Part 1

- You have a Python script that processes millions of records in a single thread. How would you optimize it to leverage multiple cores and reduce the execution time? Provide a sample code snippet.
- During a data pipeline run in Azure Data Factory, a step failed due to an invalid data format. Describe how you would debug this issue and prevent it from happening in the future.
- Write a Python script using the Azure SDK that uploads a file to an Azure Blob Storage container. Ensure the script checks if the container exists and creates it if it does not.
- Write a Python script to download logs from Azure (e.g. events from a specific resource)

Part 2 (hands on tasks using our Azure deployment)

Deploy a Virtual Machine, Ingest Data, Analyze, and Export Results

Information for Candidates:

- **Azure Subscription:** You have access to our Azure subscription named "**infrastructure-dl-dwh**".
- **Resource Group:** Your resources will be created in the resource group named **"Data_Engineer".**
- **Storage Account:** There is a storage account named "**dataengineerv1**" which has restricted access, allowing connections only from specific virtual networks and IP addresses.
- **Data File:** The CSV file named "**tourism_dataset.csv**" is stored in a container called "**raw**" within this storage account.
- **Access Setup:** To access the storage account, ensure your IP address is added to the access list. If you encounter an access error when trying to access the "raw" container, please take a screenshot of the error message and contact @Fatemeh to add your IP address.

Details

Step 1: Deploy a Virtual Machine (VM)

- **Objective:** Deploy a VM named <VM-YourName> using Python (e.g., VM-Kostas).
- **Requirements:**
    - **Create a Virtual Network (VNet):** Set up a VNet within the "Data_Engineer" resource group with your full name.
    - **Create a Subnet:** Configure a subnet within the VNet.
    - **Set Up a Network Interface Card (NIC):** Associate the NIC with the VNet and subnet you created.
    - **Deploy the VM:** Use the NIC to deploy a VM within the VNet. Ensure the VM's name follows the naming convention <VM-YourName> (e.g., VM-Kostas).

Step 2: Read Data from Azure Storage Account

- **Objective:** Read the CSV file from the Azure Storage Account using Python.
- **Requirements:**
    - **Connect to the Azure Storage Account:** Use “DefaultAzureCredentials” to connect to the storage account "dataengineerv1".
    - **Load the Data:** Use Python (preferably with the Azure SDK and Pandas library) to load the "tourism_dataset.csv" file from the "raw" container into a Pandas DataFrame.

Step 3: Perform Data Analysis

- **Objective:** Analyze the data to extract insights.
- **Requirements:**
    - **Group and Aggregate Data:** Group the data by the 'country' column and calculate the average value of the "Rate" column for each country. Please include the equivalent SQL query as a comment.
    - **Identify Top Categories:** Find the top 3 categories with the highest average rate across all countries. Please include the equivalent SQL query as a comment.

Step 4: Export Results and Save to VM

- **Objective:** Save your analysis results back to Azure Storage and your VM.
- **Requirements:**
    - **Write Results to CSV:** Save the aggregated results to a CSV file named <YourFirstName-YourLastName>.csv (e.g., Kostas-Tsirigos.csv).
    - **Upload to Azure Storage:** Create a new directory in the storage account "dataengineerv1" named <YourFirstName-YourLastName>, and upload the CSV file to this directory.
    - **Configure Networking:** Add your VM's VNet to the "Networking" settings of the storage account to enable access.
    - **Download to VM:** SSH into your VM and download the resulting CSV file from the Azure Storage Account to the home directory of your VM using Azure CLI, saving it as <result-YourName> (e.g., result-Kostas). Use the provided account key “ieLmjePYNxBcajmfHvX8TsMXa3bn8nkH3MCuaWTsA/E+G56z3KRYSPO1M5MaHNds5FhE37PsZwYm+AStsnl/lg==” for access.

Deliverables

1. **Python Scripts and Linux Scripts:** Store all your code in your GitHub repository.
2. **Screenshot of Access Error:** If you encounter an access error when trying to connect to the "raw" container, include a screenshot.
3. **Resulting CSV File:** Ensure the CSV file is saved both in Azure Storage (in your named directory) and on your VM (home directory).
4. **VM and Networking Setup:** Confirm that your VM and VNet are correctly configured to access the storage account.