Conceptual Questions

1️⃣ **Why is data cleaning important in real-time data processing?**

* Real-time data often contains missing, inconsistent, or incorrectly formatted values.
* Dirty data can cause incorrect analytics, model errors, or system failures.
* Cleaning ensures accuracy, consistency, and reliability, which is critical for timely decision-making in real-time applications.
* Example: Converting all date formats to YYYY-MM-DD ensures uniformity for time-based analysis.

2️⃣ **What are pipeline artifacts and how are they used in DevOps workflows?**

* Pipeline artifacts are files or outputs produced by a CI/CD pipeline, such as builds, logs, or processed data.
* They are stored temporarily or permanently in the pipeline for later stages or for download.
* Use cases in DevOps:

o Share processed datasets between stages of a pipeline o Pass build outputs to deployment pipelines o Keep logs or reports for auditing

3️⃣ **How would you modify the pipeline to store the cleaned data into Azure Blob Storage?**

* Instead of just publishing artifacts in Azure DevOps, you can upload files to Azure Blob Storage using Azure CLI or Python SDK.
* Example approaches:
  1. **Azure CLI task in YAML:** - task: AzureCLI@2 inputs:

azureSubscription: '<your-service-connection>' scriptType: 'bash'

scriptLocation: 'inlineScript' inlineScript: | az storage blob upload \

--account-name <storage\_account> \

--container-name <container\_name> \

--name clean\_sales\_data.csv \

--file data/clean\_sales\_data.csv \

--overwrite

* 1. **Python approach inside data\_processing.py using azure-storage-blob:**

from azure.storage.blob import BlobServiceClient

conn\_str = "<your-connection-string>" container\_name = "<container-name>"

blob\_service\_client = BlobServiceClient.from\_connection\_string(conn\_str) blob\_client = blob\_service\_client.get\_blob\_client(container=container\_name, blob="clean\_sales\_data.csv")

with open("data/clean\_sales\_data.csv", "rb") as f:

blob\_client.upload\_blob(f, overwrite=True)

* This ensures that cleaned data is stored in Azure cloud and can be accessed outside the pipeline.