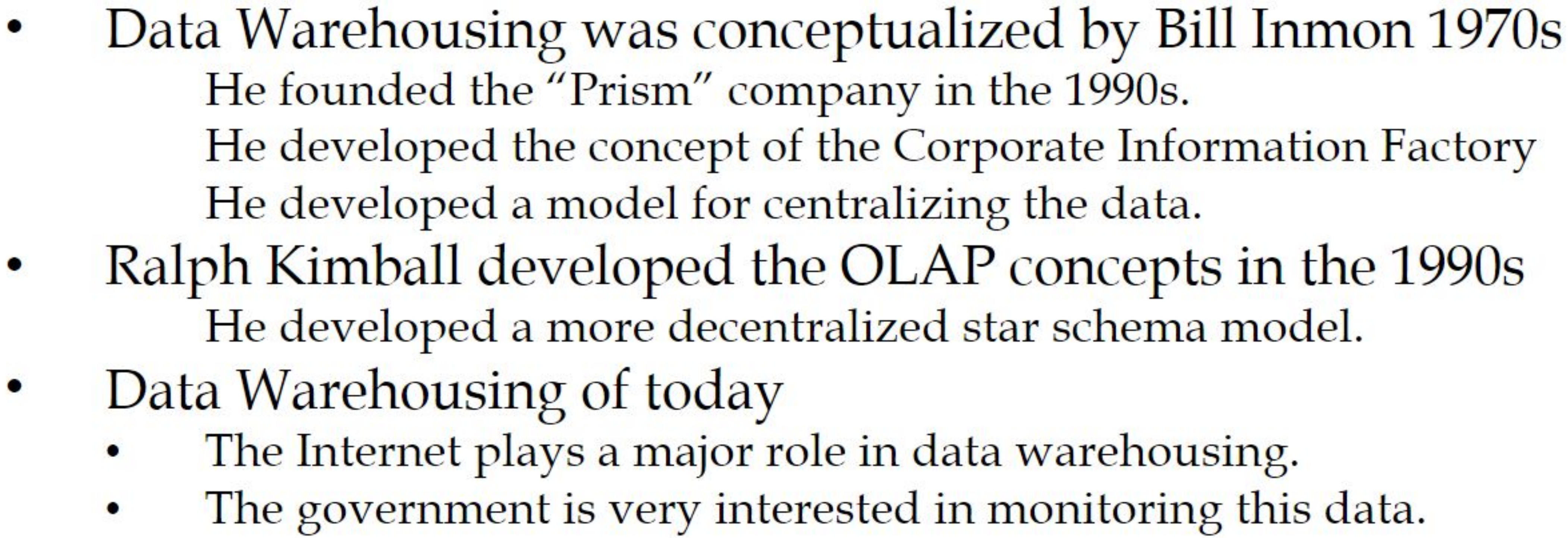


Data mart

- A data mart is a data storage system that contains information specific to an organization's business unit.
- It contains a small and selected part of the data that the company stores in a larger storage system.



History of Data warehouse



- Data Warehousing was conceptualized by Bill Inmon 1970s

  He founded the "Prism" company in the 1990s.

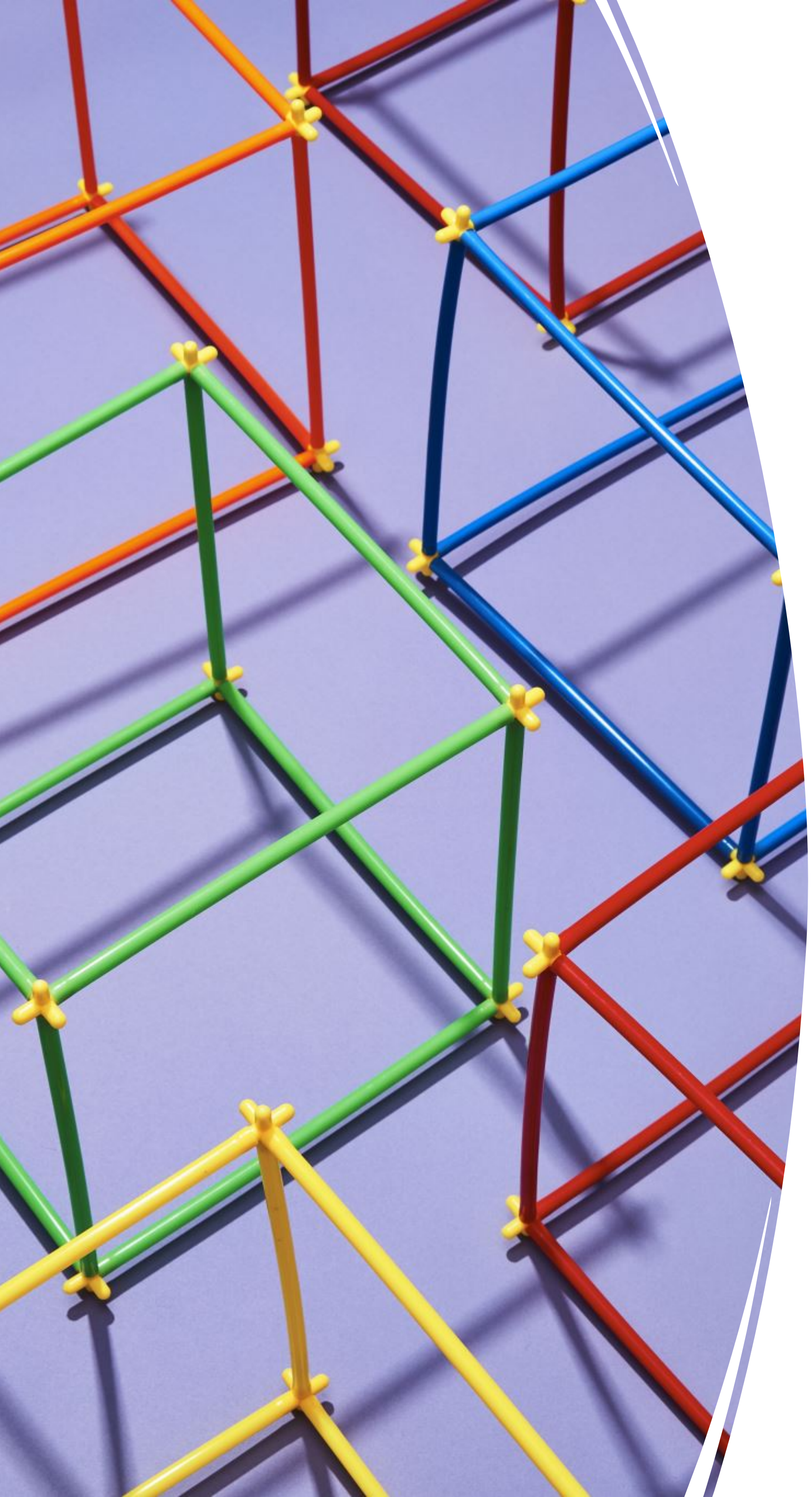
  He developed the concept of the Corporate Information Factory

  He developed a model for centralizing the data.

- Ralph Kimball developed the OLAP concepts in the 1990s

  He developed a more decentralized star schema model.

- Data Warehousing of today
  - The Internet plays a major role in data warehousing.
  - The government is very interested in monitoring this data.



