## **Assignment 1 - Data Warehousing (17.07.25)**

## **Data Warehousing:**

* A data warehouse is a big storage system where companies keep historical business data from different sources.
* The goal is to help in decision-making by collecting, organizing, and storing data in one place.
* It supports managers and analysts to find useful patterns, trends, and summaries.

## **Purpose of Data Warehouse:**

To help businesses make smarter decisions by analyzing past data.

It gives a complete view of the business at any point in time.

It supports:

* Business reporting (like sales reports)
* Trend analysis
* Forecasting and planning

## **Features:**

**1. Subject-Oriented** :  
 A data warehouse is built around specific subjects (topics) like sales, customers, products, or finance — not around day-to-day operations.  
 Instead of focusing on one single sales transaction, it stores and organizes all sales data across time to help understand trends.

**2. Integrated** :  
 Data comes from different sources (like Excel sheets, apps, databases), but in the data warehouse, it’s all combined and cleaned up so it fits together well.  
 One system says "Male", another says "M". The warehouse turns them all into one standard format: “Male”.

**3. Time-Variant:**  
 Data in a warehouse is stored with timestamps so you can see how things change over time.  
 You can ask: “What were our sales in July 2022?” or “How has customer behavior changed in 3 years?”

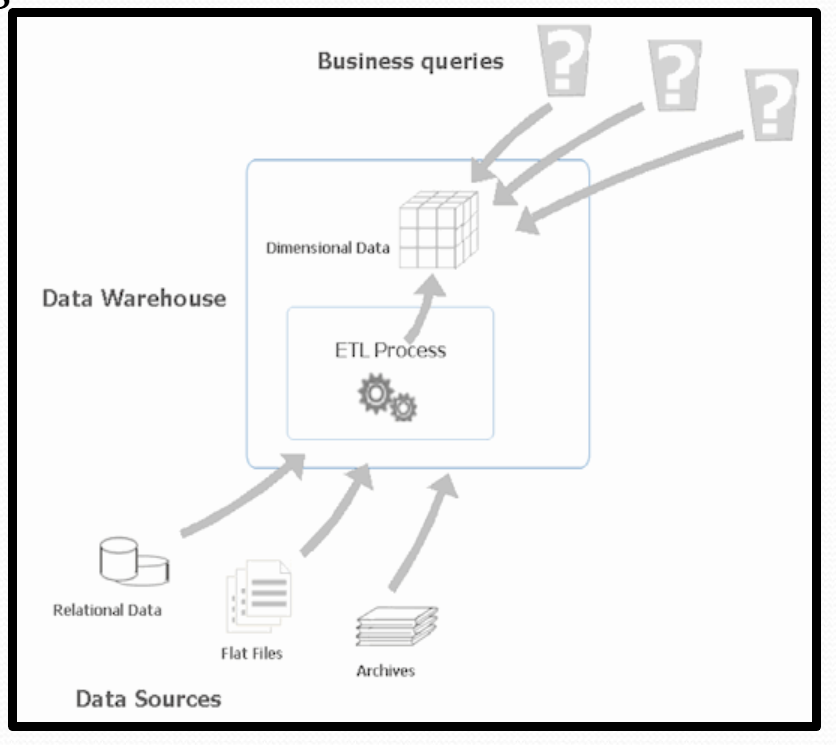
**4. Non-Volatile:**  
 Once data is in the warehouse, it's not changed or deleted. You only add new data. This keeps historical records accurate.  
 You don’t go back and edit January 2021 sales data. You keep it as it was and just add new months' data.

## **Data Warehouse Architecture (ETL-Based):**

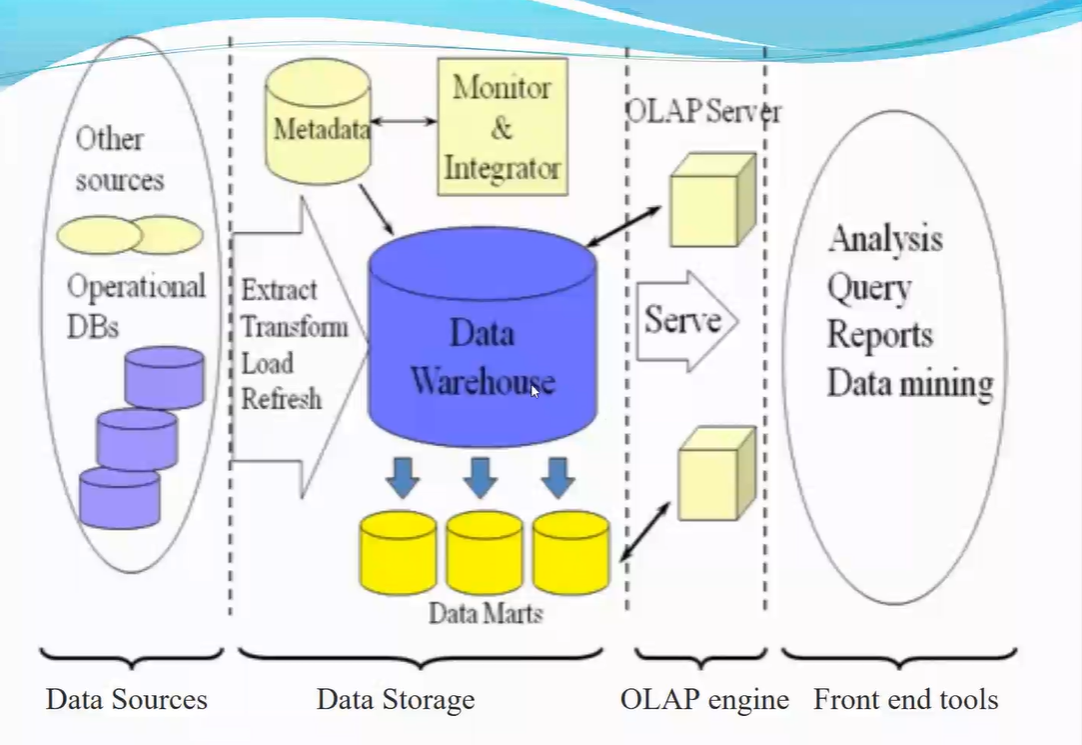
### **ETL stands for:**

* Extract – Get data from multiple sources (e.g., files, databases)
* Transform – Clean, filter, and format the data
* Load – Store the transformed data into the warehouse

