2_Datawarehouse

Tuesday, February 10, 2026 2:29 PM

What is a Data Warehouse?
A Data Warehouse is a **central repository** that stores an organization's data for **reporting, analysis, and decision-making**.
It is designed to handle **large volumes of data** coming from many different systems.
Key Characteristics
1. Historical Database

- A DWH stores **many years of historical business data**.
- This long-term data helps organizations analyze trends and make strategic decisions.

2. Read-Only for Analysis

- A DWH is built for **reading data**, not for day-to-day transactions.
- It is optimized for:
- Reporting
- Analytics
- Dashboards
- Business Intelligence (BI)

Because it is not used for inserts/updates like OLTP, it is often called a **read-only database**.
3. Decision Support System (DSS)

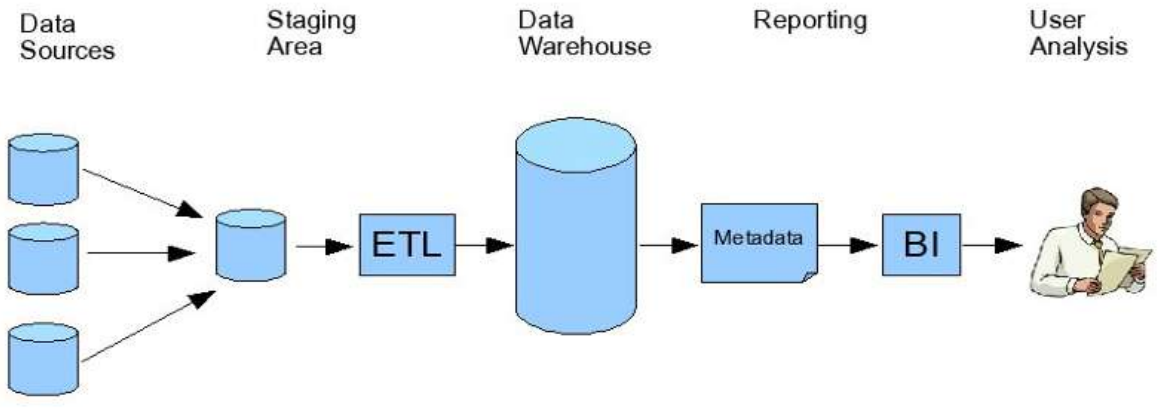
- The main purpose of a DWH is to **support business decisions**.
- It provides clean, consistent, and integrated data for:
- Executives
- Analysts
- Data scientists

This is why a DWH is also known as a **Decision Support System (DSS)**.
4. Single, Complete, and Consistent Data Store
A Data Warehouse brings data from multiple sources into **one unified place**, such as:

- OLTP systems
- CRM
- ERP
- Flat files
- Cloud storage
- APIs

After integration, the data becomes:

- **Single version of truth**
- **Consistent**
- **Cleaned and transformed**
- **Ready for business use**



Transforming data into information

ETL in Data Warehousing
ETL = Extract, Transform, Load

- **Extract** → pull data from source systems

- **Transform** → clean, standardize, and shape the data
- **Load** → store it in the Data Warehouse

ETL ensures the DW contains **clean, consistent, and analytics-ready** data.

Common ETL Tools

- **Informatica**
- **Datastage**
- **SSIS**
- **Talend**
- **Cloud ETL Tools:**
  - Azure Data Factory (ADF)
  - Matillion
  - Informatica Cloud (IICS)
- **dbt** → transformation-only tool (ELT)

The Staging Area is a **temporary workspace** used before data enters the Data Warehouse.
Its job is to **collect, clean, transform, and prepare** data coming from multiple source systems.

Key Characteristics

1. Temporary Storage for Incoming Data

- It holds raw data **only for a short time**.
- Once the data is processed and loaded into the DW, staging data is usually deleted.

2. Accepts Data from Multiple Sources

- Operational databases
- Flat files
- APIs
- Cloud systems
- Third-party applications

The staging area acts as a **landing zone** where all these different formats are brought together.

3. Structure Resembles Operational Systems

- Tables in staging look similar to the **source systems**, not the DW.
- This makes it easier to load raw data without complex transformations upfront.

4. Performs Data Preparation Tasks

The staging area handles all the "dirty work" before data reaches

- Cleaning
- Standardizing
- Transforming
- Combining data from multiple sources
- Removing duplicates
- Validating formats
- Archiving if needed

  This ensures the DW receives **clean, consistent, and ready-to-use** data.

5. Handles Timing Differences

- Data may arrive at **different times** from different systems.
- The staging area **merges and aligns** these datasets before loading them into the DW.

6. No Historical Data

- Staging is **not** a historical store.
- It keeps only the data needed for the current load cycle.
- After loading into DW, staging table

Data Acquisition

Meaning of Data Acquisition

Data Acquisition refers to the **entire ETL process**:

Extract → Transform → Load

It involves pulling data from different source systems, preparing it, and loading it into the **Staging Area**, and then into the **Data Warehouse (DWH)** or **Data Marts (DMs)**.

The Staging Area is a **temporary storage zone** used only during the acquisition process.

I. Data Extraction

Data Extraction is the process of **reading and collecting data** from various source systems.

Common source types:

- Operational databases (OLTP)
- ERP systems
- Mainframe systems (COBOL)
- NoSQL databases
- Salesforce
- Teradata
- Flat files (CSV, JSON, XML)
- Cloud applications

Goal: **Bring raw data into the Staging Area** without applying heavy transformations.

II. Data Transformation

Data Transformation converts raw source data into the **required business format** before loading into the DWH.
There are **four major types of transformations**:

1. Data Merging

Combining data from **similar sources** with similar structure and data types.
Examples:

- Joins
- Unions
- Combining multiple customer files into one dataset

Purpose: **Integrate related datasets into a unified structure.**

2. Data Cleansing

Identifying and correcting **inaccuracies, inconsistencies, and formatting issues**.
Examples:

- Converting text to proper case (Initcap)
- Lowercase / Uppercase conversions
- Replacing nulls (NVL)
- Fixing spelling or formatting issues
  Example: aMit → Amit

Purpose: Ensure **clean, standardized, and reliable** data.

3. Data Scrubbing

Deriving **new columns or definitions** from existing data.
Example:

- Adding a new column TAX in the target table
- TAX is calculated based on the SAL column from the source

Purpose: Create **business-ready attributes** that do not exist in the source.

4. Data Aggregation

Summarizing multiple detailed values into a **single summary value**.
Examples:

- Sum
- Average
- Min
- Max
- Count

Purpose: Produce **summary-level metrics** needed for reporting and analytics.

III. Data Loading

After extraction and transformation, data is loaded into:

- **Data Warehouse (DWH)**
- **Data Marts (DMs)**

Loading can be:

- **Full Load** (entire data)
- **Incremental Load** (only new or changed data)