Task # 1:

Please familiarize yourself with the basics of Data Engineering which include the following topics:

- Big Data

- Data Lake

- Database

- Data Warehouse

Big Data:

**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.

3 types of Big Data

* Structured data.
* Unstructured data.
* Semi-structured data.

7 Pillars of Big Data

Volume

Value

Variety

Velocity

Veracity

virtual,

variability

Data Lake:

A data lake stores structured, semi-structured and unstructured data, supporting the ability to store raw data from all sources without the need to process or transform it at that time.

Only when the data needs to be retrieved, will some structure need to be applied, which is ideal in the hands of data scientists and data analysis developers who can create new data models on the fly but does not provide the same reporting capabilities and ease of use for business users. Storing data in data lakes is much cheaper than in a data warehouse. Data lakes are very popular in the modern stack because of its flexibility and costs but they are not a replacement for data warehouses or relational databases.

Database:

A database is used to store, search and report on structured data from a single source. They are the simplest to create and SQL can be used to query and report on the data. There are both open source and proprietary databases, making it widely accessible to install and start using on premise or on the cloud.

A **relational database** require schemas and are not a fit for unstructured or semi-structured data. Because of this rigid schema, they are not suited to be the centralized place to store data from multiple sources where the raw data varies in format and structure. However, they are popular for data analysis and monolithic applications

Data Warehouse:

A data warehouse is used to store large amounts of structured data from multiple sources in a centralized place. Organizations invest in building data warehouses because of its ability to deliver business insights from across the company, and quickly.

The process of creating a data warehouse requires some heavy lifting in the planning and design stage of examining data structures. Data warehouses are preferred by the business and operations decision makers of the company and a good system justifies its often high costs in proprietary software and storage.