Akilesh K

k.akilesh123@gmail.com

Data engineering - Batch 1

DATA WAREHOUSING

Problem statement – what do you understand by Data warehousing

Data warehousing is a centralized repository of data collected from various sources within an organization. The purpose of a data warehouse is to support decision-making processes by providing a comprehensive and historical view of the organization's data.

Benefits of Data Warehousing:

- Improved decision making through access to timely information.
- Enhanced data quality and consistency.
- Historical analysis and trend identification.
- Support for business intelligence and reporting.

Characteristics of Data warehousing

Subject-oriented - Organizing and focusing data around a specific subject, making it more relevant and meaningful.

Integrated - Combining data from different sources to provide a comprehensive view or analysis.

Time-variant - Recognizing that data can change over time, and capturing and storing historical information.

Nonvolatile - Data that remains persistent and does not change or get overwritten easily.

Decision Support System(DSS)

A Decision Support System helps organizations make decisions by providing access to relevant data, analysis tools, facilitating the decision making process

Structured Components of DSS:

It involves organized and well defined data. This includes data stored that can be easily queried and analyzed using structured methods.

Unstructured Components of DSS:

It is less organized and more varied data types. This includes text, images, and other non-tabular formats. Analyzing and processing unstructured data often require advanced techniques.

Operational database

An operational database is a type of database specifically designed and optimized for supporting day-to-day transactional processes and activities within an organization.

OLTP (Online Transaction Processing):

OLTP is used by traditional operational systems, typically Relational Database Management Systems. It handles transactional operations, focusing on quick and efficient processing of high volumes of short, real-time transactions. Examples include order processing, inventory management, and online banking transactions.

Advantages of OLTP

- optimized for quick and efficient processing of large numbers of transactions
- ensures data integrity and consistency by enforcing database constraints and transactional rules, maintaining the accuracy of business data.

Disadvantages of OLTP

- They are not designed for complex analytical queries, reporting, or data analysis.
- It will struggle with complex queries that involve multiple joins and aggregations

OLAP (Online Analytical Processing):

OLAP, is a category of software tools and applications designed for complex analysis of large volumes of multidimensional data.

Data engineering

Data engineering is the process of designing, developing, and managing the architecture, tools, and systems for collecting, storing, and processing large volumes of data.

ETL- Extract, Transform, and Load

Types of data

Raw data - Unprocessed and unorganized information straight from the source.

Processed data - Raw data that has been cleaned, organized, and transformed to make it useful for analysis or presentation.

Cooked data – Processed data that has undergone further refinement and analysis

Big data properties

Volume - amount of data generated or collected.

Velocity - The speed at which new data is generated and how quickly it needs to be processed.

Variety - The different types and sources of data, including structured and unstructured formats.

Veracity - The accuracy and reliability of the data being collected and processed.

Processing methods

Batch processing - Handling and processing a set of data all at once, typically in large batches or groups.

Stream processing - Dealing with data in real-time as it is generated or received, allowing for continuous and immediate analysis.

Data storage

Relational database - Organizes data into tables with predefined relationships between them.

Document store - Stores and retrieves data in flexible, document-like structures without requiring a fixed schema.