**Data Warehouse**

A Data Warehouse (DW) is a centralized repository that stores integrated data from multiple sources. It is used to analyze and report structured and semi-structured data from multiple data sources, such as point-of-sale transactions, marketing automation, customer relationship management, and more.

* It supports decision-making by providing historical, current, and trend data.
* Data is usually cleaned, transformed, and loaded into the warehouse through a process known as ETL (Extract, Transform, Load).

**Benefits of a data warehouse:**

* Informed decision making
* Consolidated data from many sources
* Historical data analysis
* Data quality, consistency, and accuracy
* Separation of analytics processing from transactional databases, which improves the performance of both systems

**Key Characteristics of a Data Warehouse**

1. **Subject-oriented**

They can analyze data about a particular subject or functional area (such as sales).

1. **Integrated**

Data warehouses create consistency among different data types from disparate sources.

1. **Non-volatile**

Once data is in a data warehouse, it’s stable and doesn’t change.

1. **Time-variant**

 Data warehouse analysis looks at change over time.

**Components of Data Warehouse:**

**1. Data Sources**

* External systems that provide raw data.
* Examples: OLTP databases, flat files, APIs, CRM, ERP systems.

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

* **Extract:** Gathers data from different sources.
* **Transform:** Cleanses, formats, and transforms the data.
* **Load:** Loads the transformed data into the data warehouse.

Tools: Apache NiFi, Talend, Informatica, SSIS

**3. Staging Area**

* Temporary storage area used during ETL.
* Allows data to be cleaned and validated before loading into the warehouse.

**4. Data Warehouse Storage (Data Repository)**

* Centralized location where integrated data is stored.
* Structured in schemas like **Star**, **Snowflake**, or **Galaxy**.
* Optimized for read-heavy queries.

**5. Metadata**

* **Data about data** — describes source, structure, rules, transformations, etc.
* Helps users and tools understand how data is organized.

**6. OLAP Engine (Online Analytical Processing)**

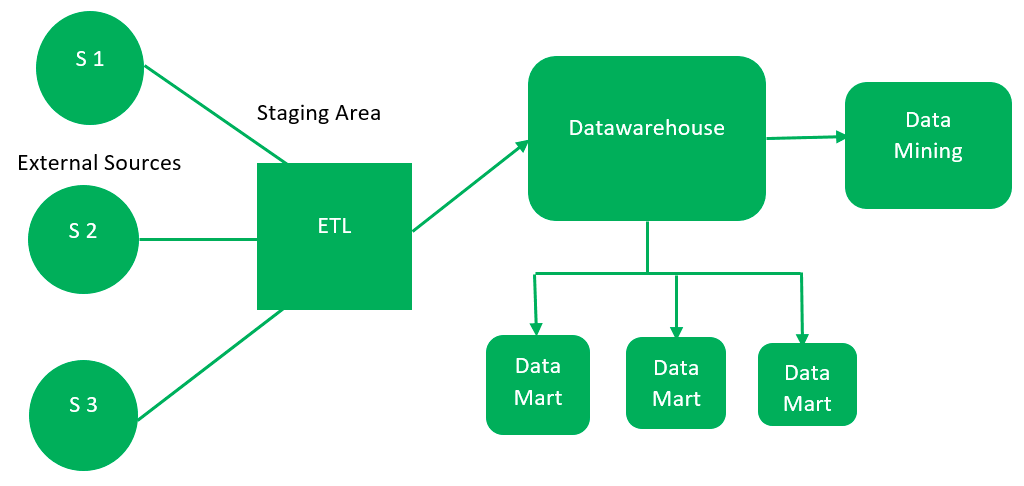
* Supports **multidimensional queries** and complex analytical operations.
* Types: MOLAP, ROLAP, HOLAP.

**7. Data Marts**

* Subsets of the data warehouse tailored to specific business units (e.g., sales, finance).
* Improves performance and access speed for specific users.

**8. Front-End Tools / Presentation Layer**

* Tools used by business users to interact with data.
* Examples: Dashboards, reports, visualization tools (Power BI, Tableau, etc.).



**Data Warehouse Architecture**

1. **Single-Tier Architecture**
   * Simplified, rarely used in practice.
2. **Two-Tier Architecture**
   * Separate client and data warehouse layers.
   * Lacks scalability.
3. **Three-Tier Architecture** (most common)
   * **Bottom Tier:** Data warehouse database server (e.g., SQL Server, Oracle).
   * **Middle Tier:**

 Acts as the **intermediary** between the database and users.

 Handles **business logic**, multi-dimensional analysis, and data processing.

 OLAP engine for analysis.

**Types of OLAP:**

* MOLAP (Multidimensional OLAP)
* ROLAP (Relational OLAP)
* HOLAP (Hybrid OLAP)
  + **Top Tier:**
    - Front-end tools for reporting.
    - User interface layer.
    - Consists of dashboards, reporting tools, ad hoc query tools.
    - Enables users to visualize and interact with data.
    - Tools include Tableau, Power BI, Looker, Excel

**Types of Data Warehouses**

The different [types of Data Warehouses](https://www.geeksforgeeks.org/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.

**Popular Data Warehouse Tools & Technologies:**

* Amazon Redshift
* Google BigQuery
* Snowflake
* Microsoft Azure Synapse
* IBM Db2 Warehouse