**DATA ENGINEERING**

S.YAZHINI

B.E - CSE

**1.AN INTRODUCTION TO DATA WAREHOUSING:**

* Data Warehouse:
* These are the bulk number of constructed machines to store data.
* It dumps are stores raw data.
* Data represents the company’s history and the operational data that represents near-term history are also added.
* It is :

**1. Subject-oriented :**

* + arranged according to the subject instead of the application.
  + focuses on the schema of data for decision-makers.

**2. Integrated:** organized by multiple different data sources like relational databases flat files, and online transaction records.

**3.Time-variant:** The period for the data is longer than the operational systems

eg: more than 5-10 yrs

**Non-volatile :**

* + Updation cannot be done.
  + the data entered are never removed.

**2. Purpose of Data Warehouse:**

The main purpose of a data warehouse is to build a historical record that can be invaluable to data scientists and business analysts. Because of these capabilities, a data warehouse can be considered an organization’s “single source of truth”

**Types:**

* **Enterprise Data Warehouse**: A centralized warehouse that provides decision support service across the enterprise.
* **Operational Data Store:** A data store is required when neither data warehouse supports the organization's reporting needs.
* **Data Mart**: A subset of an enterprise's data that is designed to serve a particular business unit or department.

**3. Data Warehouse Architecture:**

A data warehouse is a type of data management system that aggregates data from different sources into a single, central, consistent to store of data.

Data Warehouse Architecture is divided into four types:

* **Source Layer:** This layer contains the data from various sources, including operational systems, external data sources, and other data warehouses.
* **Integration Layer:** This layer is responsible for integrating the data from the source layer and transforming it into a format that can be used by the data warehouse.
* **Storage Layer:** This layer stores the integrated and transformed data in a format that is optimized for querying and analysis.
* **Access Layer:** This layer provides access to the data warehouse for end-users, including business analysts, data scientists, and other stakeholders.

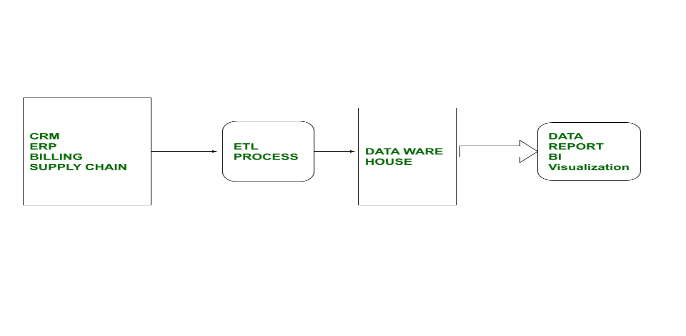
**4. Operational Data Store:**

* An operational data store is a type of database that is used as an interim logical area for a data warehouse, and is designed to integrate data from multiple sources for lightweight data processing activities such as operational reporting and real-time analysis.
* Operational data stores are commonly used in online transaction processing applications, which involve processing transactional data. These applications use the quick, lightweight processing that Operational Data Store tools provide.
* An ODS is connected to multiple data sources and pulls data into a central location.

**5. OLTP Vs Warehouse Applications:**

**Warehouse Applications:**

* Data Warehousing is a technique that gathers or collects data from different sources into a central repository
* It is designed for the decision-making process.
* In Data warehousing, the size of the database is around 100GB-2TB.
* It uses Query processing



**Online-Transaction processing (OLTP) :**

Online-Transaction Processing is a technique used for detailed day-to-day transactions of data that continuously chain on an everyday basis.



**6.Data Marts:**

A data mart is a data storage system that contains information specific to an organization's business unit.

Companies use a data mart to analyze department-specific information more efficiently.

**Types of Data Marts:**

* Dependent
* Independent
* Hybrid

**Data Mart Benefits:**

* Easier access to data
* Faster insight and decisions
* Lower Cost
* Better data access control

**7. Data marts Vs Data Warehouses:**

**Data Ware House :**

* Data warehouse is a Centralized system.
* In the data warehouse light denormalization takes place.
* It is a top-down model.
* Data warehouse is flexible.

**Data marts:**

* It is a decentralized system.
* In data mart highly denormalization takes place.
* It is a bottom-up model.
* Data mart is not flexible.

**Data Warehouse Life cycle:**

Data Warehousing is a flow process used to gather and handle structured and unstructured data from multiple sources into a centralized repository to operate actionable business decisions.

