Question 1: Describe a Data Engineer role in an organisation and its main responsibilities.

Answer:

A data engineer is an IT worker whose primary job is to prepare data for analytical or operational uses. These software engineers are typically responsible for building data pipelines to bring together information from different source systems. They integrate, consolidate and cleanse data and structure it for use in analytics applications. They aim to make data easily accessible and to optimize their organization's big data ecosystem.

The amount of data an engineer works with varies with the organization, particularly with respect to its size. The bigger the company, the more complex the analytics architecture, and the more data the engineer will be responsible for. Certain industries are more data-intensive, including healthcare, retail, and financial services.

Data engineers work in conjunction with data science teams, improving data transparency and enabling businesses to make more trustworthy business decisions.

The data engineer role

Data engineers focus on collecting and preparing data for use by data scientists and analysts. They take on three main roles as follows:

- **Generalists.** Data engineers with a general focus typically work on small teams, doing end-to-end data collection, intake, and processing. They may have more skill than most data engineers, but less knowledge of systems architecture. A data scientist looking to become a data engineer would fit well into the generalist role.

A project a generalist data engineer might undertake for a small, metro-area food delivery service would be to create a dashboard that displays the number of deliveries made each day for the past month and forecasts the delivery volume for the following month.

- **Pipeline-centric engineers**. These data engineers typically work on a midsize data analytics team and more complicated data science projects across distributed systems. Midsize and large companies are more likely to need this role.

A regional food delivery company might undertake a pipeline-centric project to create a tool for data scientists and analysts to search metadata for information about deliveries. They might look at distance driven, and drive time required for deliveries in the past month, then use that data in a predictive algorithm to see what it means for the company's future business.

- **Database-centric engineers**. These data engineers are tasked with implementing, maintaining and populating analytics databases. This role typically exists at larger companies where data is distributed across several databases. The engineers work with pipelines, tune databases for efficient analysis and create table schemas using extract, transform, load (ETL) methods. ETL is a process in which data is copied from several sources into a single destination system.

A database-centric project at a large, multistate, or national food delivery service would be to design an analytics database. In addition to creating the database, the data engineer would write the code to get data from where it's collected in the main application database into the analytics database.

Data engineer responsibilities

Data engineers often work as part of an analytics team alongside data scientists. The engineers provide data in usable formats to the data scientists who run queries and algorithms against the information for predictive analytics, machine learning and data mining applications. Data engineers also deliver aggregated data to business executives and analysts and other end users so they can analyse it and apply the results to improving business operations.

Data engineers deal with both structured and unstructured data. Structured data is information that can be organized into a formatted repository like a database. Unstructured data -- such as text, images, audio, and video files -- doesn't conform to conventional data models. Data engineers must understand different approaches to data architecture and applications to handle both data types. A variety of big data technologies, such as open-source data ingestion and processing frameworks, are also part of the data engineer's toolkit.