Name: Muqaddas Fatima

**Data Engineering Track**

**Task 1**

**Big Data:**

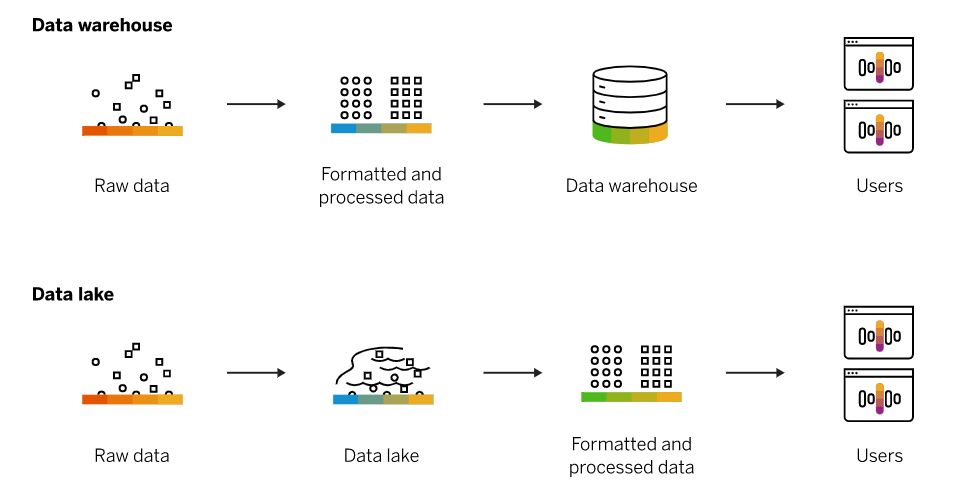
Big Data refers to extremely large sets of data that can be analyzed to reveal patterns, trends, and associations. This data is often too large and complex to be processed using traditional methods, so special tools and techniques are used to extract insights from it.

* Big Data is characterized by the "three Vs": Volume (the amount of data), Velocity (the speed at which data is generated and processed), and Variety (the diversity of data types and sources).
* To manage and process Big Data, Data Engineers use a variety of tools and technologies, such as Hadoop, Spark, and NoSQL databases. These tools allow for distributed processing and storage of data across multiple machines and systems.
* Building a data pipeline is an important part of managing Big Data. It involves collecting data from various sources, transforming it into a format that can be easily analyzed, and loading it into a centralized repository, such as a Data Warehouse or Data Lake.
* Big Data is used in a variety of applications, including business analytics, machine learning, and artificial intelligence. It allows organizations to gain valuable insights from their data and make data-driven decisions.

**Data Lake:**

A Data Lake is a centralized repository where all of an organization's raw data is stored in its original format. Unlike a Data Warehouse, which structures data for easy querying, a Data Lake stores data in its native format, allowing for more flexibility in analysis and discovery. E.g. DW: Guestroom, DL: store room

* Data Lakes are designed to store large amounts of data, including structured, semi-structured, and unstructured data, such as log files, social media posts, and sensor data. The data is often stored in a distributed file system, such as Hadoop HDFS.
* Unlike a traditional data warehouse, which uses a schema on write approach, a Data Lake uses a schema on read approach. This means that data is stored in its original format and is only structured when it is read by an application or tool.
* Data Lakes can be used for a variety of processing tasks, including batch processing, stream processing, and machine learning. Tools like Apache Spark and Apache Flink are commonly used for processing data in Data Lakes.



**Database:**

A Database is an organized collection of data stored electronically. It allows data to be easily accessed, managed, and updated. Databases can be used to store information about customers, products, transactions, and more.

* Databases are typically structured using tables, with each table representing a different entity, such as customers or products. Each table contains columns, which represent the different attributes of the entity, such as name or address, and rows, which represent individual instances of the entity.
* Databases are designed to manage data, including adding new data, modifying existing data, and deleting data. They also provide mechanisms for querying and searching data, allowing users to retrieve specific information.
* The most common type of database is a Relational Database, which uses a set of tables with defined relationships between them. These relationships allow for efficient querying and searching of data.
* There are also non-relational databases, such as document-oriented databases or key-value stores, which store data in a more flexible and scalable manner. These databases are often used for managing large amounts of unstructured or semi-structured data.

**Data Warehouse:**

A Data Warehouse is a central repository of data that is used for analysis and reporting. It's designed to support business intelligence and decision-making activities by providing a structured way to organize and analyze data from various sources.

* Data Warehouses are built using data integration techniques, which involves extracting data from various sources, transforming it into a standardized format, and loading it into the warehouse.
* Data Warehouses are structured using a dimensional model, which involves organizing data into fact tables and dimension tables. Fact tables contain the measures or metrics of interest, such as sales or revenue, while dimension tables contain descriptive information, such as product or customer information.
* Data Warehouses typically store historical data, allowing users to analyze trends and patterns over time. This historical data is often partitioned by time, such as by day, month, or year.
* Because Data Warehouses are used for business intelligence and analytics, it's important to ensure that the data is accurate and consistent. This requires strong data quality processes, including data cleansing and validation.