**Bytewise fellowship: Data Engineering**

**Task 1:**

**Big Data:**

Big data primarily refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many entries offer greater statistical power, while data with higher complexity may lead to a higher false discovery rate.

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

Example: stock exchange data, Facebook data and other social media data.

**Data Lake:**

A Data Lake is a storage repository that can store large amount of structured, semi-structured, and unstructured data. It is a place to store every type of data in its native format with no fixed limits on account size or file. It offers high data quantity to increase analytic performance and native integration.

Data Lake is like a large container which is very similar to real lake and rivers. Just like in a lake you have multiple tributaries coming in, a data lake has structured data, unstructured data, machine to the Data Lake democratizes data and is a cost-effective way to store all data of an organization for later processing. Research Analyst can focus on finding meaning patterns in data and not data itself.

**Database:**

A database is information that is set up for easy access, management and updating. Computer databases typically store aggregations of data records or files that contain information, such as sales transactions, customer data, financials and product information.

Databases are used for storing, maintaining and accessing any sort of data. They collect information on people, places or things. That information is gathered in one place so that it can be observed and analyzed. Databases can be thought of as an organized collection of information.

**Data Ware house:**

A data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data. The data

within a data warehouse is usually derived from a wide range of sources such as application log files and transaction applications.