ByteWise Fellowship – Week 1

Module: DE Basics [1] Date: 14-March-2023

Question#1

**What is BigData?**

Big Data is referred as enormous amount of data produced which cannot be processed conventional method and tools.

**Need for BigData:**

The conventional tools are unable to process the data as the rate at which it processes the data currently is very slow and it can lead to inaccurate results as well. This puts a limit on data analytics, to solve these issues related to processing large amount of data “Big Data” comes into play.

They work on the principle of **5V’s**

* Volume -> Ability to store exabytes+ data
* Velocity -> Processing large amount of data
* Variety -> Storing Structured/Semi-Structured/Unstructured data
* Veracity -> Accurate Processing
* Value -> Improved data analytics results

**Tools for Big Data:**

* Apache Hadoop
* Apache Spark
* Cassandra

**How does BigData work?**

Firstly, the data is stored and then it is broken down into smaller chunks via “File Distribution System” and then its processed on multiple machines. The system is made in such a way that if one machine fails the others keep working and replicas of the data is also made for quick recovery and processing.

Hadoop works on the same principle explained above called HFDS (Hadoop file distribution system)

Question#2

**What is Data Lake?**

It is a data storage repository that helps to store data in any form (Structured/Semi-Structured/Unstructured data) that can be easily accessed as well.

**Why was it needed?**

Data Lake came into being when large amount of data was to be stored,processed and then put up for consumers at a faster and economical rate than the conventional tools. Big Data can store data multiple times more than a data lake but it cannot be accessed as rapidly as a data lake.

**Working of a Data Lake:**

* Ingest (Data Collection and Ingestion)
* Store (Data Storage and Management)
* Process (Data Processing and Transformation)
* Consume (Data Access and Retrieval)

**Tools of Data Lake:**

* Apache Kafka (Ingestions)
* Apache Hadoop, Amazon S3 (Storage)
* Apache Spark (Process)

Question#3

**What is Data Warehouse?**

A data warehouse is a central repository of data that is used to support business intelligence activities such as reporting, analysis, and data mining

It is designed to store and manage large volumes of historical and current data from various sources in a structured and easily accessible format.

**Need for a DWH:**

Data warehouses are designed to solve the problem of data by large number of producers by integrating data from various sources, cleaning and transforming it into a consistent format, and storing it in a centralized location. This makes it easier for business users to access and analyze the data, leading to better decision-making and improved business outcomes.

**Working:**

* ETL Pipeline
* Data Modelling - designing the structure of the data warehouse, including the schema and relationships between tables.
* Data Storage - stores data in a structured format that is optimized for querying and analysis
* Data Access

**Tools:**  
ETL Tools – Informatica

DBMS – MySQL Server

BI Tools – Power BI

Question#4

**What is Database?**

A database is a structured collection of data that is organized in a way that helps in:

* Storage
* Retrieval
* Manipulation

Databases are used to manage large amounts of data in a variety of settings, including businesses, governments, and scientific research.

**Key Components of a database:**

* Tables – That include Rows and Columns
* Indexes – For faster data retrieval
* Primary and Foreign Key

**Types of Databases:**

* NoSQL databases - are designed to handle large volumes of unstructured or semi-structured data. They do not use tables and are typically used for storing data that cannot be easily organized into a table structure, such as documents, images, or videos.
* Object-oriented databases - are based on the object-oriented programming and are used to store complex objects and their relationships.
* Graph databases - are used to store data that has complex relationships between entities.