ASSIGNMENT NO: 1

Name: Pratik Wani

Problem Statement: Explain the Data Warehouse concept and What you have learned in today's session (17th Jan 2024)

Solution:

➢ Data Warehouse:-

- Data Warehouse is a huge amount of data storage which collects the data from multiple heterogeneous sources like Flat files, RDBMS, etc.
- Data Warehouse is different from normal databases, Ordinary databases can only perform Transactional operations but we can't use them for analytical purpose that is why organization need a data warehouse which can keep there historical data for making strategic decisions and analyzing different business trends.
- Data Warehouse is Subject Oriented, Integrated, Time variant and Non Volatile collection of data

➢ Features of Data Warehouse:-

- Subject oriented:
  - Data is stored on the basis of subjects and the main focus in on modeling and analysis of data for decision making
- Integrated:
  - Data is collected from various heterogeneous sources like Binary files, RDBMS and flat files.
  - It maintains the Consistency in data.
- Time-variant:
  - Data Warehouse stores historical data of the organization hence its time horizon is usually large. Ex. Placement ratio of college over last 10 years
- Non-volatile:
  - Once the data entered into the data warehouse, we can't remove that data
  - We can't update the data

➢ Why Data Engineering?

- In today's world there is no monotony in businesses, In each and every section there is competition that's the reason organizations needs to take decisions fast and accurate to become successful.
- According to different researches, many analytical projects failed due to *improper data* here data engineering comes into picture. Data engineers make data accessible so that organizations can use it to evaluate and optimise their performance.
- The main goal is to collect, manage and convert the raw data into *useful information* so that data scientists and analysts can use it to build a systems that helps in company's growth.

➢ Data Types:

- Raw Data:-
  - Unprocessed data
  - No scheme applied

- Processed Data:-
  - Scheme applied
  - Stores data in tables or data pipelines

- Cooked Data:-
  - It is selected Processed data that can be used for analytical purpose

➢ Big Data Properties:

- Volume: amount of data
- Velocity: How fast data is receiving
- Variety: different types of data
- Veracity: Reliability and Consistency in data

➢ Batch processing:

- Data → Storage → Analysis → Information System

➢ Stream Processing:

- Data comes in chunks and then streams . Ex. Netflix

➢ How can we store data?

- RDBMS (Table format data)
- No-SQL (Key-Value pair data)

➢ Decision Support System:

- Organizations need to take the decisions fast to become successful that's the reason they need some information system which helps them in quick decision making and also supports different data sources.

➢ Components of DSS:

- Structured components:-
  - Can directly use for further analytical purpose

- Unstructured components:-
  - Need some processing on data also need some human interactions after that they can be use for further analysis

➢ DSS architectural styles:

- OLAP (online application processing):
  - Primary goal is *data analysis* not data processing
  - Use by data warehouse
  - provides an environment to get insights from the database retrieved from multiple database systems at one time.

- OLTP (online transaction processing):
  - Main goal is *data processing* not data analysis
  - Use by RDBMS
  - Handles the day-to-day transactions in any organization
  - Ex. ATM machine

- OLTP Architectural workflow:

  Raw data → Data Warehouse → Reporting, Mining, Analytical Purpose

- Benefits of OLTP:
  - Simple: Simple to handle the queries
  - Less paper work
  - Faster and Accurate results
  - It has different constraints to main the consistency and integrity in database

- Drawbacks of OLTP:
  - Need to update it frequently
  - We can't use these table format data for analytical purpose
  - To get some simple data we need to use subquery and joins
  - Because of these reasons we need some central data house which keeps the huge historical data and can be used for analytical projects that's reason *Data warehouses* comes in picture