History Leading to Data Warehousing

- Improvement in database technologies, especially relational DBMSs
- Advances in computer hardware, including mass storage and parallel architectures
- Emergence of end-user computing with powerful interfaces and tools
- Advances in middleware, enabling heterogeneous database connectivity
- Recognition of difference between operational and informational systems

Need for Data Warehousing

- Integrated, company-wide view of high-quality information (from disparate databases)
- Separation of operational and informational systems and data (for improved performance)
- Online Analytical Processing
- Decision Making Systems
- Data mining

Issues with Company-Wide View

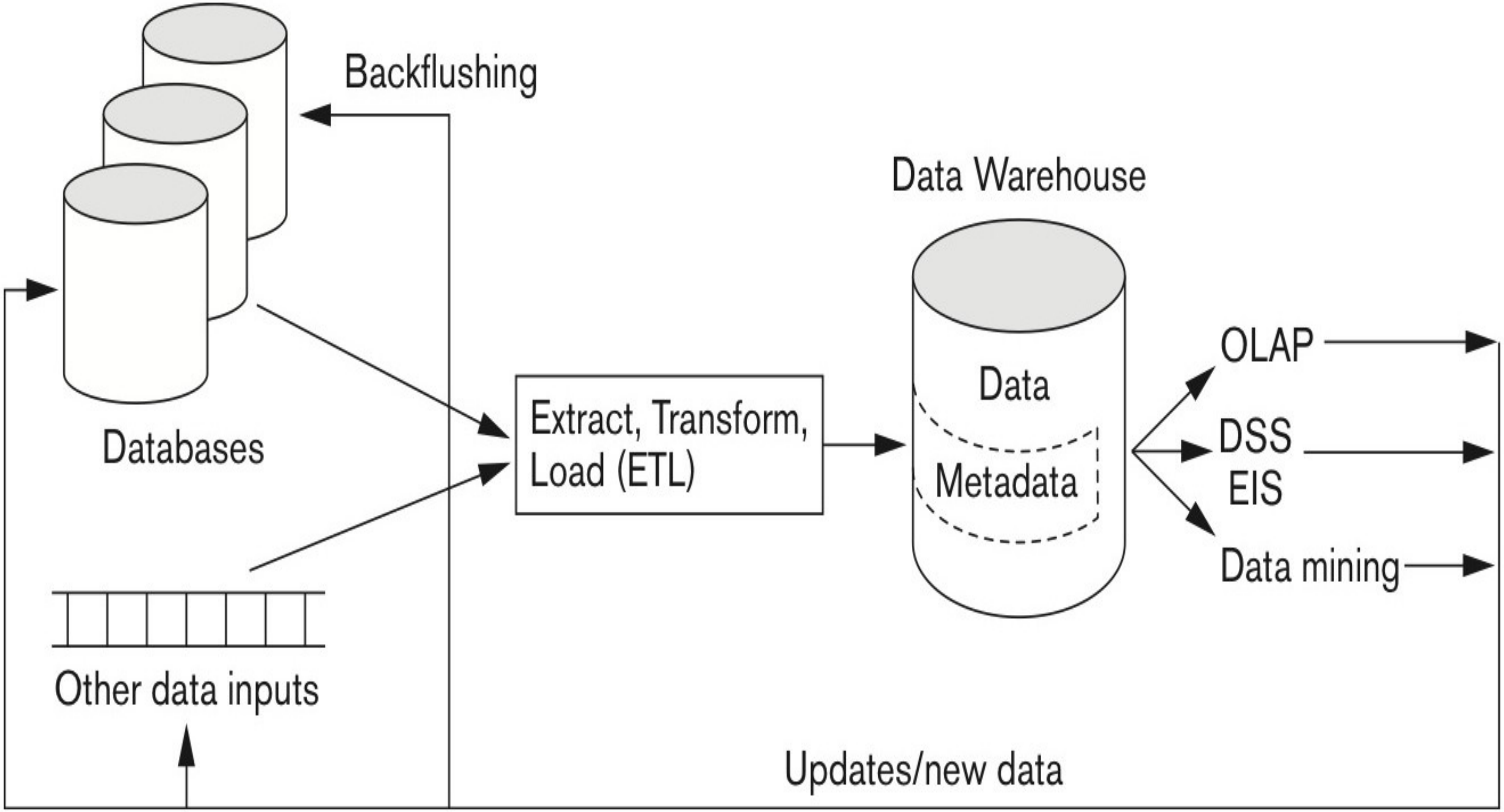
- Inconsistent key structures
- Synonyms
- Free-form vs. structured fields
- Inconsistent data values
- Missing data

Datawarehouse Architecture

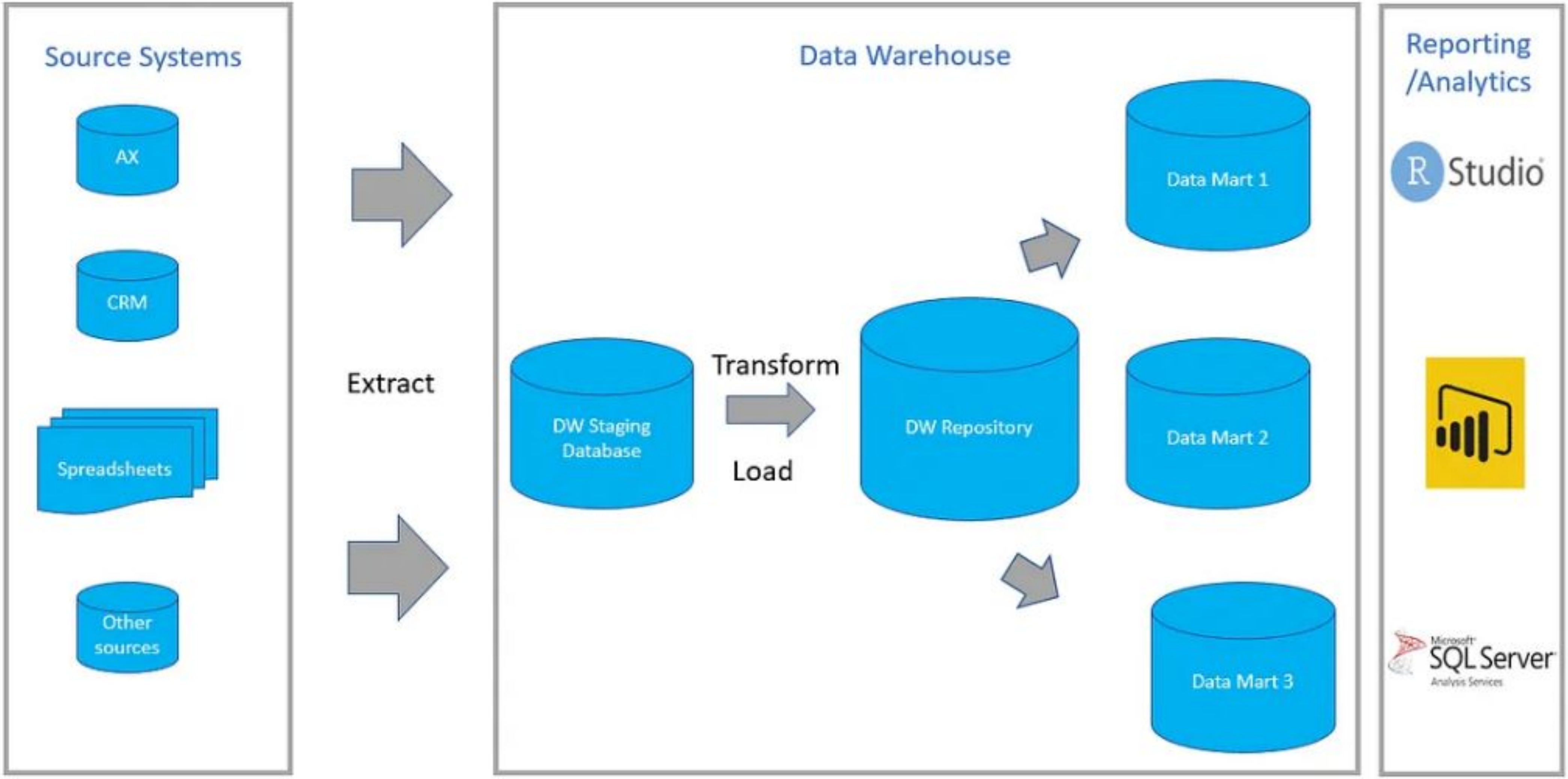
Data Warehouse processing involves

- Data Sources
- ETL (Extract, Transform, Load)
- OLAP – Data Analytics
- Data Mining

