**Data Engineering Assignment 1**

# **An Introduction to Data Warehousing**

A Data Warehouse (DW) is a system that is specially designed to support management in decision-making. It stores a large amount of historical data in a way that provides a clear picture of business activities over time.

According to W.H. Inmon, a data warehouse is subject-oriented, integrated, time-variant, and non-volatile.

### **Features of Data Warehousing**

### **1. Subject-Oriented**

Data is organized based on key subjects like sales, finance, and marketing.  
This helps managers focus on decision-making instead of daily operations.

### **2. Integrated**

Data is collected from multiple sources and made consistent.  
It ensures uniform naming, formats, and measurements across all data.

### **3. Time-Variant**

A data warehouse stores historical data over long periods.  
This allows analysis of trends and business performance over time.

### **4. Non-Volatile**

Once data is stored, it is not updated or deleted.  
Only new data is added, maintaining a stable historical record.

# **Purpose of a Data Warehouse**

To help businesses make better decisions by looking at past and current data in one place.

It gives a clear and complete picture of the business.

It is useful for:

● Creating reports (like sales or profit reports)

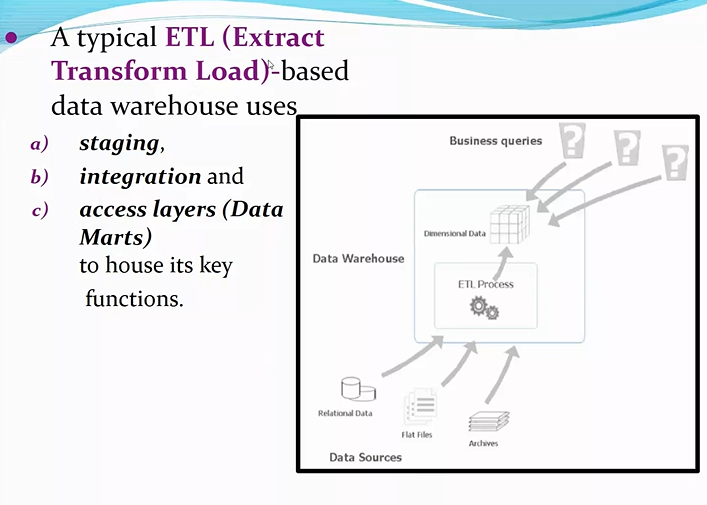
● Finding patterns and trends

● Planning for the future

# **Data Warehourse Architecture (ETL Based)**

ETL stands for Extract, Transform, Load.  
It is a process used to collect data from different sources, clean and change it into the right format, and then store it in a data warehouse.

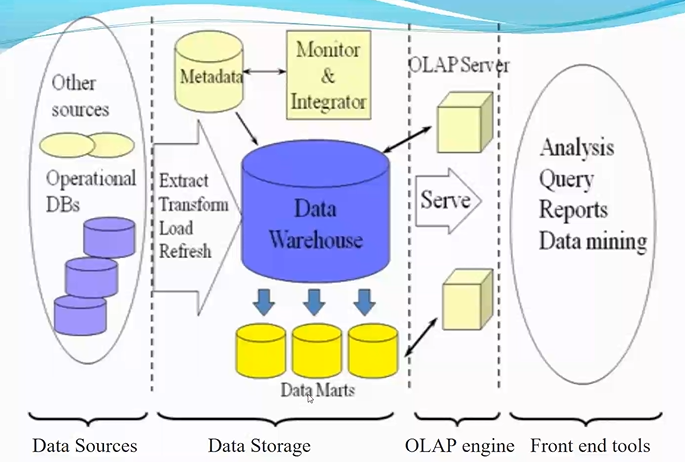
* Extract – Get the data from different places (like databases, files, or applications).
* Transform – Clean the data and change it into a format that can be used.
* Load – Store the final data into the data warehouse for analysis.



### **Architecture Includes**

1. **Staging Layer:** Collects and temporarily stores raw data from various sources.
2. **Integration Layer:** Cleans, transforms, and combines data to make it consistent and usable.
3. **Access Layer (Data Marts):** Provides organized data for users to run reports and perform analysis.

### **Architecture**



# **Operational Data Store**

An Operational Data Store (ODS) is like a bridge or middle layer between:

* Real-time, live systems (like apps, websites, transactional databases)
* And the Data Warehouse (used for long-term storage and analysis)

It collects and integrates current data from multiple sources.

