**MUHAMMAD HUZAIFA DILSHAD**

**DATA ENGINEERING**

**TASK NO 1**

**BIG DATA**

Big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs. Big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can’t manage them. But these massive volumes of data can be used to address business problems you wouldn’t have been able to tackle before.

**THREE V’S OF BIG DATA**

**Volume:**

With big data, you’ll have to process high volumes of low-density, unstructured data. This can be data of unknown value, such as Twitter data feeds, clickstreams on a web page or a mobile app, or sensor-enabled equipment. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes.

**Velocity:**

Velocity in big data refers to speed at which data is processed and analyzed. With the growth in data, traditional data processing methods have become insufficient to handle this volume and variety of data. As a result new technologies and frameworks have been developed to process large amount of data.

**Variety:**

Variety refers to the many types of data that are available. Traditional data types were structured and fit neatly in a [relational database](https://www.oracle.com/in/database/what-is-a-relational-database/). With the rise of big data, data comes in new unstructured data types such as text, audio, and video, require additional preprocessing to derive meaning and support metadata.

**HISTORY OF BIG DATA**

The concept of big data is new but its origin of large data sets in 1960’s and 70’s. Around 2005, people began to realize just how much data users generated through Facebook, YouTube, and other online services. Hadoop was developed that same year. NoSQL also began to gain popularity during this time

While big data has come far, its usefulness is only just beginning. Cloud computing has expanded big data possibilities even further. The cloud offers truly elastic scalability, where developers can simply spin up ad hoc clusters to test a subset of data. And [graph databases](https://www.oracle.com/in/autonomous-database/what-is-graph-database/) are becoming increasingly important as well, with their ability to display massive amounts of data in a way that makes analytics fast and comprehensive.

**BIG DATA CHALLENGES**

First, big data is…big. Although new technologies have been developed for data storage, data volumes are doubling in size about every two years. Organizations still struggle to keep pace with their data and find ways to effectively store it. But it’s not enough to just store the data. Data must be used to be valuable and that depends on curation. Clean data, or data that’s relevant to the client and organized in a way that enables meaningful analysis, requires a lot of work. Data scientists spend 50 to 80 percent of their time curating and preparing data before it can actually be used.

**DATA LAKE:**

A data lake is a centralized repository designed to store, process, and secure large amounts of structured, semi structured, and unstructured data. It can store data in its native format and process any variety of it, ignoring size limits.

**DATABASE:**

A database is an organized collection of structured information, or data, typically stored electronically in a computer system. Information is organized in rows and columns. Some of the examples of database are oracle, MySQL. Data in database is accessed through SQL.

**DATA WAREHOUSE:**

It is the special type of database that is used for analytic and reporting purposes. This analytic can be used for further enhancing your business. The data of database is archived In data warehouse to be used further. Data is transferred from database to data warehouse through ETL. Data in database is accessed through power bi and excel.