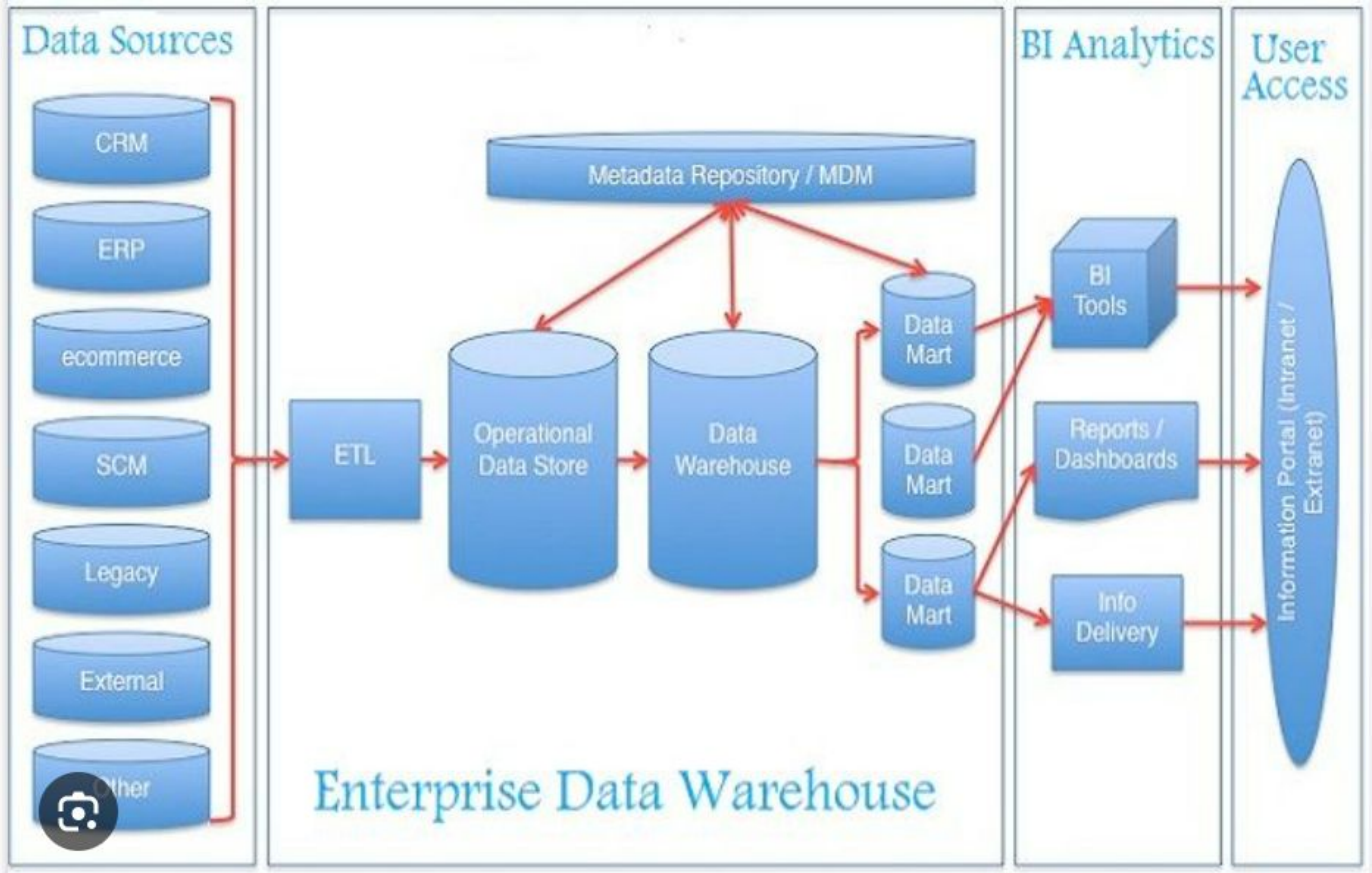
Datawarehouse Architecture (Common)



Datawarehouse Architecture



Datawarehouse Architecture



The major components of a data warehousing process

- **Data sources:** internal, external (data provider), OLAP, ERP, Web data
- **Data extraction and loading:** using custom-written or commercial software called (ETL) and loaded into a staging area to be transformed and cleansed, then loaded into the warehouse
- **Metadata:** to ease indexing and search
- **Middleware tools:** to enable access to DW. It includes data mining tools, OLAP, reporting tools, and data visualization tools.

The major components of a data warehousing process

- EIS- Executive Information System

Works closely with data warehouses in monitoring business performance as well as identifying problems

- DSS- Decision Support System

computer program that makes decisions based on data from

internal databases or external databases

Extraction, transformation, and load (ETL)

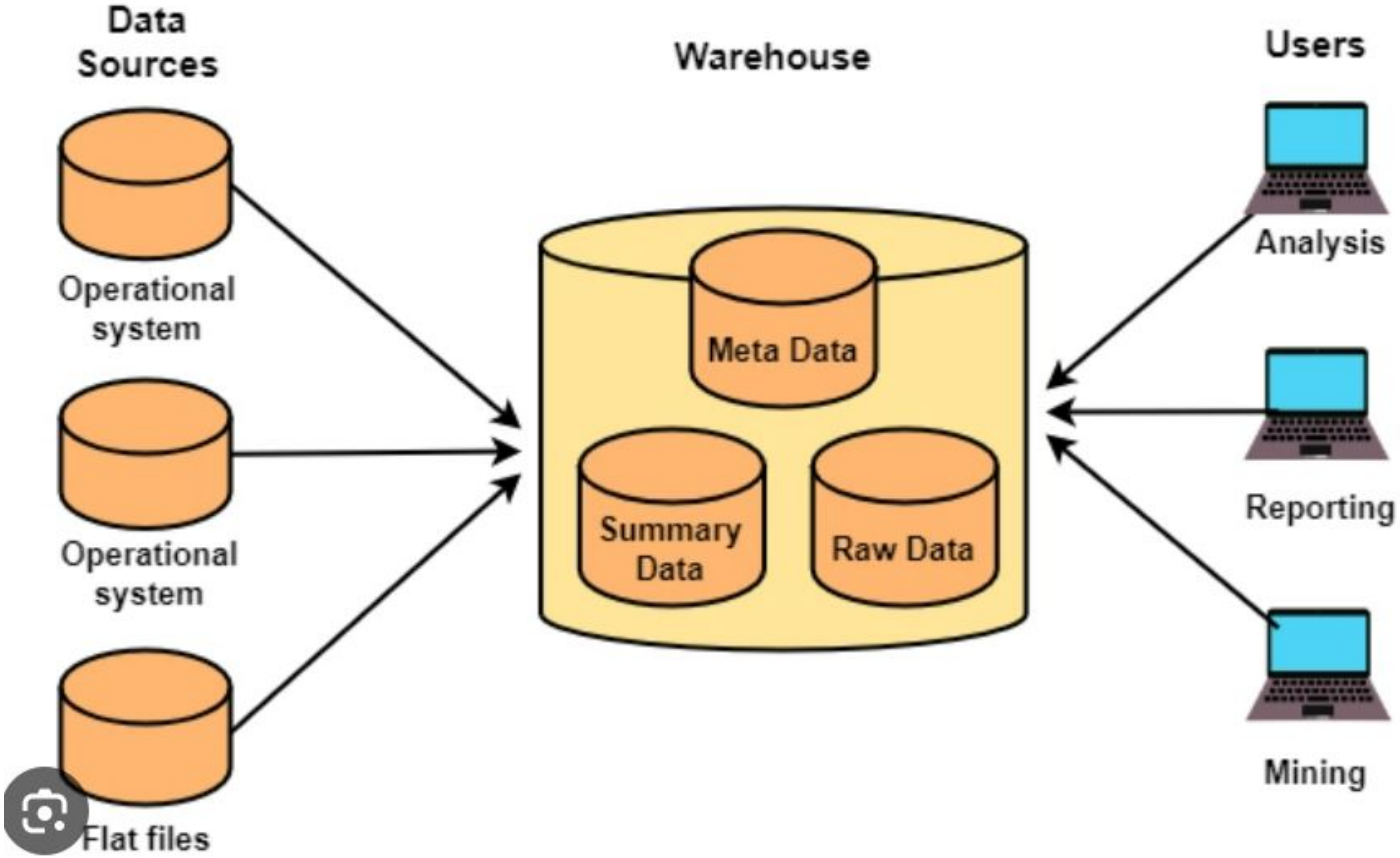
A data warehousing process that consists of:

- Extraction (i.e., reading data from a database),
- Transformation (i.e., converting the extracted data from its previous form into the form in which it needs to be so that it can be placed into a data warehouse or simply another database), and
- Load (i.e., putting the data into the data warehouse)
- During extraction process, the input files are written to a set of staging tables, to facilitate the load process.

