

***DATA ENGINEERING***

***TASK-01***

*Submitted to*

***Sir Muhammad Ahtisham***

*Submitted by*

***Aabra Malik***

***Big Data***

Big data refers to extremely large and complex data sets that cannot be easily processed, stored, or analyzed using traditional data management tools or methods. Big data typically refers to data sets that are too large to be handled by a single machine and require distributed computing frameworks and tools.

Big data is characterized by three key characteristics: **volume, velocity, and variety.**

***Volume*** refers to the large amount of data generated and collected from various sources, such as social media, sensors, and mobile devices. These data sets can range from terabytes to petabytes or even exabytes in size, making it difficult to store and process.

***Velocity*** refers to the high speed at which data is generated and must be processed in real-time or near real-time. This is particularly true for data streams, such as social media feeds, stock market data, or sensor data, which can generate huge amounts of data in a very short time.

***Variety*** refers to the different types of data that make up a big data set. These data types can be structured, semi-structured, or unstructured, and can come from a variety of sources, including text, images, video, and audio. This adds complexity to the processing and analysis of big data, as different tools and techniques may be required to handle each data type.

***Storing and processing big data*** requires specialized tools and technologies that can handle the large volume, high velocity, and diverse variety of data. Here are some of the common methods and technologies used to store and process big data:

Distributed file systems: These are designed to store and manage data across multiple nodes in a cluster. Examples include the Hadoop Distributed File System (HDFS) and Apache Cassandra.

Cloud computing platforms: Cloud-based storage and processing platforms like Amazon Web Services (AWS) and Microsoft Azure provide scalable and cost-effective solutions for storing and processing big data.

***Data Lake***

A data lake is a centralized repository that stores raw, unstructured, semi-structured, and structured data at scale.

Data lakes are designed to handle vast amounts of data, ranging from terabytes to petabytes, and are intended to support a wide range of use cases, such as data exploration, machine learning, and analytics.

In data engineering, data lakes are typically built using distributed file systems, such as Hadoop Distributed File System (HDFS), and are accompanied by a suite of tools for data ingestion, processing, and analysis, such as Apache Spark.

Data lakes differ from traditional data warehouses in that they are not pre-defined and rigidly structured, but rather flexible and adaptable to changing data requirements. This makes them an ideal solution for organizations that deal with large volumes of diverse and rapidly changing data.

***Database***

A database is a structured collection of data that is organized in a way that allows efficient storage, retrieval, and manipulation of information. Databases are used to store information for a wide range of applications, including business, scientific research, education, and more.

In a database, data is organized into tables, which consist of rows and columns. Each row in a table represents a single record or instance of an entity, while each column represents a specific attribute or characteristic of that entity. For example, in a database for an e-commerce website, a table might store customer information, with columns for each customer's name, address, email, and order history.

***Data Warehouse***

A data warehouse is a large-scale, centralized repository of data that is specifically designed to support business intelligence (BI) activities, such as data analysis, reporting, and decision-making.

Data warehouses are typically constructed from data that has been extracted, transformed, and loaded (ETL) from a variety of disparate sources, such as transactional databases, web services, and other data sources. The data is then transformed and organized into a common data model that is optimized for querying and analysis.

Overall, data warehouses are an essential tool for businesses and organizations that need to make data-driven decisions based on large volumes of complex data.

***\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\****