**Purpose of ODS**

It holds this data temporarily for:

* Quick reporting
* Real-time decisions
* Operational tasks

After that, the data is cleaned and transferred to the data warehouse for long-term storage and analysis.

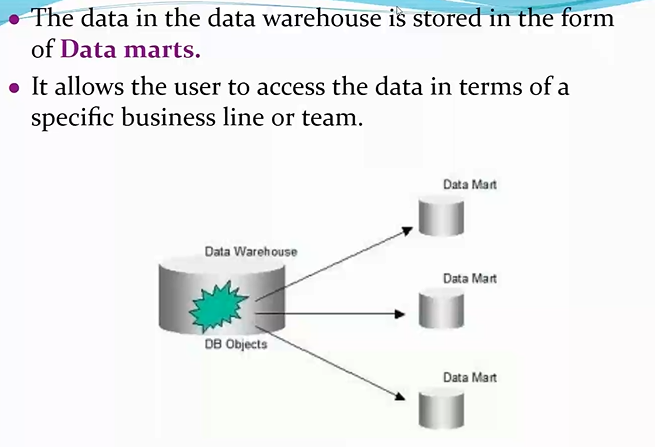
# **OLTP vs Data Warehouse Applications**

| **Feature** | **OLTP** | **Data Warehouse (OLAP)** |
| --- | --- | --- |
| Purpose | Day-to-day operations | Data analysis and decision-making |
| Data Type | Current, real-time | Historical |
| Operations | Insert, Update, Delete | Complex Queries, Read-only mostly |
| Structure | Normalized (many tables) | Denormalized (fewer, larger tables) |
| Users | Frontline staff, customers | Analysts, Managers |
| Example | ATM, Booking apps | Sales dashboards, BI tools |
| Speed | Very fast for small queries | Fast for big reports, but not for updates |
| Data Volume | Smaller, live data | Very large, multi-year data |
| Pitfall | Not good for analysis | Not good for frequent updating |

# **Data Marts**

A **Data Mart** is a smaller part of a **Data Warehouse**. It is created for a specific department or team in a company, like Sales, Marketing, or Finance.

Instead of giving access to the whole data warehouse, a data mart only gives the data needed by that particular team. This makes it easier and faster for them to find the information they need.



### **Features of a Data Mart:**

* It is a **subset** of the data warehouse.
* It is **focused on one business area** (like sales or marketing).
* It helps teams make **quick and smart decisions**.
* It is **faster and cheaper** to use than accessing the full data warehouse.

### **Example:**

A retail company has a big data warehouse with all its business data.  
The Sales team only needs sales-related data, so they get a **Sales Data Mart**.  
The Marketing team gets a **Marketing Data Mart**.  
Each team uses its own data mart to do its work easily.

**Data Marts vs Data Warehouse**

| **Feature** | **Data Warehouse** | **Data Mart** |
| --- | --- | --- |
| **Definition** | A large storage system that holds all business data | A smaller part of a data warehouse for a specific team or subject |
| **Size** | Very large | Smaller |
| **Scope** | Covers the entire organization | Focuses on one department or business area |
| **Users** | Used by data analysts, management, across departments | Used by individual teams (like sales, marketing) |
| **Data Source** | Takes data from multiple operational systems | Usually gets data from the data warehouse |
| **Purpose** | For enterprise-wide analysis and reporting | For team-level or department-level analysis |
| **Cost** | More expensive to build and maintain | Cheaper and easier to manage |
| **Performance** | Can be slower (due to size) | Faster (handles less data) |
| **Maintenance** | Needs more time and resources | Easier to maintain |

# **Data Warehouse lifecycle**

### **1. Requirement Specification**

* Business needs gathered by analysts
* Requirements collected from client side
* Acts as blueprint for data warehouse

### **2. Data Modelling**

* Designs structure for storing data
* Organizes and connects related data
* Uses Star, Snowflake, Galaxy schema

### **3. ETL Design and Development**

* Extract data from source systems
* Transform data into right format
* Load data into data warehouse

### **4. OLAP Cubes**

* Stores data in multiple dimensions
* Supports fast and deep analysis
* Becoming outdated for real-time use

### **5. UI Development**

* Builds interface for user access
* Helps users generate reports easily
* Tools like Tableau, Power BI

### **6. Maintenance**

* Update schema or warehouse structure
* Add or remove data attributes
* Ensure system runs without issues

### **7. Test and Deployment**

* Check if data is complete
* Validate data from different sources
* Deploy warehouse for user access