ASSIGNMENT-1

QUESTION:

Explain Data Engineering and its importance & Explain Properties of Big Data & Explain OLTP & Explain Data Warehouse.

SOLUTION:

Data Engineering:

Data Engineering is a process of designing, building and scaling systems that allows users to collect and analyse data from multiple sources existing in different forms.

Data is of three types. They are: 1) Raw Data
2) Processed Data
3) Cooked Data

- Raw Data is unprocessed, unorganized, and unstructured data that has been collected directly from various sources.
- Processed Data is information that has undergone a series of manipulations to make it more structured, organized, and suitable for specific purposes.
- Cooked Data is the processed data that has been summarized.

Importance of Data Engineering:

Data Engineering is important to solve various critical business problems. It is used in Data Availability and Accessibility, Data Quality Assurance, Efficient Data Processing, Scalability, Security and Cost Optimization. It helps in making better decisions.

Properties of Big Data:

1) Volume
2) Velocity
3) Variety
4) Veracity

1) Volume: Volume refers to quantity of data generated or collected from various sources that is available in different formats.

2) Velocity: Velocity refers to the speed at which data is generated, collected, and processed.

3) Variety: Data with different data types is stored from different sources which includes structured data, semi-structured data and unstructured data.

4) Veracity: It refers to accuracy and reliability of data.

OLTP:

1. OLTP refers to Online Transaction Processing.
2. It allows users to access large amount of data.
3. OLTP systems are designed for processing day-to-day, operational transactions that occur in real-time and used by traditional operational systems i.e., RDBMS.
4. These transactions typically involve adding, updating, or retrieving small amounts of data, and they are critical for the daily operations of an organization.
5. OLTP systems ensures to maintain ACID properties i.e., Atomicity, Consistency, Isolation and Durability.
6. OLTP handles the daily transactions and interactions with customers.
7. They ensure that data is accurately and promptly recorded.
8. OLTP systems are mainly used e-commerce, banking and many other applications where real time processing is required.

   Example: **Online Book Store**
   Consider an online bookstore where customers can browse, search, and purchase books.
   1) **Inserts:** When a new customer creates an account or places an order, new data is added to the system.
   2) **Updates:** If a customer updates their shipping address or modifies their order, the system needs to reflect those changes immediately.
   3) **Retrieves:** When a customer searches for a book or checks the status of their order, the system retrieves and displays the relevant data.

Advantages of OLTP

1) Data Manipulation is easy.
2) The tasks include insertion, updation, or deletion of data which allows users to use efficiently.
3) It supports bigger databases.
4) It ensures to maintain ACID properties.
5) Maintain Data Integrity.
6) Faster query processing.

Disadvantages of OLTP

1) OLTP requires instant update.
2) Data obtained from OLTP is not suitable for analysis.

DATA WAREHOUSE:

Data warehouse is a database that allows organizations to store, manipulate and analyze large amounts of data which is collected from various sources . It mainly helps in making effective decisions by the organizations.

Elements of Data Warehouse:

1) Relational database (to store and manage data).
2) ETL (Extraction, Transformation and Load)
3) Statistical analysis, reporting, and data mining capabilities
4) Client analysis tools for visualizing and presenting data to business users.

Characteristics of Data Warehouse:

Characteristics of data warehouse include subject-oriented, Integrated, Time Variant and non-volatile.

1) **Subject-Oriented:**
   A data warehouse is also subject-oriented, which means that the data is organized around specific subjects, such as customers, products, or sales. This allows for easy access to the data relevant to a specific subject, as well as the ability to track the data over time.

2) **Integrated:**

Data from different sources is integrated into the data warehouse in a standardized format. Integration ensures consistency and accuracy, allowing users to analyze data across the organization without worrying about issues that occur.

3) **Time Variant:**

The data is stored with a time dimension. This allows users for easy access to data for specific time periods, such as last quarter or last year. This makes it possible to track trends and patterns over time. Data warehouses include historical data and support time-based analysis.

4) **Non-Volatile:**

Data in the warehouse is never updated or deleted. This is important because it allows for the preservation of historical data, making it possible to track trends and patterns over time. This ensures data consistency for reporting and analysis.

Name: Vuthur Sriganga
Hexaware Data Engineering Batch-1