**HEXAWARE TRAINING**

**ASSIGNMENT-1**

**DATA WAREHOUSE:**

****A **Data Warehouse** is a centralized repository that stores data collected from multiple sources. It is specifically designed to facilitate querying and analysis, rather than transaction processing. This data is typically structured, historical, and subject-oriented, which helps organizations make informed business decisions.

Data warehouses are optimized for **read-heavy operations**, where the focus is on retrieving large amounts of data efficiently. Unlike operational databases, which handle real-time updates and deletions, data warehouses support complex queries, trend analyses, and business intelligence tasks.

Key characteristics of a data warehouse include:

* **Subject-Oriented:** Focused on key business subjects like sales, customers, or inventory.
* **Integrated:** Consolidates data from different sources into a unified format.
* **Non-Volatile:** Data is stable; it’s rarely deleted or updated.
* **Time-Variant:** Stores historical data for analysis over different time periods.

**DATA WAREHOUSE ARCHITECTURE:**

The architecture of a data warehouse typically includes the following components:

**1. Data Sources**

Data originates from various **heterogeneous sources**, which may include:

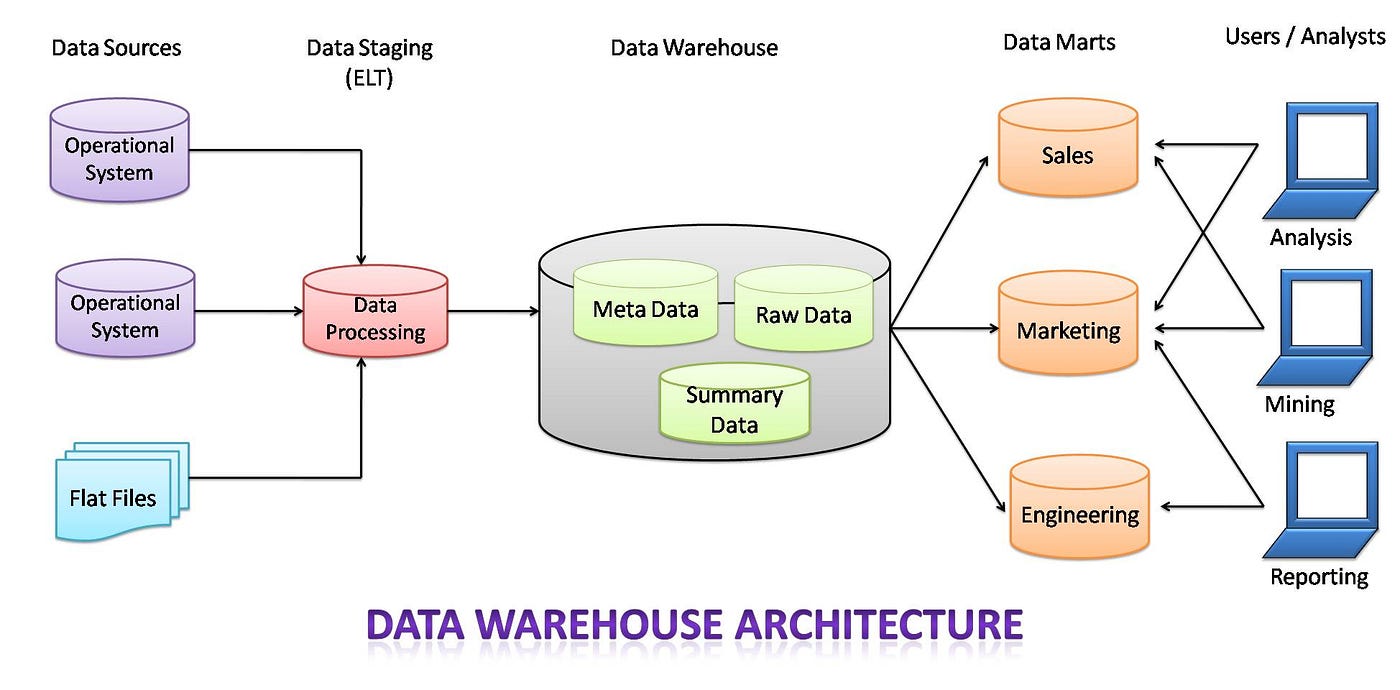
* **Operational Systems:** Such as ERP, CRM, or legacy systems that manage day-to-day transactions.
* **Flat Files:** Including CSV, XML, or spreadsheet files containing semi-structured or unstructured data.

These sources provide the raw data needed for decision-making and analysis.