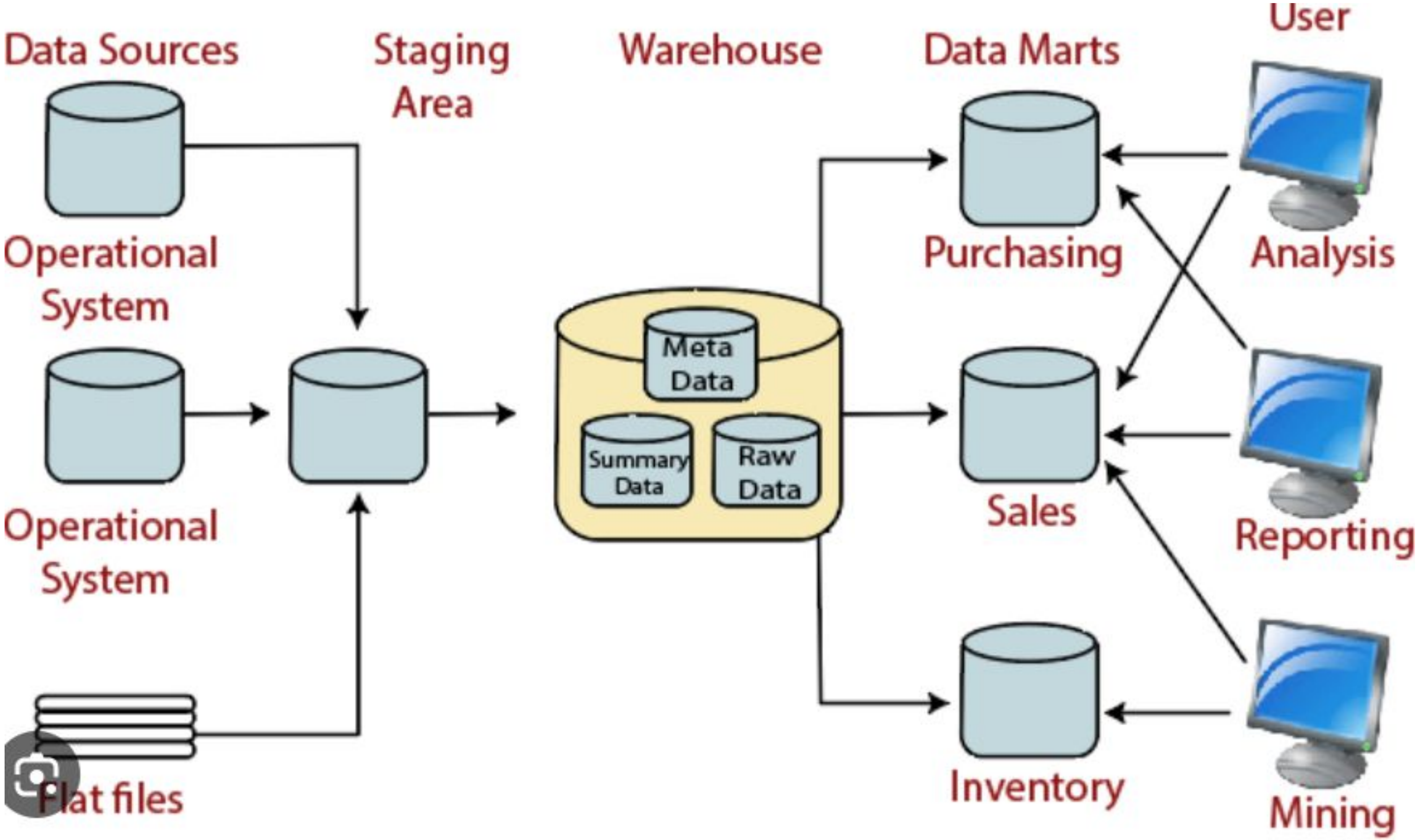
Datawarehouse Architecture types

- Data Warehouse Architecture: Basic
- Data Warehouse Architecture: With Staging Area
- Data Warehouse Architecture: With Staging Area and Data Marts

