**TASK NO: 1**

**BIG DATA:**

Big Data is a collection of huge and complex data which keeps growing exponentially with time and can’t be processed using traditional methods.

Big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs.

## **The three Vs of big data:**

**Volume:**

The amount of data matters. 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 is the fast rate at which data is received and (perhaps) acted on. Normally, the highest velocity of data streams directly into memory versus being written to disk. Some internet-enabled smart products operate in real time or near real time and will require real-time evaluation and action.

**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/database/what-is-a-relational-database/). With the rise of big data, data comes in new unstructured data types. Unstructured and semi-structured data types, such as text, audio, and video, require additional preprocessing to derive meaning and support metadata.

## **The value—and truth—of big data:**

Two more Vs have emerged over the past few years: **value** and **veracity**. Data has intrinsic value. But it’s of no use until that value is discovered. Equally important: How truthful is your data—and how much can you rely on it?

Both value and veracity help define the quality and insights gathered from data.

**X---------------------------------------X**

**DATABASE:**

A database is a collection of data or information.

A database stores the current data required to power an application.

**Database characteristics:**

A variety of [database types](https://www.mongodb.com/databases/types) have emerged over the last several decades. All databases store information, but each database will have its own characteristics. Relational databases store data in tables with fixed rows and columns. Non-relational databases (also known as No SQL databases) store data in a variety of models including [JSON (JavaScript Object Notation), BSON (Binary JSON)](https://www.mongodb.com/json-and-bson), key-value pairs, tables with rows and dynamic columns, and nodes and edges. Databases store structured and/or semi-structured data, depending on the type.

### Why use a database?

If your application needs to store data (and nearly every interactive application does), your application needs a database. Applications across industries and use cases are built on databases.

**X---------------------------------------X**

**DATA WAREHOUSE:**

A data warehouse stores current and historical data from one or more systems in a predefined and fixed schema, which allows business analysts and data scientists to easily analyze the data**.**

### Data warehouse characteristics:

Data warehouses typically have a pre-defined and fixed relational schema. Therefore, they work well with structured data. Some data warehouses also support semi-structured data.

Once the data is in the warehouse, business analysts can connect data warehouses with [BI tools](https://www.mongodb.com/cloud-explained/business-intelligence-bi-tools). These tools allow business analysts and data scientists to explore the data, look for insights, and generate reports for business stakeholders.

### Why use a data warehouse?

Data warehouses are a good option when you need to store large amounts of historical data and/or perform in-depth analysis of your data to generate business intelligence. Due to their highly structured nature, analyzing the data in data warehouses is relatively straightforward and can be performed by business analysts and data scientists.

**X---------------------------------------X**

**DATA LAKE:**

A data lake stores current and historical data from one or more systems in its raw form, which allows business analysts and data scientists to easily analyze the data.

### Data lake characteristics:

Data lakes store large amounts of [structured, semi-structured, and unstructured data](https://www.mongodb.com/unstructured-data). They can contain everything from relational data to JSON documents to PDFs to audio files.

Data does not need to be transformed in order to be added to the data lake, which means data can be added (or “ingested”) incredibly efficiently without upfront planning.

The flexible nature of data lakes enables business analysts and data scientists to look for unexpected patterns and insights. The raw nature of the data combined with its volume allows users to solve problems they may not have been aware of when they initially configured the data lake.

### Why use a data lake?

Data lakes are a cost-effective way to store huge amounts of data. Use a data lake when you want to gain insights into your current and historical data in its raw form without having to transform and move it. Data lakes also support machine learning and predictive analytics.