DATA WAREHOUSE

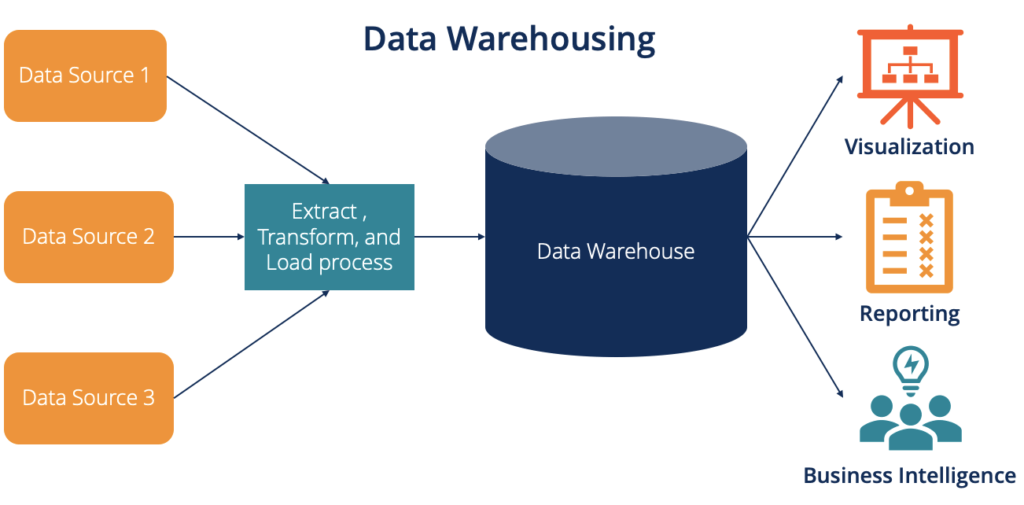
Ahmed Sherif

**Definition:**

Data Warehouse (DW) is a centralized repository that stores large volumes of structured data, semi-structured data and unstructured data from multiple sources. It enables Business Intelligence (BI), analytics, and reporting.

It is a system designed for query and analysis rather than transaction processing. It consolidates historical and current data from disparate sources to support decision-making.

Data can feed into a warehouse from multiple databases, including customer relationship, inventory, point of sale (POS) and supply chain management systems.



Business intelligence initiatives have been supported by data warehousing systems, also known as enterprise data warehouse (EDW) systems. Their primary responsibilities include gathering information from various sources, cleaning and preparing it, and loading and preserving it.

**History of Data Warehousing:**

The increasing demand for structured data analysis in businesses has fueled the development of data warehousing over several decades. Here is a timeline of significant events:   
  
1. Prior to the 1970s:

* Flat files and hierarchical databases (like IBM's IMS) were the mainstays of early data processing businesses.
* Mainframe systems with little analytical power produced the reports.

2. The emergence of relational databases in the 1970s

* The relational database model was first presented by IBM's Edgar F. Codd in 1970.
* Structured querying (SQL) was made possible by databases such as IBM DB2 and Oracle (1979).
* Databases were first used for analytics in Decision Support Systems (DSS).

3. Data warehousing's inception in the 1980s

* In the early 1980s, data warehousing pioneer Bill Inmon formalized the idea.
* A "subject-oriented, integrated, non-volatile, time-variant" data store is what was defined as a DWH.
* The first commercial DWH appliance for large-scale analytics was introduced by Teradata in 1984.

4. Rapid Adoption and Technological Developments in the 1990s

* Dimensional modeling (star schema, snowflake schema) was first presented by Ralph Kimball.
* Tools for online analytical processing, or OLAP (such as Microsoft Analysis Services), became available.
* For data integration, ETL (Extract, Transform, Load) has become a standard procedure.
* Specialized DWH databases for analytics were first introduced by Red Brick Systems in the 1990s.

5. The 2000s: Growth and the Impact of Big Data

* DWHs for unstructured data were supplemented by data lakes (Hadoop, 2006).
* Cloud-based DWHs became available (Google BigQuery, 2010; Amazon Redshift, 2012).

