Question 1

A data engineer serves as an architect and builder for a company's data infrastructure. Their primary responsibility lies in creating systems and pathways that facilitate the smooth flow of information, ensuring it's organized and easily accessible for analytical or operational purposes. They are deeply involved in developing, building, and managing data pipelines, with a critical task of ensuring the quality and accuracy of output data. Typically, data engineers work extensively with large datasets and databases, necessitating a combination of technical expertise and a profound understanding of business needs.

The data engineering process begins with the collection of data from sources such as Secure File Transfer Protocol (SFTP) or APIs. The SFTP process involves establishing a secure connection using Secure Shell between the client and the server. Once the secure connection is confirmed, data transfer commences. Data engineers utilize SFTP to pull data from a server by writing a program or script that runs on their server, establishing a secure connection with the server they intend to retrieve data from. This process involves extracting files from one directory and placing them in their designated destination directory.

Concurrently, data engineers are expected to write code, scripts, or programs that make API calls exposed by the application team, commonly using languages like Python and libraries such as Requests. For data engineers who manage applications, they collect data from their personal databases and store it as a file on their server. After extracting the data, data engineers must prepare it to meet the standards of data storage, involving tasks like changing data types and adjusting timestamps.

Subsequently, data engineers employ various libraries, such as Pandas and Spark in Python, to transform the data into useful information, applying business logic to solve given problems. Finally, they create and frequently update table reports, vital for business intelligence engineers who use tools like Power BI and Tableau to craft informative dashboards. Once the data is clean, data scientists run their models on this refined data, aiding organizational decision-making.