By: Zia Ur Rehman

Date: 14-mar-2023

**Big Data**

Every second millions of gigabytes of data is generated through different data sources like social media, online transactions etc. These datasets cannot be processed with traditional data processing applications and technologies.

Big data is characterized by the 5Vs:

* Volume: extremely large amounts of Data.
* Velocity: the high speed at which it is generated and processed often times in real time
* Veracity: the accuracy and quality of data
* Variety: the different types of data i.e. structured, unstructured and semi-structured
* Value: the usability and effectiveness of the data

Big data is stored and processed differently than small amounts of data. Data engineers use Hadoop, spark and NoSQL databases to store and process it. For example Hadoop uses Hadoop Distributed File System (HDFS) to store the data. HDFS instead of storing one single file on one computer, it breaks it down in chunks and stores it on a number of different computers. This means if one computer breaks down the data is safe in other computers. It also use MapReduce to breakdown tasks and process them in a parallel fashion, this is called “Parallel Computing”.

Big data analytics is used to extract insights and make informed decisions in various fields such as business, healthcare, finance, and science.

**Data Lake**

A data lake is a central repository that allows organizations to store all their structured and unstructured data at any scale. It allows data to be stored in its raw form, without requiring it to be structured or transformed upfront, making it easier to store and manage large volumes of data. Data lakes are designed to support big data processing and analytics, and can store data from a variety of sources. They provide a single location for data storage, enabling organizations to analyze and extract insights from data in real-time.

Data lakes can be accessed and queried using a variety of tools and technologies (e.g. SQL, Spark etc.), depending on the specific needs of the organization.

**Database**

A database is an organized collection of data that is designed to be easily accessed, managed, and updated. Databases can be used to store information about various entities, such as customers, products, employees, transactions, and more. Databases are typically organized into tables, which consist of rows and columns. Each row represents a single record, and each column represents a specific attribute of the record. Such databases are also called relational databases.

Databases are typically managed by a database management system (DBMS), which is a software that allows users to create, update, and query data in the database. Common DBMSs include Oracle, MySQL, Microsoft SQL Server, and PostgreSQL. All of these are Relational Database Management systems (RDBMS).

There are also non-relational databases like NoSQL databases. NoSQL (Not Only SQL) databases are a type of database management system that are designed to handle large volumes of unstructured or semi-structured data. Unlike traditional Relational Database Management Systems (RDBMS), NoSQL databases are schema-free and allow for flexible and scalable data models. They are often used in modern web applications, mobile apps, and big data systems. Examples of such databases are MongoDB, Cassandra and Redis.

**Data Warehouse**

A Data Warehouse is a centralized repository that stores all the best quality data in one place. This data is historical in nature and is brought in from various sources. The data is then transformed and served to its intended users.

Data warehouses are designed to support analytical queries and reporting, rather than transactional processing. They are optimized for read-heavy workloads, and are typically updated on a regular basis (e.g. daily or weekly) rather than in real-time.

Some of the key components of a data warehouse include:

* Extraction, Transformation, and Loading (ETL) processes: These processes are used to extract data from various sources, transform it, and load it into the data warehouse.
* Data modeling: Data in the data warehouse is typically organized into dimensional models, which are optimized for reporting and analysis.
* Data storage: Data in the data warehouse is typically stored in a highly optimized and compressed format, such as columnar storage.
* Query and reporting: Data warehouses typically provide powerful query and reporting capabilities, allowing users to analyze and visualize data in a variety of ways.