Data Warehousing

A **Data Warehouse (DWH)** is a centralized repository that stores integrated, historical data from multiple sources. It supports **Business Intelligence (BI)**, analytics, and reporting by providing a unified view of organizational data.

Key Characteristics:

Characteristic	Description
Subject-Oriented	Organized by business subjects (e.g., sales, customers).
Integrated	Combines data from multiple sources into a consistent format.
Time-Variant	Tracks changes over time (historical data).
Non-Volatile	Data is read-only; once stored, it doesn't change.

Data Warehouse Architecture:

1 Data Sources

- OLTP databases (e.g., MySQL, PostgreSQL).
- Flat files (CSV, Excel).
- APIs, IoT devices, web logs.

2 ETL Process

- **Extract**: Pull data from sources.
- **Transform**: Clean, standardize, aggregate.
- **Load**: Store in the DWH.

3 Data Warehouse Database

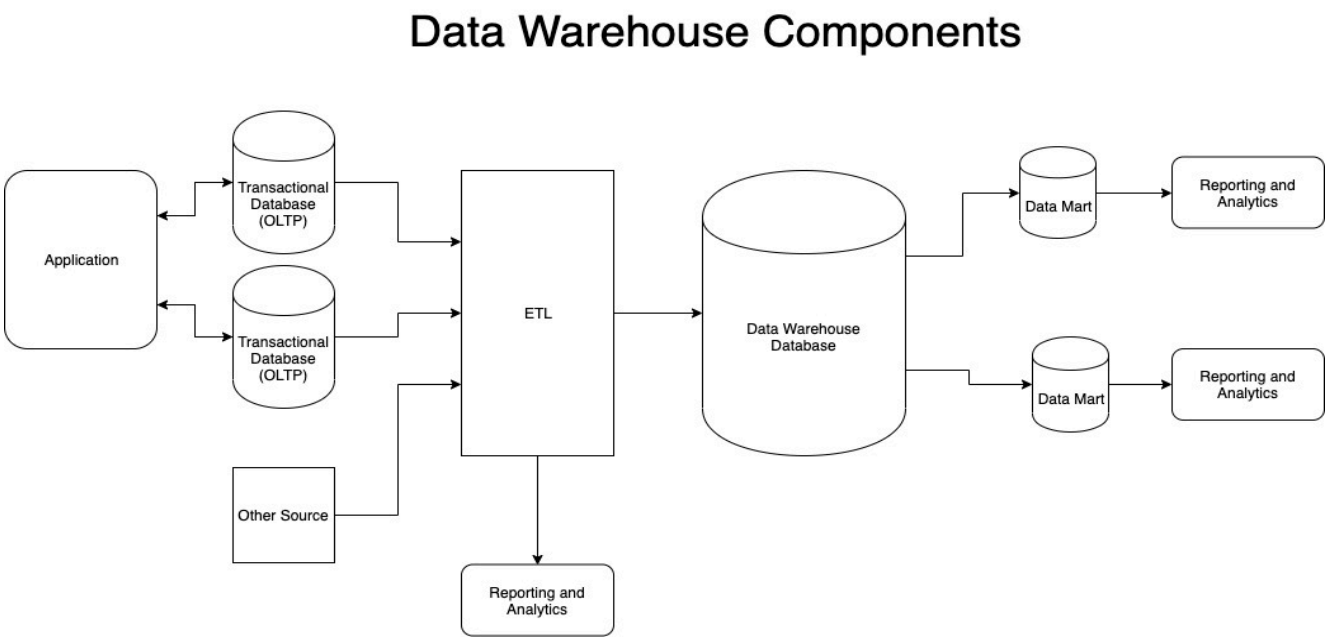
- Optimized for **OLAP** (Online Analytical Processing).
- Uses **star/snowflake schemas**.

4 Data Marts

- Subsets for departments (Sales, HR).
- Types: **Dependent** (from DWH) or **Independent** (standalone).

5 Reporting & Analytics

- Tools: Power BI, Tableau, Looker.
- Dashboards, ad-hoc queries.



