TASK 2.1

Write a short note differentiating big and small data

Big Data and Small Data are terms used to describe the scale and scope of data that organizations work with, and they have significant differences in terms of their characteristics and applications. Here's a short note differentiating the two:

Big Data:

* Volume: Big Data refers to a massive volume of data that exceeds the capacity of traditional databases and data processing systems. It typically involves terabytes, petabytes, or even larger datasets.
* Variety: Big Data is characterized by its diverse nature, encompassing structured, semi-structured, and unstructured data. This includes text, images, videos, sensor data, social media interactions, and more.
* Velocity: The data in Big Data scenarios is generated and collected at high speeds in real-time or near real-time. The rapid influx of data requires efficient processing and storage solutions.
* Complexity: Dealing with Big Data often involves complex processing and analysis techniques like data mining, machine learning, and artificial intelligence to derive meaningful insights.
* Use Cases: Big Data is commonly used for large-scale analytics, predictive modeling, and business intelligence, where finding patterns and trends in massive datasets can lead to valuable insights and decision-making.

Small Data:

* Volume: Small Data refers to a relatively small amount of data that can be easily handled by traditional databases and simple data analysis tools. It typically involves gigabytes or smaller datasets.
* Variety: Small Data is usually structured and uniform, making it easier to organize and analyze compared to the diverse and unstructured nature of Big Data.
* Velocity: The rate of data generation and collection in Small Data scenarios is moderate and does not require specialized real-time processing capabilities.
* Complexity: Analyzing Small Data is often less complex, and traditional statistical methods and basic analytics techniques are sufficient to gain insights from the data.
* Use Cases: Small Data is commonly used for specific, targeted analyses, such as market research, customer surveys, and individual behavior analysis. It is particularly useful when detailed and in-depth information about a limited number of entities is required.

TASK 2.2

DATA LIFE CYCLE

The data life cycle refers to the different stages that data goes through from its initial collection or creation to its eventual deletion or archiving. It encompasses the entire journey of data within an organization or system, including its acquisition, storage, processing, analysis, and disposal. Understanding the data life cycle is crucial for effectively managing data and ensuring its quality, security, and usefulness. The data life cycle typically consists of the following stages:

* Data Collection: This is the first stage of the data life cycle, where data is gathered from various sources. It can be collected through manual data entry, sensors, web scraping, APIs, surveys, transactions, or other means. The data collected should align with the organization's objectives and be relevant to the intended analysis.
* Data Storage: Once collected, the data needs to be stored securely and efficiently. Data storage can be in various forms, such as databases, data warehouses, data lakes, or cloud storage. Proper data management practices are essential to maintain data integrity and accessibility.
* Data Processing: In this stage, the collected data is preprocessed, cleaned, and transformed to prepare it for analysis. Data processing involves tasks like data normalization, deduplication, imputation of missing values, and data formatting.
* Data Analysis: Once the data is processed and prepared, it is analyzed to extract insights and meaningful information. This stage involves applying various statistical, machine learning, or data mining techniques to identify patterns, trends, correlations, and relationships within the data.
* Data Visualization: Data visualization is an important part of the data life cycle, where the analyzed data is presented in a visually understandable format, such as charts, graphs, dashboards, or reports. Effective data visualization aids in conveying complex information in a more accessible and actionable manner.
* Data Interpretation and Decision Making: After analyzing and visualizing the data, the insights gained are interpreted and used to make informed decisions, formulate strategies, or solve problems.
* Data Archiving and Retention: Data that is no longer actively used but may have historical value or legal requirements is archived and stored for future reference. Data retention policies define how long data should be kept before it is deleted or disposed of securely to comply with regulations and privacy standards.
* Data Deletion: Data that has reached the end of its life cycle and is no longer needed or legally required should be securely deleted to avoid any potential privacy or security risks.