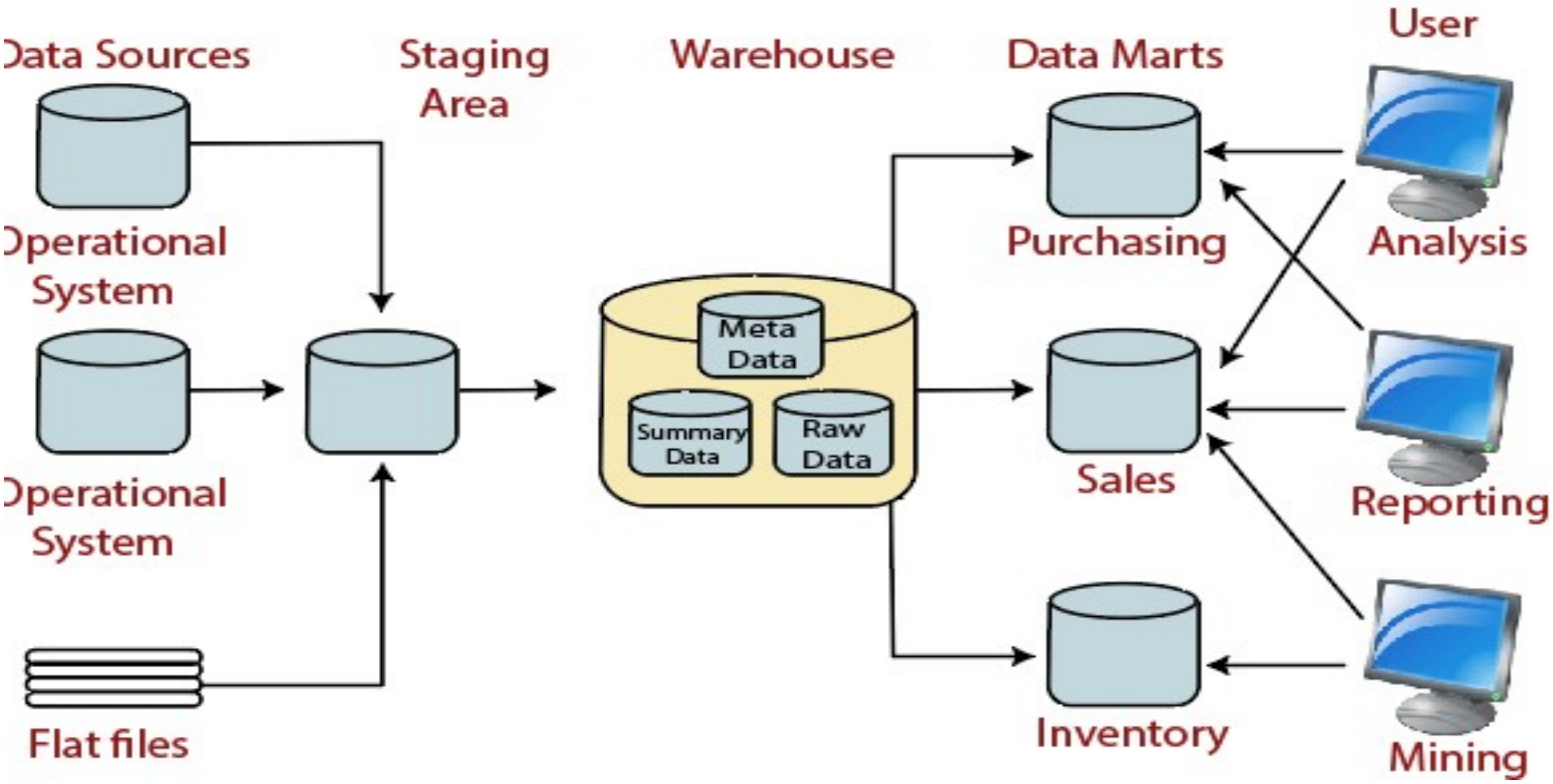
Data Warehouse Architecture: Basic



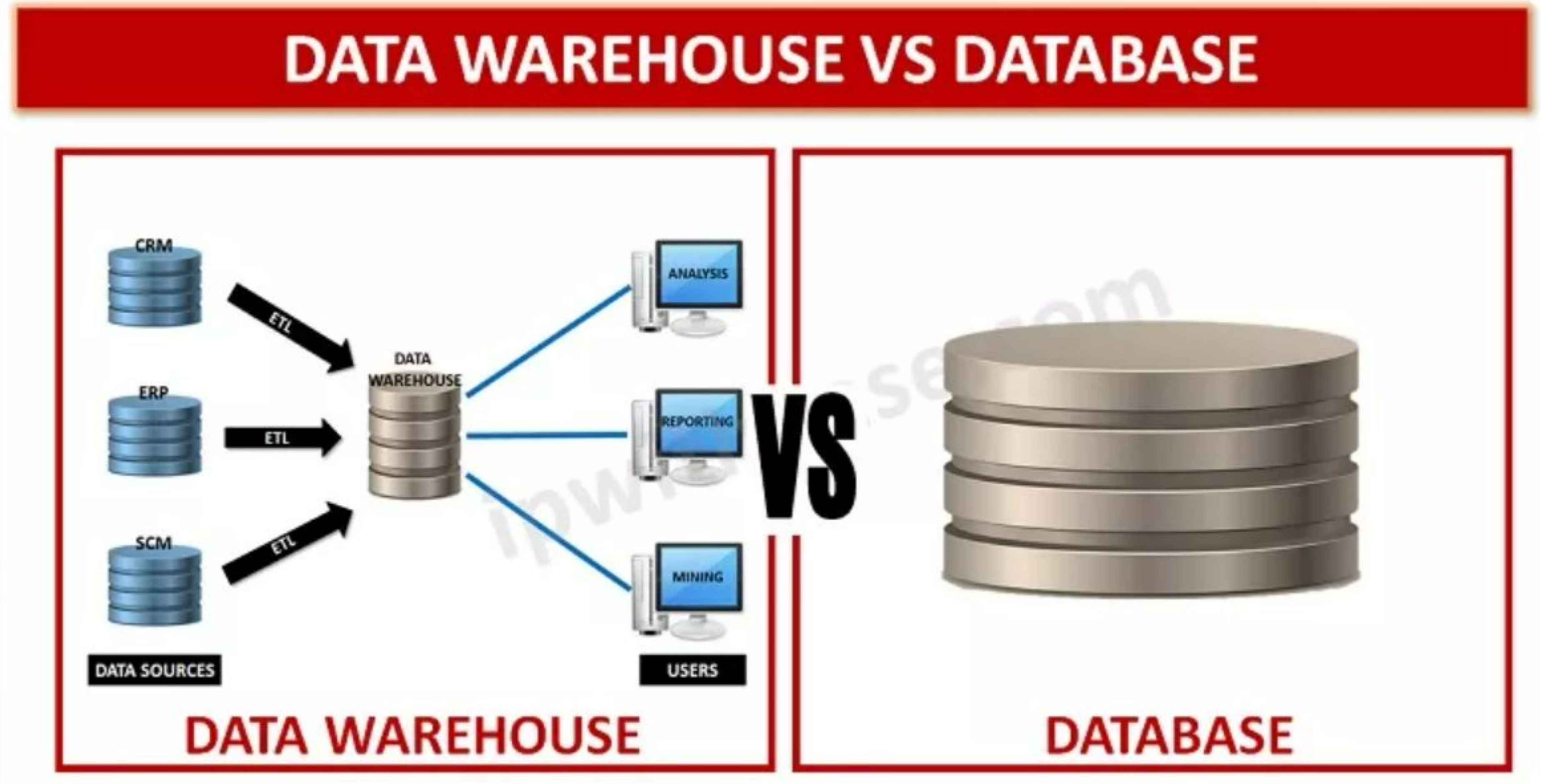
Data Warehouse Architecture: With Staging Area



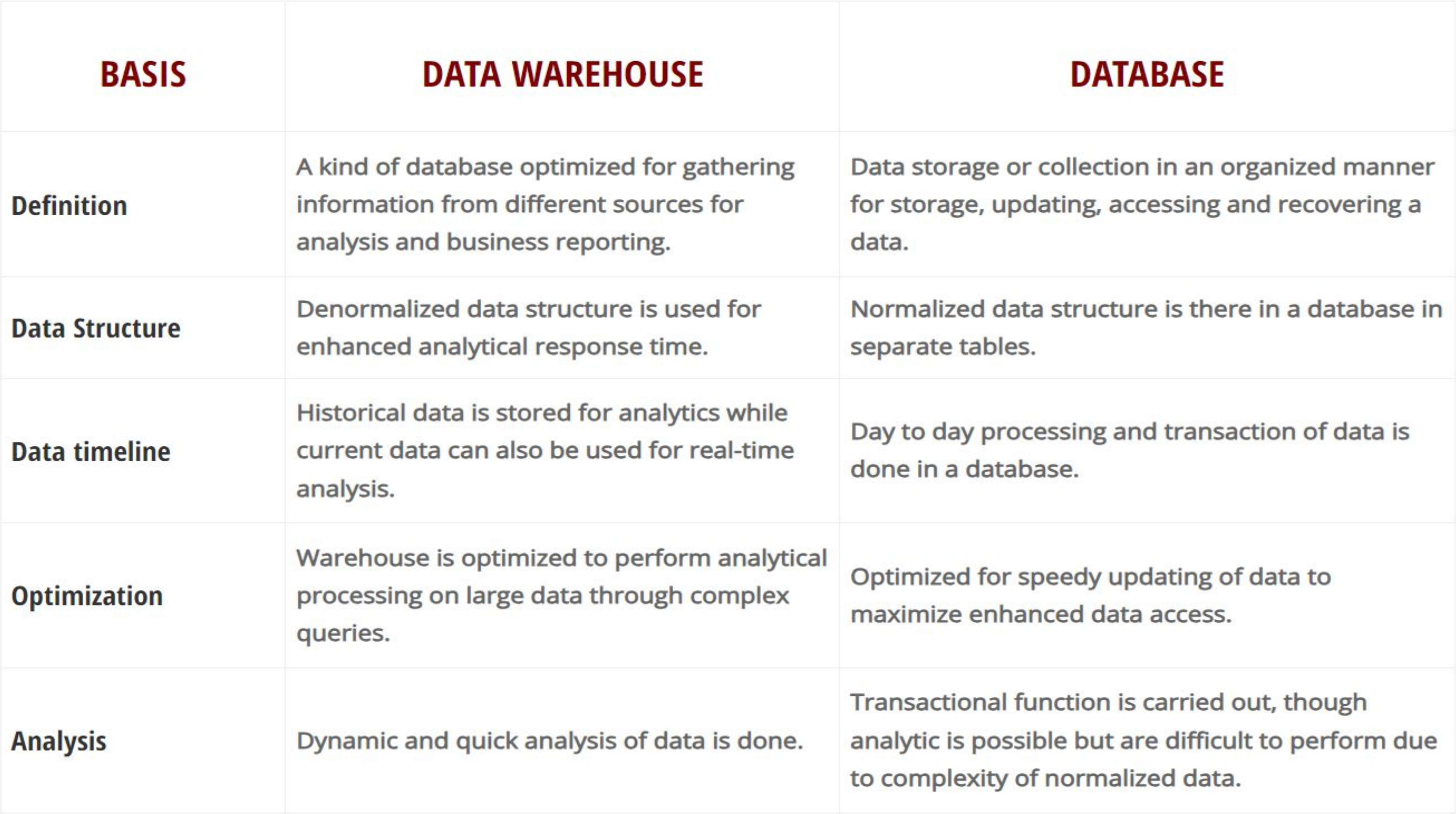
Data Warehouse Architecture: With Staging Area and Data Marts



Difference between Database and Data warehouse



Difference between Database and Data warehouse



BASIS	DATA WAREHOUSE	DATABASE
Definition	A kind of database optimized for gathering information from different sources for analysis and business reporting.	Data storage or collection in an organized manner for storage, updating, accessing and recovering a data.
Data Structure	Denormalized data structure is used for enhanced analytical response time.	Normalized data structure is there in a database in separate tables.
Data timeline	Historical data is stored for analytics while current data can also be used for real-time analysis.	Day to day processing and transaction of data is done in a database.
Optimization	Warehouse is optimized to perform analytical processing on large data through complex queries.	Optimized for speedy updating of data to maximize enhanced data access.
Analysis	Dynamic and quick analysis of data is done.	Transactional function is carried out, though analytic is possible but are difficult to perform due to complexity of normalized data.

Characteristics of Data Warehouses

The following characteristics of Data Warehouses were identified:

- Multidimensional conceptual view
- Unlimited dimensions and aggregation levels
- Unrestricted cross-dimensional operations
- Client-server architecture
- Multiuser support
- Accessibility
- Transparency
- Intuitive data manipulation
- deductive analysis
- Flexible distributed reporting

Data Mart

- A Data Mart is a subset of a directorial information store, generally oriented to a specific purpose or primary data subject which may be distributed to provide business needs.
- Data Marts are analytical record stores designed to focus on particular business functions for a specific community within an organization.
- Data marts are derived from subsets of data in a data warehouse, though in the bottom-up data warehouse design methodology, the data warehouse is created from the union of organizational data marts.

