Task 1

14 March 2023

Big Data :

1. Big Data is a term used to describe data that is extremely large and complex and cannot be processed or analyzed using traditional processing frameworks.
2. It is often described using three v’s: volume, velocity and variety.
3. Volume refers to the sheer amount of data that is generated and can range from terabytes to petabytes and even more. The data is collected from variety of sources such as web, images, videos, iot devices etc. Everyday a huge amount of data is being produced which is estimated to be around 2.5 quintillion bytes per day.
4. Velocity refers to the speed at which the data is generated from these sources. With the arrival of sources which generate data in real time, the velocity of data being generated is growing at exceptional rate. It is getting more and more difficult for organisations to keep up with this and extract valuable insights from it.
5. Variety refers to different types of data that can be structured,semi-structured or non structured. Structured data is data which is organised and is easily searchable such as data in tables, spreadsheets etc. Non structured data refers to images, video, audio and text which cannot be ordered and organised. Semi structured data comprises of files such as JSON and XML. It falls between structured and non structured data.
6. In addition to these three v’s of big data there are more v’s as well but they are not recognised or accepted as the original three.
7. To effectively analyze and process big data, new technologies such as parallel processing, distributed computing, machine learning are used. The insights gained from big data can be used in decision making and improve operations efficiency in organisations.

Data Lake:

1. Data lake is a centralised repository which is used to store all kind of data whether it be structured, semi structured or non structured.
2. Unlike data warehouse the schema is not needed to be defined, meaning that you can store all data without worrying about the schema/structure of data.
3. The data in the data lake doesnt need to be preprocessed or structured before it is stored. Instead the data is stored in its original raw format from where the rest of the operations on it can be performed.
4. They are often used for machine learning,data analytics and other advanced analytical applications.
5. They can be hosted on-premise or on cloud. Many organisations use Amazon’s S3, Google cloud storage or HDFS(on premise).
6. Data lakes are becoming increasingly important because of all the data at one place, all autorized people can access the data and perform broad data exploration and discovery.

Database:

1. Database is a collection of structured data that is organised and stored in a way that is easier to manage, retrieve and manipulate.It is widely used to store and manage data by various organisations throughout the world.
2. The database consists of tables where each table comprises of rows and columns.The data is stored in it according to the schema defined.
3. Relational databases are the most widely used type of database and are based on the relational model, which organizes data into tables with relationships between them.
4. Databases are managed by a database management system (DBMS), which is a software system that controls the organization, storage, and retrieval of data from the database. Some popular DBMSs include MySQL, Oracle, Microsoft SQL Server, and PostgreSQL.

Datawarehouse:

1. A data warehouse is a large, centralized repository of data that is designed to support business intelligence (BI) activities, such as reporting, analytics, and data mining. It is a database that is specifically optimized for querying and analyzing large volumes of data quickly and efficiently.
2. The data reaches it though a process called ETL(Extract Transform Load), the data is cleaned and prepared for BI activities.
3. It stores summarised data from various sources such as databases and employ OLAP to analyze data. The benefits of a data warehouse include better data quality, faster business insights, smarter decision-making, and gaining and growing competitive advantage. Amazon’s Redshift is a good example of data warehouse.