# **Data Warehousing**

It is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from heterogeneous sources.

The primary goal of data warehousing is to provide a unified and consistent view of an organization's data for analytical and reporting purposes.

**FEATURES OF DATA WAREHOUSE**

**Subject-oriented –** A data warehouse is always a subject oriented as it delivers information about a theme instead of organization’s current operations.

* It also delivers an easy and precise demonstration around particular theme by eliminating data which is not required to make the decisions.

**Integrated -** A data warehouse combines data from various sources. These may include a cloud, relational databases, flat files, structured and semi-structured data, metadata, and master data.

* Integration of data warehouse benefits in effective analysis of data.

**Time-Variant -** which means that the data is stored with a time dimension. This allows for easy access to data for specific time periods, such as last quarter or last year.

* which means that the data is stored with a time dimension. This allows for easy access to data for specific time periods, such as last quarter or last year.

**Non-Volatile –** This means that the data in the warehouse is never updated or deleted, only added to.

* This is beneficial in analysing historical data and in comprehension the functionality.

**Decision support systems**

* A decision support system (DSS) is an interactive information system that analyzes large volumes of data for informing business decisions.
* A DSS is a combination of raw data, documents, personal knowledge, and/or business models to help users make decisions.

**Structured and Unstructured components of DSS**

* Every task has a structured and unstructured component.
* Structured components directly help in decision-making, while unstructured components require human interaction with the DSS.

**OLTP**

* OLTP stands for  **Online Transaction Processing**
* Online Transaction Processing is a type of data processing that consists of executing a number of transactions occurring concurrently.
* It refers to systems that facilitate and manage transaction-oriented applications, typically for data entry and retrieval transaction processing.
* Usually run simple queries involving just one or a few records
* OLTP systems implement concurrency control mechanisms to manage multiple transactions happening simultaneously and ensure data consistency.
* Examples include order processing, inventory management, and online banking.

# **OLAP Architecture**

* **OLAP** stands for **OnLine Analytical Processing**.
* It is based on the multidimensional data model.
* It is a technology that enables analysts to extract and view business data from different points of view.
* OLAP data is typically stored in a Star Schema, which is a combination of dimensions and fact tables.
* OLAP systems handle a lower volume of transactions compared to OLTP. They are optimized for read-intensive operations and complex analytical queries.
* OLAP systems are generally not as concerned with concurrent write operations as OLTP systems. The focus is on supporting simultaneous read-intensive analytical queries.

**Data Engineering**

* Data engineering is the process of designing and building systems that let people collect and analyze raw data from multiple sources and formats.
* Data engineering is the discipline which creates data collection, storage, transformation, and analysis processes for large amounts of raw data, structured data, semi-structured data, and unstructured data (e.g., [Big Data](https://www.mongodb.com/basics/big-data-explained)) so that data science professionals can draw valuable insights from it.

**ETL**

* ETL stands for **Extract Transform Load.**
* ETL is the process of combining data from multiple sources into a large, central repository called a data warehouse
* The mechanism of extracting information from source systems and bringing it into the data warehouse is commonly called ETL

**The process of ETL can be broken down into the following three stages:**

1. **Extract :** The first stage in the ETL process is to extract data from various sources such as transactional systems, spreadsheets, and flat files. This step involves reading data from the source systems and storing it in a staging area**.**
2. **Transform :** In this stage, the extracted data is transformed into a format that is suitable for loading into the data warehouse. This may involve cleaning and validating the data, converting data types, combining data from multiple sources, and creating new data fields.
3. **Load :** After the data is transformed, it is loaded into the data warehouse. This step involves creating the physical data structures and loading the data into the warehouse.

**4V’s**

Big data refers to large amounts of data that can inform analysts of trends and patterns related to human behavior and interactions. There are four major components of big data.

* **Volume :**

Volume refers to the amount of data a company can receive from its customers, potential users, indirect clients, suppliers, employees, competitors, markets and operations.

 A large volume of data demands better technology to collect, process, store and analyse it.

* **Velocity :**

Velocity refers to the speed with which multiple sources produce data, which means how fast a source is generating data a computer can process instantaneously.

Data is generated and collected at high speeds. In some cases, data streams in real-time, requiring rapid processing to derive meaningful insights.

* **Variety :**

Variety refers to the type of data an organisation may receive and process. There are three types of data, which include unstructured, semi-structured and structured.

A system manages information according to the type of data is receiving, which means it can apply different algorithms and require distinct types of storage based on the data.

* **Veracity :**

Veracity refers to the reliability and quality of the data. Big data often includes data from diverse sources, and ensuring accuracy can be challenging.

It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.