ETL Process:

Step	Tools	Example
Extract	Apache NiFi, SSIS	Pull sales data from SQL Server.
Transform	Python (Pandas), dbt	Remove duplicates, calculate revenue.
Load	Snowflake, Redshift	Load into fact/dimension tables.

ETL vs. ELT:

- **ETL**: Transform before loading (traditional).
- **ELT**: Load raw data, transform later (modern, cloud-based).

Data Marts:

Benefits:

- Faster queries for specific teams.
- Simpler access than a full DWH.

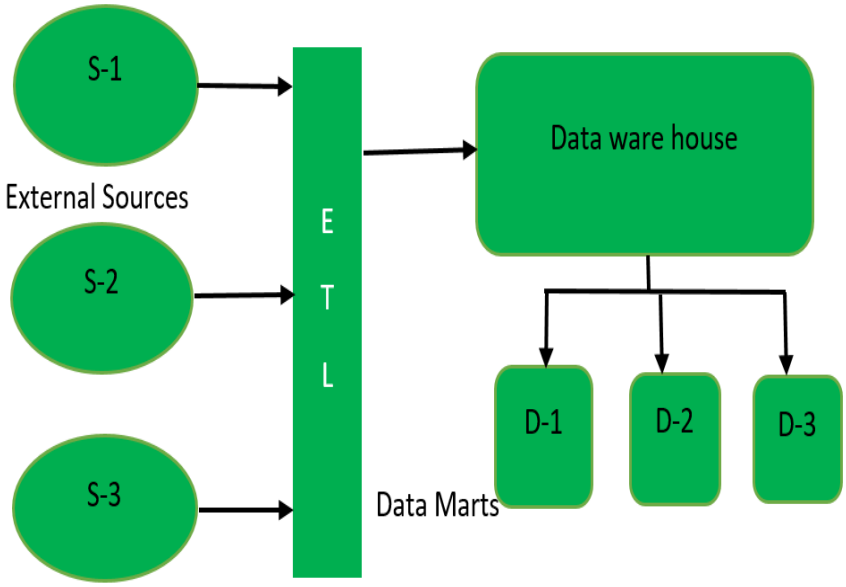
Design Considerations:

- Align with business needs.
- Ensure data consistency with the main DWH.

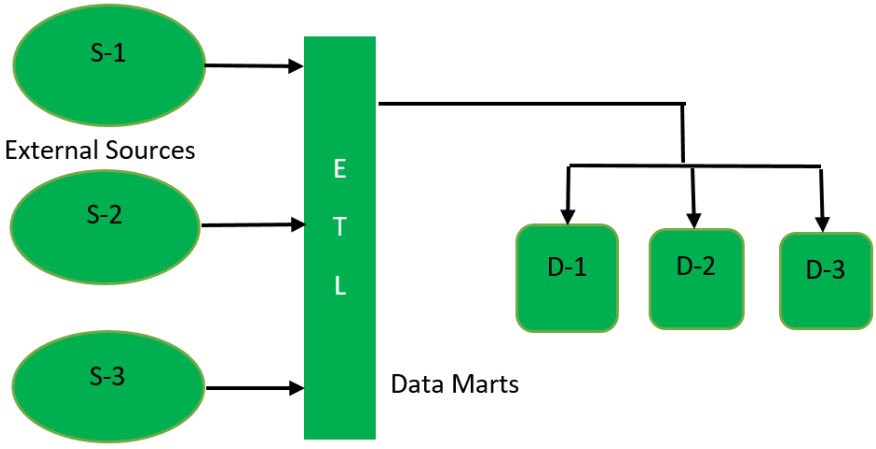
Types of Data Marts:

- Dependent Data Mart: Derived from a central Data Warehouse.
- Independent Data Mart: Built without using a Data Warehouse.

- **Dependent Data Mart:**



- **Independent Data Mart:**



Advantages & Challenges:

Pros	Cons
Single source of truth	High implementation cost
Historical analysis	Complex ETL pipelines
Scalable for big data	Requires maintenance

Use Cases:

- **Retail**: Demand forecasting.
- **Healthcare**: Patient trend analysis.
- **Finance**: Fraud detection.

Conclusion:

Data warehousing is essential for modern analytics. Proper architecture, ETL, and data marts ensure efficient data utilization.