6. Current Cloud & AI-Powered Warehousing, 2010s–Present

* A scalable, fully cloud-native DWH was first presented by Snowflake (2012).
* Real-time analytics (streaming ETL, Apache Kafka) became essential.
* Integration of AI and ML (e.g., Google BigQuery ML's automated insights).
* Traditional DWH models were challenged by data mesh and decentralized architectures.
* Prominent Data Warehousing Figure Bill Inmon promoted the top-down strategy, which prioritizes enterprise DWH.

**Key Characteristics:**

* Subject-Oriented: Organized around business subjects (e.g., sales, customers).
* Integrated: Combines data from different sources into a unified format.
* Non-Volatile: Once stored, data doesn’t change (read-only for analysis).
* Time-Variant: Maintains historical data for trend analysis.

**Architecture:**

Data Warehouse follows a structured architecture to efficiently store, process, and retrieve large volumes of data for analytics and business intelligence.

**1)Bottom Tier (Data Storage & ETL Layer)**

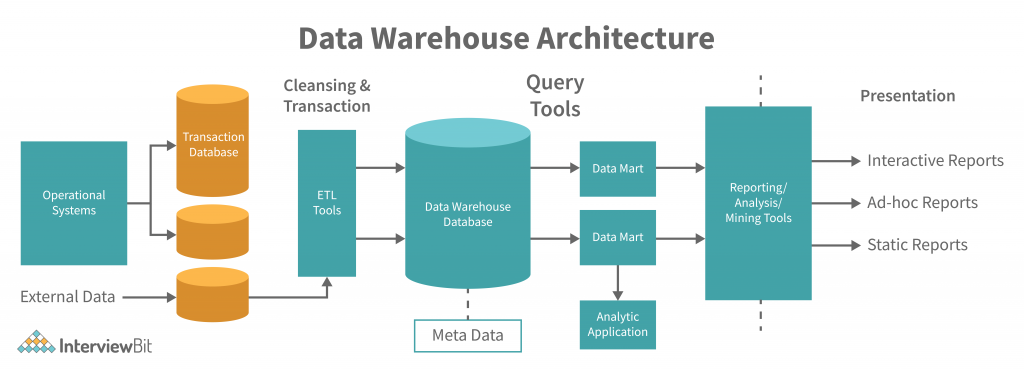
* Collects raw data from various sources and prepares it for storage.
* **Components**:
  + **Source Systems** (Databases, ERP, CRM, logs, APIs, flat files).
  + **ETL (Extract, Transform, Load) :**
    - **Extract**: Pulls data from multiple sources.
    - **Transform**: Cleans, filters, and standardizes data.
    - **Load**: Moves processed data into the DWH.
  + **Data Lake (Optional)**: Stores raw/unstructured data before ETL.

**B. Middle Tier (OLAP Server & Processing Layer)**

* Organizes data for efficient querying and analytical processing.
* **Components**:
  + OLAP (Online Analytical Processing) Engine
  + Data Warehouse Database

**C. Top Tier (Front-End & Reporting Layer)**

* Stores structured data in optimized schemas:
* **Components**:
  + **BI & Visualization Tools** (Power BI, Tableau, Looker).
  + **Ad-hoc Query Tools** (SQL clients, custom dashboards).
  + **AI/ML Integration** (Predictive analytics, automated insights).



**Data Mart:**

A data mart is a specialized subset of a data warehouse designed to serve the analytical needs of a specific business unit, department, or function. It contains curated, subject-oriented data optimized for fast querying and reporting.

**Types of Data Marts:**

1. Dependent Data Mart

* Source: Built from an existing DW (top-down approach).
* Example: Finance team extracts data from the central DW.
* Advantage: Ensures consistency with enterprise data.

2. Independent Data Mart

* Source: Standalone, not linked to a DW (bottom-up approach).
* Example: Marketing creates a mart directly from CRM data.
* Risk: May lead to silos if not integrated later.

3. Hybrid Data Mart

* Combines data from DW and external sources.
* Example: Sales mart enriched with social media trends.

**Advantages:**

* Improves decision-making with consolidated data.
* Enhances data quality and consistency.
* Supports historical analysis and forecasting.

**Uses:**

* Business Intelligence and reporting
* Historical Data Analysis
* Trend forecasting
* Customer and Market behavior analysis
* Real-Time analytics
* AI and Automation integration.