Reasons for creating a data mart

- Creates collective data by a group of users
- Easy access to frequently needed data
- Ease of creation
- Improves end-user response time
- Lower cost than implementing a complete data warehouses
- Potential clients are more clearly defined than in a comprehensive data warehouse
- It contains only essential business data and is less cluttered.

Types of Data Marts

There are mainly three approaches to designing data marts.

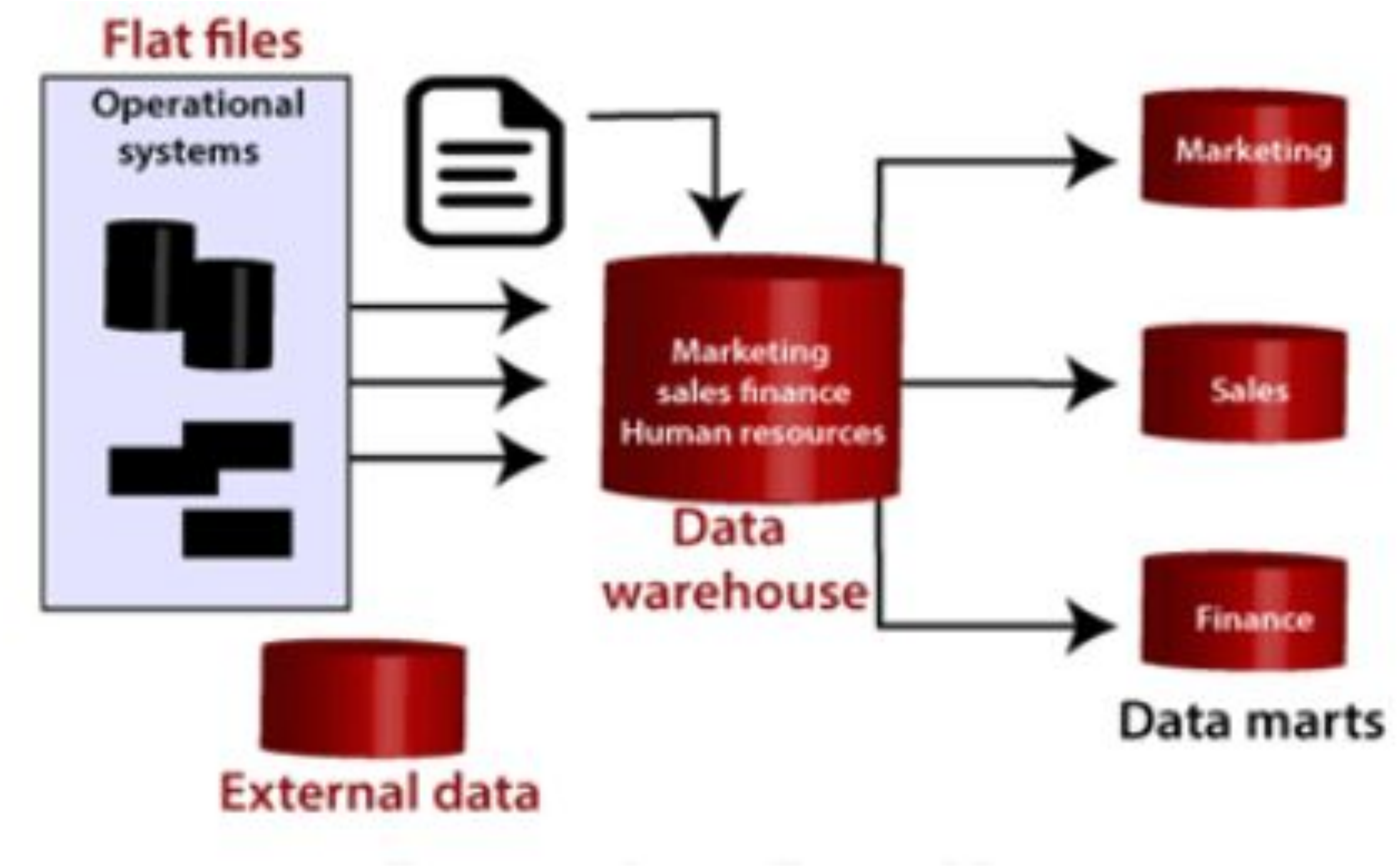
These approaches are

- Dependent Data Marts
- Independent Data Marts
- Hybrid Data marts

Dependent Data Marts

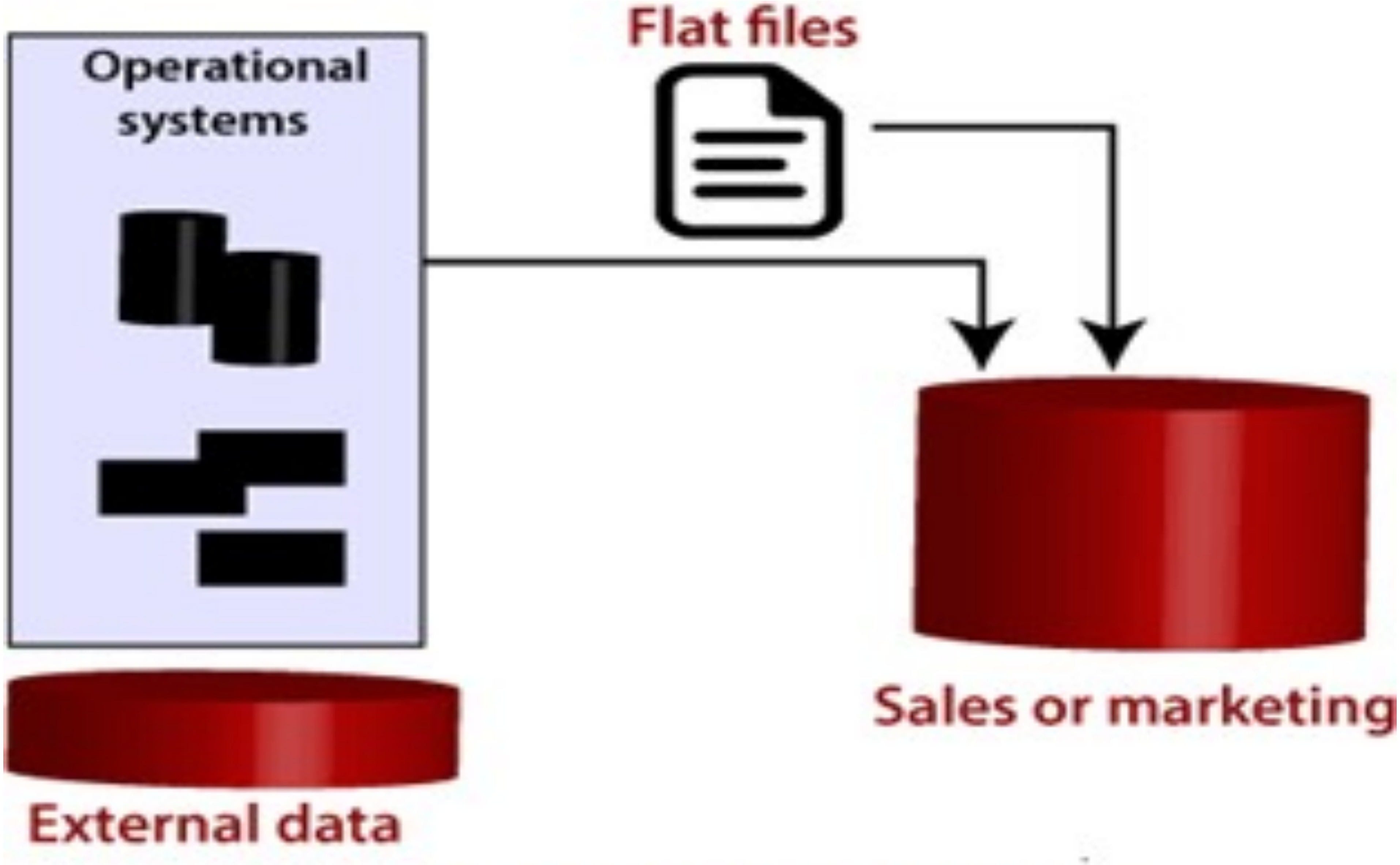
- A dependent data marts is a logical subset of a physical subset of a higher data warehouse.
- According to this technique, the data marts are treated as the subsets of a data warehouse.
- In this technique, firstly a data warehouse is created from which further various data marts can be created.
- These data mart are dependent on the data warehouse and extract the essential record from it.
- In this technique, as the data warehouse creates the data mart; therefore, there is no need for data mart integration.
- It is also known as a top-down approach.

