Name: Moiz Zulfiqar

Day1: Task 1

Big Data

Big data refers to large, complex and diverse data sets that are difficult to manage and process using traditional methods. The concept of big data gained momentum in the early 2000s when industry analyst Doug Laney defined it as the three V's: volume, velocity, and variety. Volume refers to the sheer amount of data, velocity refers to the speed at which data is generated and processed, and variety refers to the different types of data.

Big data has become increasingly important as organizations seek to analyze the data to gain insights and make data-driven decisions. The value of big data lies not only in the amount of data, but also in how it is analyzed and used. By analyzing big data, businesses can streamline resource management, improve operational efficiencies, optimize product development, drive revenue growth, and enable smart decision-making.

Big data is changing the way organizations manage and derive insights from data. The use of big data technologies and cloud computing has made it more cost-effective to handle all types of data. The success of big data projects also relies on the collaboration of various roles, including data scientists, data engineers, and data officers.

In summary, big data refers to the large, diverse, and complex data sets that inundate businesses on a daily basis. By analyzing big data, organizations can gain valuable insights to make data-driven decisions and improve operational efficiencies. The use of big data technologies and cloud computing has made it more cost-effective to handle all types of data.

Data Lake

A data lake is a centralized repository that stores and processes large amounts of data in its native format, including structured, semi-structured, and unstructured data. It offers a scalable and secure platform that enables organizations to ingest any type of data from any system at any speed, store any volume of data in full fidelity, process data in real-time or batch mode, and analyze data using various programming languages or third-party data analytics applications.

Unlike traditional data warehouses, the data lake emphasizes the flexibility and availability of data. Data can reside in its native format until it is needed, and users can determine the specific data types and sources they need, how much they need, when they need it, and the types of analytics they need to derive.

The data lake is a new concept, and there are different schools of thought regarding its importance and applicability to enterprises. Some view it as essential for data-driven companies, especially for handling vast streams of unstructured data and achieving faster response times and analytics. However, skeptics consider it a new name for an old concept with limited applicability.

In summary, the data lake is a promising solution for storing, processing, and analyzing large amounts of data in a flexible and scalable way. Its future looks bright, as organizations increasingly need to integrate smaller data with big data and answer critical business questions faster than traditional data storage and reporting tools can provide.

Database

A database is an organized collection of structured information or data, usually stored electronically in a computer system. It is typically controlled by a database management system (DBMS). Databases are commonly modeled in rows and columns in a series of tables to facilitate processing and querying of data. Most databases use Structured Query Language (SQL) for writing and querying data. SQL is a programming language used by nearly all relational databases to query, manipulate, and define data and to provide access control.

Databases differ from spreadsheets in terms of how data is stored and manipulated, who can access the data, and how much data can be stored. Databases are designed to hold much larger collections of organized information and allow multiple users to access and query data simultaneously.

There are several types of databases, each with unique features and functions. The types of databases include relational databases, object-oriented databases, distributed databases, data warehouses, NoSQL databases, graph databases, and OLTP databases. Other less common databases are tailored to specific functions. Newer types of databases include open source databases, cloud databases, multimodel databases, document/JSON databases, and self-driving databases.

Database software is used to create, edit, and maintain database files and records. It handles data storage, backup, and reporting, multi-access control, and security. Database software simplifies data management by allowing users to store data in a structured form and access it through a graphical interface.

A database management system (DBMS) is a comprehensive database software program that serves as an interface between the database and its end users or programs, allowing users to retrieve, update, and manage data. The DBMS also provides a security system that regulates access to the database.

Data Warehouse

A data warehouse is a type of data management system that stores large amounts of historical data derived from various sources, such as application log files and transaction applications. It is designed to enable and support business intelligence activities, especially analytics, and to allow organizations to derive valuable business insights from their data to improve decision-making. A data warehouse centralizes and consolidates large amounts of data from multiple sources and is considered an organization’s “single source of truth.”

There are four unique characteristics that define a data warehouse: it is subject-oriented, integrated, nonvolatile, and time-variant. A well-designed data warehouse will perform queries quickly, deliver high data throughput, and provide enough flexibility for end-users to analyze the data for closer examination.

The architecture of a data warehouse is determined by the organization's specific needs, but it generally includes a relational database to store and manage data, an extraction, loading, and transformation (ELT) solution for preparing the data for analysis, statistical analysis, reporting, and data mining capabilities, and client analysis tools for visualizing and presenting data to business users.

A modern data warehouse includes a converged database that simplifies management of all data types, self-service data ingestion and transformation services, support for SQL, machine learning, graph, and spatial processing, multiple analytics options that make it easy to use data without moving it, and automated management for simple provisioning, scaling, and administration.

The expansion of big data and the application of new digital technologies are driving change in data warehouse requirements and capabilities. AI and machine learning are transforming almost every industry, service, and enterprise asset, and data warehouses are no exception.

References

- Big Data: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html
- Data Lake: https://cloud.google.com/learn/what-is-a-data-lake#:~:text=A%20data%20lake%20is%20a,of%20it%2C%20ignoring%20size%20limits, https://www.sas.com/en_us/insights/articles/data-management/data-lake-and-data-warehouse-know-the-difference.html
- Database: https://www.oracle.com/pk/database/what-is-database/#:~:text=A%20database%20is%20an%20organized,database%20management%20system%20(DBMS)
- Data Warehouse: https://www.oracle.com/pk/database/what-is-a-data-warehouse/