**Task 01**

**Abdul Razaque**

**Data Engineering**

* **Big Data:**

Big Data refers to a collection of data that is so huge and complex that none of the traditional data management tools are able to store it or process it efficiently.

**Fact and figure:**

* Analysts predict that by 2020, there will be 5,200 gigabytes of data in every person worldwide.
* On average, people send about 500 million tweets per day.
* Walmart processes one million customer transactions per hour.
* Amazon sells 600 items per second.
* On average, each person using email receives 88 emails daily and sends 34.
* **Data Lake:**
* A data lake is a centralized repository designed to store, process, and secure large amounts of structured, semi-structured, and unstructured data.
* A data lake is a large repository or storage system that holds vast amounts of raw, unstructured, and structured data in its native format. The data in a data lake can come from a variety of sources, such as weblogs, social media, sensors, and transactional systems.
* Unlike traditional data warehousing systems, data lakes do not require data to be transformed or processed before being stored.
* Data lakes are often built on Hadoop Distributed File System (HDFS) or cloud-based storage systems like Amazon S3, Microsoft Azure Blob Storage, or Google Cloud Storage.
* They can be accessed using a variety of tools and technologies, including SQL-based query engines, data processing frameworks like Apache Spark and Apache Flink, and machine learning libraries like TensorFlow and PyTorch.

**Enterprises rely on data lakes in key ways to help:**

* Lower the total cost of ownership
* Simplify data management
* Prepare to incorporate artificial intelligence and machine learning
* Speed up analytics
* Improve security and governance
* **Database:**
* A database is an organized collection of data stored and accessed electronically. It can contain a wide variety of information, such as customer records, sales transactions, employee information, and more.
* A database is an organized collection of structured information, or data, typically stored electronically in a computer system.
* Databases are used to store, manage, and retrieve data efficiently and effectively.

They can be categorized into several types based on their structure and function, including:

* **Relational databases**: These are the most common type of database and organize data into tables with predefined relationships between them.
* **NoSQL databases**: These databases do not use the tabular structure of relational databases and instead store data in a more flexible and unstructured manner.
* **Object-oriented databases**: These databases store data in the form of objects, which can be manipulated using object-oriented programming techniques.
* **Graph databases**: These databases use graph structures to represent and store data, allowing for more efficient querying of complex relationships.
* **Data Warehouse**:

A data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics.

A data warehouse is a large, centralized repository that stores data from various sources and is designed to support business intelligence and analytics activities.

It is a type of database specifically optimized for analysis and reporting, rather than transactional processing.

A typical data warehouse often includes the following elements:

* A relational database to store and manage data.
* An extraction, loading, and transformation (ELT) solution for preparing the data for analysis.
* Statistical analysis, reporting, and data mining capabilities.
* Client analysis tools for visualizing and presenting data to business users.
* Other, more sophisticated analytical applications that generate actionable information by applying data science and artificial intelligence (AI) algorithms, or graph and spatial features that enable more kinds of analysis of data at scale.

**Benefits of a Data Warehouse:**

Data warehouses offer the overarching and unique benefit of allowing organizations to analyze large amounts of variant data and extract significant value from it, as well as to keep a historical record.

Four unique characteristics (described by computer scientist William Inmon, who is considered the father of the data warehouse) allow data warehouses to deliver this overarching benefit. According to this definition, data warehouses are

* **Subject-oriented**. They can analyze data about a particular subject or functional area (such as sales).
* **Integrated**. Data warehouses create consistency among different data types from disparate sources.
* **Nonvolatile**. Once data is in a data warehouse, it’s stable and doesn’t change.
* **Time-variant**. Data warehouse analysis looks at change over time.