Dependent Data Marts



Independent Data Marts

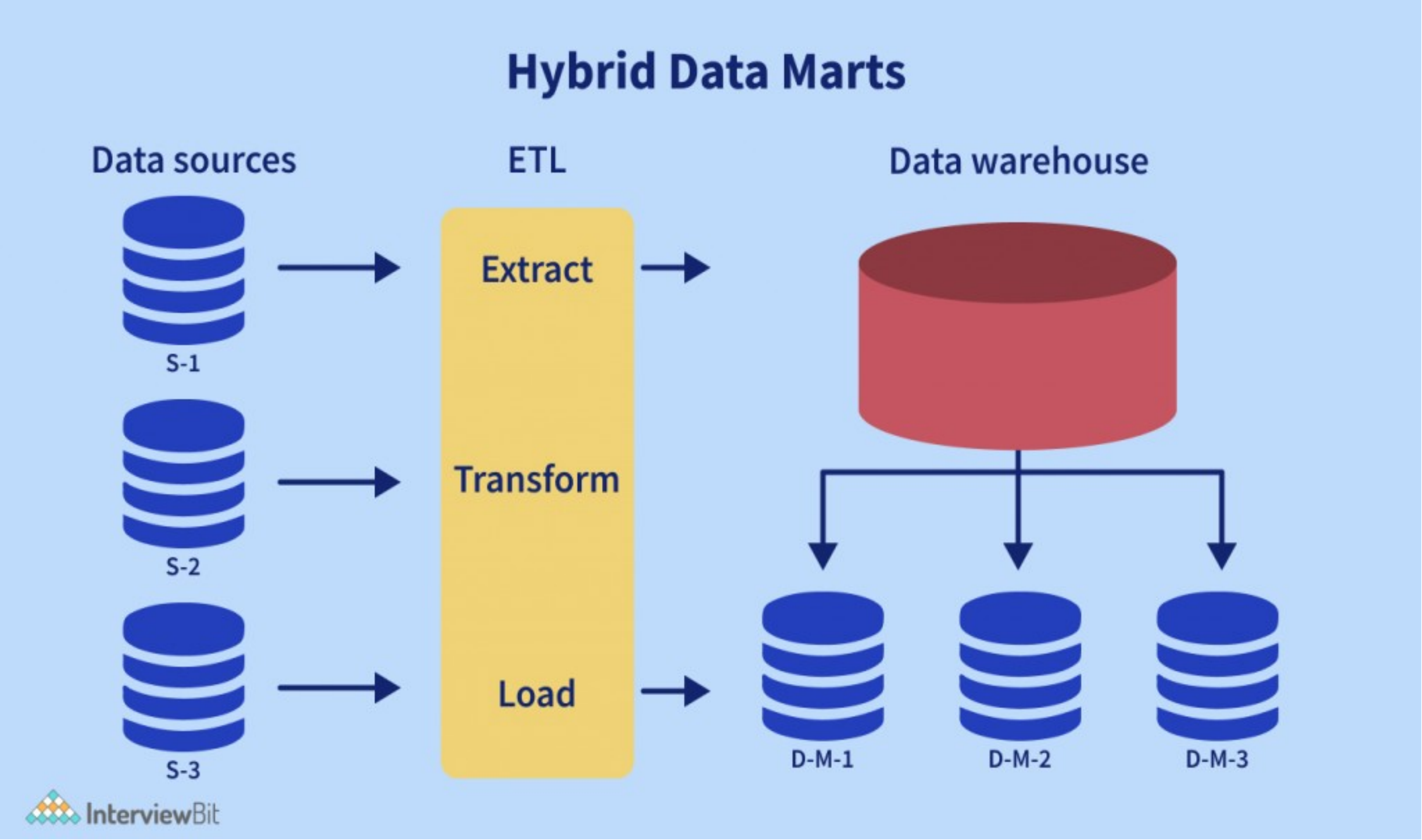
- The second approach is Independent data marts (IDM) Here,
- Firstly independent data marts are created, and then a data warehouse is designed using these independent multiple data marts.
- In this approach, as all the data marts are designed independently;
- therefore, the **integration of data marts is required.**
- It is also termed as a bottom-up approach as the data marts are integrated to develop a data warehouse.



Independent Data Mart

Hybrid Data Marts

- It allows us to **combine input from sources** other than a data warehouse.
- This could be helpful for many situations; especially when Adhoc integrations are needed, such as after a new group or product is added to the organizations.
- It forms the Datamart's directly from the data source and also from the Datawarehouse.


Hybrid Data Marts
Data sources
ETL
Data warehouse
Extract
Transform
Load
S-1
S-2
S-3
D-M-1
D-M-2
D-M-3
InterviewBit

Steps in Implementing a Data Mart

- Designing
- Constructing
- Populating
- Accessing
- Managing

Designing

- The design step is the first in the data mart process.
- This phase covers all of the functions from initiating the request for a data mart through gathering data about the requirements and developing the logical and physical design of the data mart.

Designing

It involves the following tasks:

- Gathering the business and technical requirements
- Identifying data sources
- Selecting the appropriate subset of data
- Designing the logical and physical architecture of the data mart.

Constructing

This step contains creating the physical database and logical structures associated with the data mart to provide **fast** and **efficient access** to the data.

Constructing

It involves the following tasks:

- Creating the physical database and logical structures such as tablespaces associated with the data mart.
- creating the schema objects such as tables and indexes describe in the design step.
- Determining how best to set up the tables and access structures.

Populating

This step includes all of the tasks related to the getting data from the source, cleaning it up, modifying it to the right format and level of detail, and moving it into the data mart.

It involves the following tasks:

- Mapping data sources to target data sources
- Extracting data
- Cleansing and transforming the information.
- Loading data into the data mart
- Creating and storing metadata

Accessing

This step involves putting the data to use:

querying the data, analyzing it, creating reports, charts and graphs and publishing them.

Accessing

It involves the following tasks:

- Set up and intermediate layer (Meta Layer) for the front-end tool to use.
- This layer translates database operations and objects names into business conditions so that the end-clients can interact with the data mart using words which relates to the business functions.
- Set up and manage database architectures like summarized tables which help queries agree through the front-end tools execute rapidly and efficiently.

