Task # 2:

Familiarize yourself with the following topics:

- **Data Marts**
  Data Marts is the subset of Datawarehouse. It talks about the specific subject area. It contains the processed data that can be analyzed in the later stage. Teams within the organization create a data mart from the data warehouse and then discontinue/stop when it fulfills the need. It also saves time because we don't need to search the whole warehouse instead, we target one subject line for analysis. Dart marts are just the components of a data warehouse, collectively they make the warehouse.

- **Data Lakehouse**
  Data Lakehouse is top of a line management system for the data lake. It provides flexibility and cost-effectiveness. When we dump raw data onto the data lake, we cannot process it because of its large size, and unstructured behavior but with data warehouse capabilities we are adding features that are lacking in data lakes themselves.

- **Data Mesh**
  Data Mesh is a decentralized data platform that allows end users to access important data without transporting it to a data warehouse or data lake. Organizations can set up data meshes for either reporting needs or data sharing across the enterprise. Data Mesh provides ownership to the producer to gain insight from it. The idea is to make data available by connecting different data owners, data consumers, or producers.

- **DWH vs Data Lake**
  - Data Warehouse is the collection of several different databases whereas a Data Lake is the one centralized repository for dumping large amounts of data.
  - A data warehouse contains structured data whereas Data Lake is containing raw and unstructured data.
  - Both support OLAP architecture.
  - We perform the ETL process before dumping it in the warehouse and in Data Lake we use a Glue crawler to define the schema of the data.

- **OLTP vs OLAP**
  - OLTP systems record data in real-time whereas OLAP systems keep track of historical data.
  - OTLP systems handle a small amount of data whereas OLAP systems handle a large amount of data.
  - Databases have the characteristics of OLTP whereas Data Warehouse/Data Lake followed OLAP.

Task # 3:

After you complete these topics, please answer the following questions in your document:

Can a database be used as DWH?

No, by the concept they serve different purposes. We need a database for our day-to-day records and if that record holds on for a longer period, we dump it in the warehouse in a structured format so that it can be analyzed. Also, a Data warehouse is a collection of databases that target different subject areas of any organization so that we can keep track of the things happening and happening. So, a database cannot fulfill the need of a warehouse instead we use it for transactional data.

Major differences between structured and Unstructured data?

Structured data eliminate the anomalies of querying the data whereas unstructured data contains raw data in various forms which is difficult to search. Unstructured data requires processing to understand. DWH/Relational Databases are an example of structured data and Data Lakes are non-relational databases.

What are the duties of a data engineer? (high-level)

Data engineers work in a variety of settings to build systems that collect, manage, and convert raw data into usable information. Data engineers design, build and optimize systems for data collection, storage, access, and analytics at scale. Their goal is to make data accessible so that organizations can use it to evaluate and optimize their performance. They create data pipelines used by data scientists, data-centric applications, and other data consumers. Today modern Data Engineers use cloud services to perform the task on resources.