Storage

The resulting tables have to be stored in an appropriate format, for example storing the data in a relational database or CSV. You must describe why you chose the storage format you have chosen (How/why is this format helpful, what are the advantages/disadvantages)

Ans:

If using a relational database, I would create tables for each data frame and define schema and relationships between the tables. This would make it easier to query the data and perform complex joins and aggregations. If using CSV, I would store each data frame in a separate CSV file, which would make it easier to share the data with others.

Optimization

Imagine that the volume of the data has become very large and that we're working with multiple data sources and tables, the data pipeline is starting to take longer and longer to complete and we need to optimize it. Without implementing any code, explain what approaches you would take to optimize the performance of a data pipeline.

Ans:

1-Parallelization: I would parallelize the data processing pipeline by breaking the data into smaller chunks and processing them in parallel

2-Compression: I would compress the data to reduce the amount of disk I/O and network bandwidth required to transfer the data.

3-Stream processing: If the data is continuously arriving, I would consider using stream processing frameworks such as Kafka or Spark Streaming to process the data in real-time.