### **Architecture Includes:**

1. Staging Layer: Temporary storage for raw data before it's cleaned.
2. Integration Layer: Data is processed and made ready for use.
3. Access Layer (Data Marts): End-users can query and analyze data here.

## **Architecture**



## **Operational Data Store (ODS)**

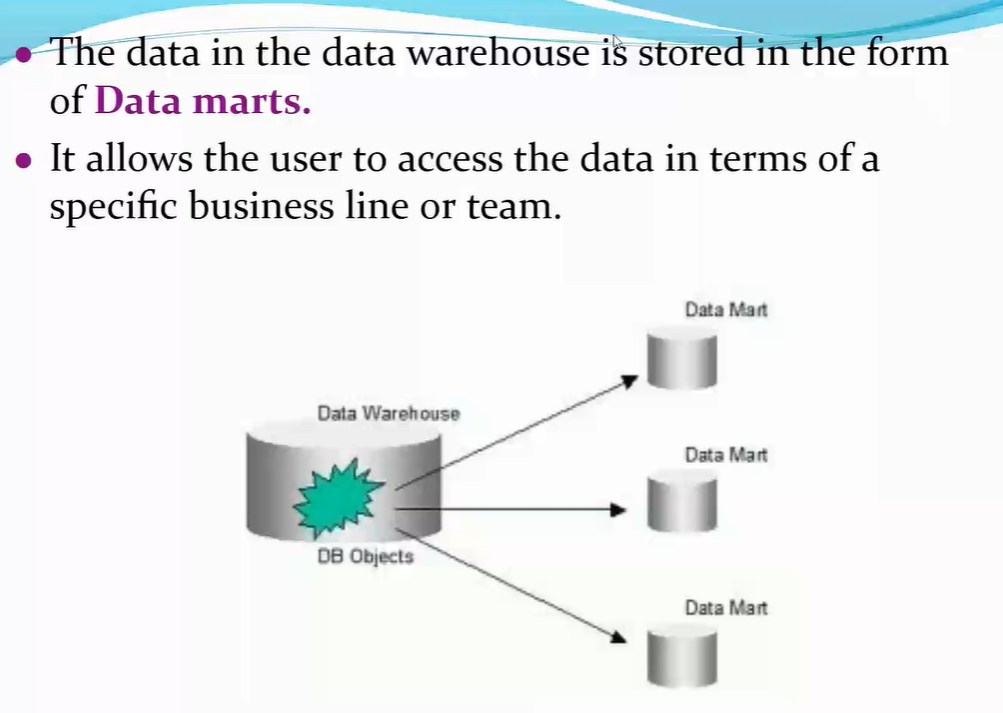
* ODS is a middle step between live systems (like apps or websites) and the data warehouse.
* It stores real-time or current operational data before it's sent to the warehouse.
* Useful for quick access and updates.
* Data in ODS is:  
  + Integrated from many sources
  + Used for day-to-day operations
  + Later passed to the data warehouse for long-term analysis

## **OLTP vs Data Warehouse Applications**

| **Feature** | **OLTP (Online Transaction Processing)** | **Data Warehouse (OLAP)** |
| --- | --- | --- |
| Purpose | Run day-to-day operations | Analyze historical data |
| Data Type | Real-time, current data | Historical, read-only data |
| Operations | Insert, update, delete | Read, analyze, summarize |
| Users | Staff, clerks | Managers, analysts |
| Queries | Short and simple | Complex and analytical |

## **Data Marts :**

* A Data Mart is a smaller, focused version of a data warehouse.
* It holds data for a specific department or team (like Sales, HR, Finance).
* Faster and cheaper than a full data warehouse.
* Easier for small groups to access relevant data.



## **Data Mart vs Data Warehouse :**

| **Feature** | **Data Mart** | **Data Warehouse** |
| --- | --- | --- |
| Size | Small | Large |
| Scope | Department level | Organization wide |
| Users | Specific team (e.g Sales) | Whole company |
| Complexity | Simple | Complex |
| Build Time | Short | Long |

## **Data Warehouse Life Cycle :**

### **1. Requirement Analysis**

* Understand what the business needs.
* What kind of reports, data, dashboards are expected?
* Talk to stakeholders like managers, analysts, etc.

**2. Data Modeling**

* Design how the data will be stored.
* Use diagrams (like ER diagrams or star/snowflake schemas).
* Decide tables, relationships, dimensions, facts.

**3. ETL Process (Extract, Transform, Load)**

* **Extract** data from different sources.
* **Transform** it into the correct format (cleaning, formatting).
* **Load** it into the data warehouse.

**4. Data Warehouse Deployment**

* Make the warehouse live.
* Set up data marts if needed (department-specific areas).

**5. Testing & Validation**

* Check if data is accurate, up to date, and useful.
* Validate performance, speed, and reliability.

**6. Maintenance & Updates**

* Regularly update data.
* Fix issues, add new sources or reports.
* Ensure security and performance.