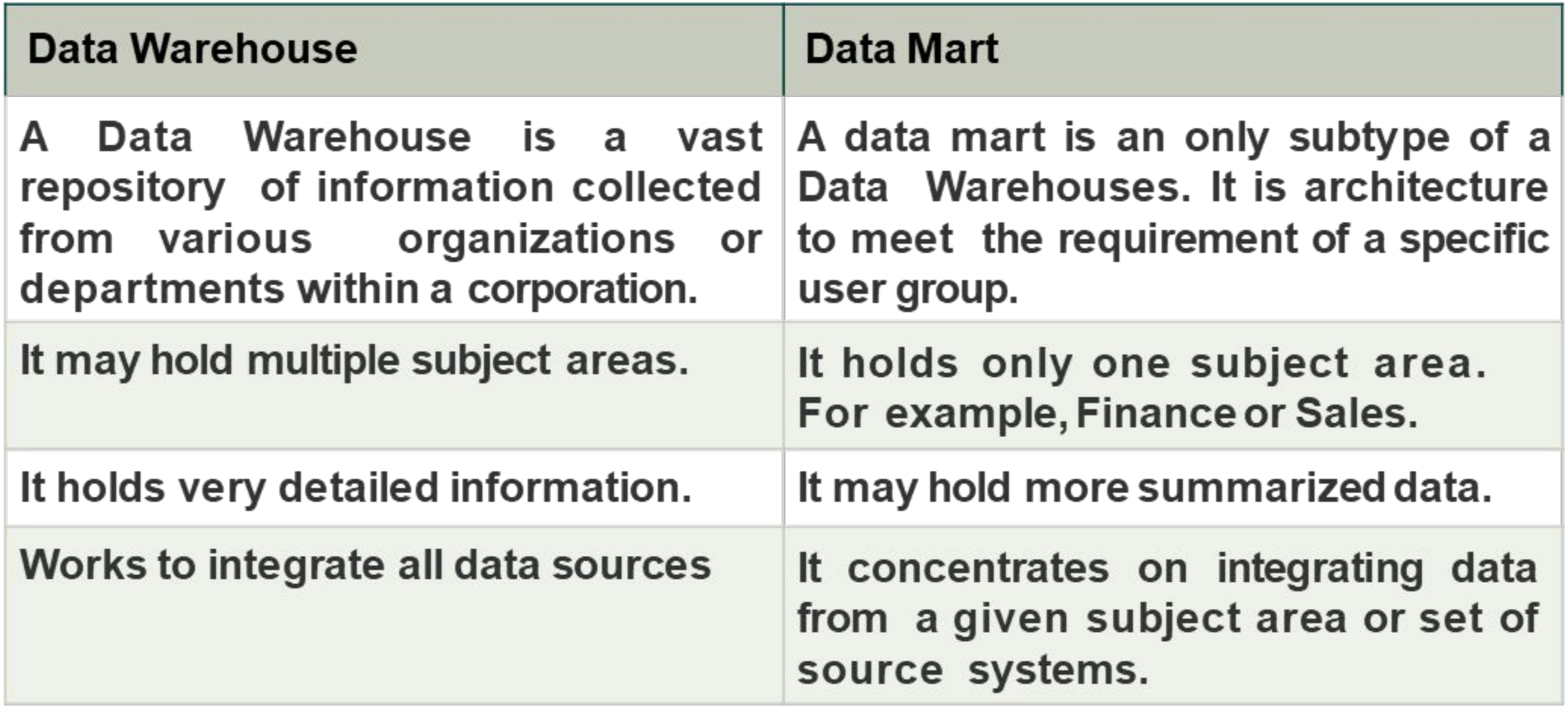
Managing

This step contains **managing the data mart** over its lifetime.

In this step, management functions are performed as:

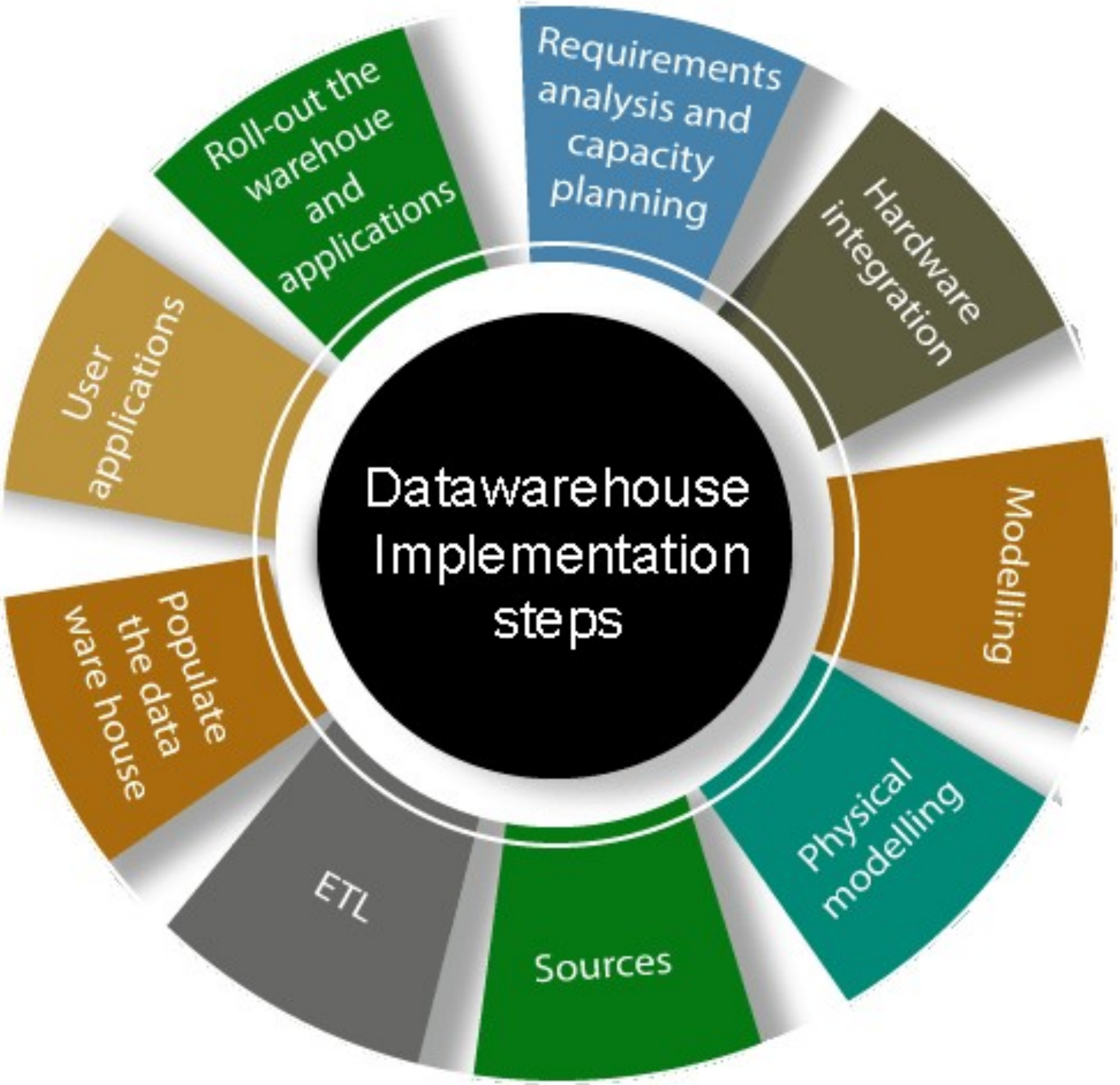
- Providing secure access to the data.
- Managing the growth of the data.
- Optimizing the system for better performance.
- Ensuring the availability of data event with system failures.

Datawarehouse vs Datamart



Data Warehouse	Data Mart
A Data Warehouse is a vast repository of information collected from various organizations or departments within a corporation.	A data mart is an only subtype of a Data Warehouses. It is architecture to meet the requirement of a specific user group.
It may hold multiple subject areas.	It holds only one subject area. For example, Finance or Sales.
It holds very detailed information.	It may hold more summarized data.
Works to integrate all data sources	It concentrates on integrating data from a given subject area or set of source systems.

Data Warehouse Implementation


Datawarehouse Implementation steps
Requirements analysis and capacity planning
Hardware integration
Modelling
Physical modelling
Sources
ETL
Populate the data ware house
User applications
Roll-out the warehoue and applications

Gathering the business requirement or Requirements analysis:

The first process in data warehousing involves defining enterprise needs, defining architectures, carrying out capacity planning, and selecting the hardware and software tools. This step will contain be consulting senior management as well as the different stakeholder.

Hardware integration:

Once the hardware and software has been selected, they require to be put by integrating the servers, the storage methods, and the user software tools.

Modeling:

Modelling is a significant stage that involves designing the warehouse schema and views. This may contain using a modeling tool if the data warehouses are sophisticated.

Physical modeling or selecting OS and selecting the database software:

For the data warehouses to perform efficiently, physical modeling is needed. This contains designing the physical data warehouse organization, data placement, data partitioning, deciding on access techniques, and indexing. The operating system and software also to be decided properly based on the requirement

Sources:

The information for the data warehouse is likely to come from several data sources. This step contains identifying and connecting the sources using the gateway, ODBC drives, or another wrapper.

ETL:

The data from the source system will require to go through an ETL phase. The process of designing and implementing the ETL phase may contain defining a suitable ETL tool vendors and purchasing and implementing the tools. This may contains customize the tool to suit the need of the enterprises.

Populate the data warehouses:

Once the ETL tools have been agreed upon, testing the tools will be needed, perhaps using a staging area. Once everything is working adequately, the ETL tools may be used in populating the warehouses given the schema and view definition.

User applications and select the end user tools:

For the data warehouses to be helpful, there must be end-user applications and tools. This step contains designing and implementing applications required by the end-users.

Roll-out the warehouses and applications:

Once the data warehouse has been populated and the end-client applications tested, the warehouse system and the operations may be rolled out for the user's community to use.

Data Warehouse Readiness Assessment

examines the state of an organization relative to the delivery of data warehouse solutions