**2. Data Staging Area (ETL / ELT Process)**

The data staging area serves as an intermediate zone where data undergoes **ETL (Extract, Transform, Load)** or **ELT (Extract, Load, Transform)** processes:

* **Extract:** Data is gathered from various source systems.
* **Transform:** Data is cleaned, validated, and reformatted to meet warehouse standards.
* **Load:** The transformed data is loaded into the data warehouse for storage.

This step ensures the consistency, accuracy, and completeness of data before it enters the warehouse.

**3. Data Warehouse Storage**

Once processed, the data is stored in a **centralized data warehouse**, which consists of three main components:

* **Meta Data:** Describes data definitions, source details, and usage.
* **Raw Data:** Stores detailed, unprocessed data for backup or auditing.
* **Summary Data:** Pre-aggregated, high-level data for faster querying and reporting.

This layered storage enables both detailed and high-level analysis.

**4. Data Marts**

**Data Marts** are specialized repositories derived from the data warehouse. Each data mart is focused on a specific business domain or department, such as:

* **Sales**
* **Marketing**
* **Engineering**

They are optimized for departmental access and improve query performance by narrowing the data scope.

**5. Users / Analysts**

Different user groups utilize the data for various analytical purposes:

* **Analysis:** Business analysts explore patterns and metrics.
* **Mining:** Data scientists use algorithms to extract insights.
* **Reporting:** Executives and managers generate reports for strategic decisions.

These users interact with the system through dashboards, query tools, or business intelligence platforms.

## ****NEED FOR DATA WAREHOUSE:****

**1. Handling Large Volumes of Data**: Traditional databases can only store a limited amount of data (MBs to GBs), whereas a data warehouse is designed to handle much larger datasets (TBs), allowing businesses to store and manage massive amounts of historical data.

**2. Enhanced Analytics**: Transactional databases are not optimized for analytical purposes. A data warehouse is built specifically for data analysis, enabling businesses to perform complex queries and gain insights from historical data.

**3. Centralized Data Storage**: A data warehouse acts as a central repository for all organizational data, helping businesses to integrate data from multiple sources and have a unified view of their operations for better decision-making.

**4. Trend Analysis**: By storing historical data, a data warehouse allows businesses to analyze trends over time, enabling them to make strategic decisions based on past performance and predict future outcomes.

**5. Support for Business Intelligence**: Data warehouses support business intelligence tools and reporting systems, providing decision-makers with easy access to critical information, which enhances operational efficiency and supports data-driven strategies.

## TYPES OF DATA WAREHOUSES:

The different types of Data Warehouses are:

1. **Enterprise Data Warehouse (EDW)**: A centralized warehouse that stores data from across the organization for analysis and reporting.
2. **Operational Data Store (ODS)**: Stores real-time operational data used for day-to-day operations, not for deep analytics.
3. **Data Mart**: A [subset](https://www.geeksforgeeks.org/data-marts-storage-component-of-hdfs/) of a data warehouse, focusing on a specific business area or department.
4. **Cloud Data Warehouse**: A data warehouse hosted in the cloud, offering scalability and flexibility.
5. **Big Data Warehouse**: Designed to store vast amounts of unstructured and structured data for big data analysis.
6. **Virtual Data Warehouse**: Provides access to data from multiple sources without physically storing it.
7. **Hybrid Data Warehouse**: Combines on-premises and cloud-based storage to offer flexibility.
8. **Real-time Data Warehouse**: Designed to handle real-time data streaming and analysis for immediate insights.

**REAL-TIME APPLICATIONS:**

**1. Banking**

**Use Case:** Fraud detection, risk management, and customer insights.

* + Consolidates transaction data from ATMs, mobile apps, and branches.
  + Detects unusual patterns in transactions to prevent fraud.
  + Assesses credit risk before approving loans.
  + Generates reports for regulatory compliance.

**2. E-Commerce**

**Use Case:** Personalized recommendations, order tracking, and logistics.

* + Stores clickstream data to analyze user browsing behavior.
  + Integrates inventory, shipping, and user data to optimize delivery.
  + Powers product recommendations based on past purchases and search history.
  + Enables dynamic pricing based on demand.

**3. Healthcare**

**Use Case:** Patient records, treatment history, and diagnostics.

* + Centralizes patient data from labs, diagnostics, and consultations.
  + Tracks treatment effectiveness across time and regions.
  + Supports predictive analytics for early disease detection.
  + Assists in operational planning like staff allocation and equipment use.

**4. Telecommunications**

**Use Case:** Customer churn prediction, network optimization, and billing.

* + Tracks call records, data usage, and complaint logs.
  + Identifies customers likely to switch providers (churn).
  + Optimizes network usage by analyzing peak traffic zones.
  + Consolidates billing data for accurate invoicing.