Task # 1:

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

- **Big Data**
  Big Data is defined as a combination of structured, unstructured, and semi-structured data collected by organizations. Big data is often described by five characteristics i.e., volume, velocity, variety, veracity, and value. Data is growing on a large scale and eventually faster than we thought so we need tools that store them in their appropriate manner so that we extract the best out of it. Big data by the name itself indicates a problem that has been solved by different provided tools like Hadoop, and Cassandra but now a modern stack of handling the big data is much more reliable for example AWS, Azure, GCP, etc.
- **Data Lake**
  Data Lake is the centralized repository where we can dump raw data i.e., it will be structured, or unstructured. There is one storage level where we store all the data. We see the data lake concept more in AWS. Data lake traditionally separates the storage layer from the compute layer. S3 makes up the storage layer and compute layer can be made up of several different services i.e., glue, and lambda.
- **Database**
  The database is the traditional storage of data where we store a bunch of records. When we talk about Databases, they are relational and follow OLTP system architecture. There are different anomalies within the database we need to configure. There can be unstructured data that might cause the problem in the analysis. Different tools like MySQL Workbench, PostgreSQL, MSSQL, etc. offer database management systems.
- **Data Warehouse**
  Data Warehouse is the process of getting into structured data from unstructured data. A data warehouse is traditionally the OLAP system because it maintains both transactional and analytical records for analysis. There are four stages where we need to pass on to get it into the warehouse i.e., data sources, data staging, data storage, and data marts. Typical tools including Teradata, Snowflake, Azure, etc. are offering services to hold the data.