Introduction to Big Data
Graded Assignment 1

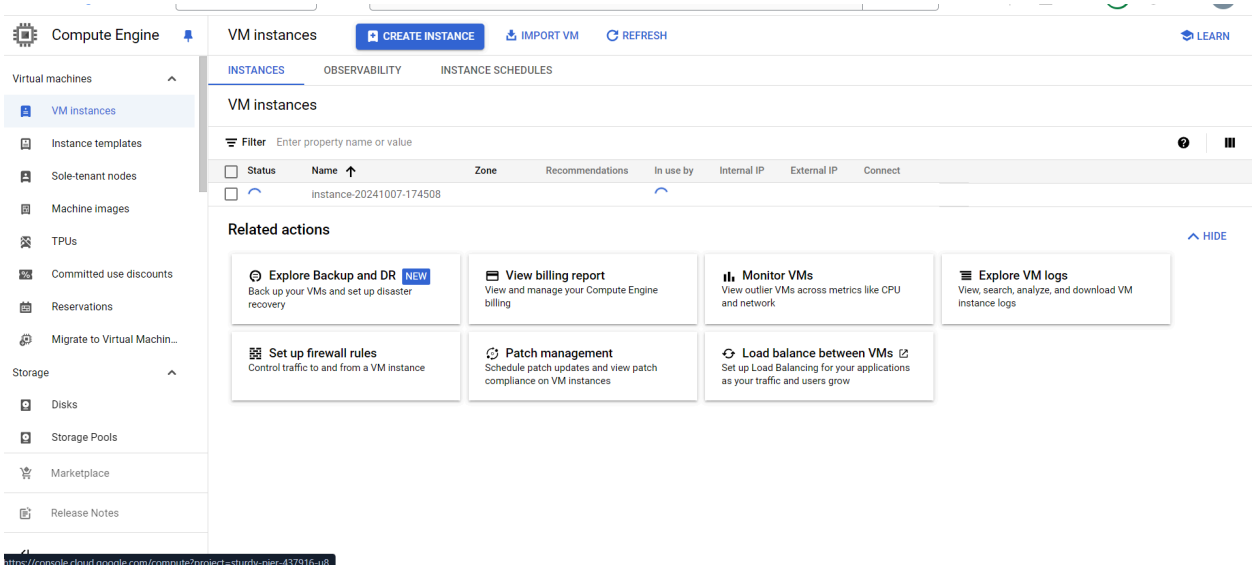
Name - Avijeet Palit
Roll - 21f1005675
Date - 07/10/2024

This report outlines the workflow for creating a Virtual Machine (VM) in Google Cloud Platform (GCP) and setting up a Cloud Storage bucket for uploading a text file. Additionally, it details the steps to upload a Python script that counts the number of lines in the text file.

To accomplish the tasks in the graded assignments, follow these steps:

Step 1: Create a VM Instance on GCP

1. **Access Google Cloud Console**: Go to the Google Cloud Console.
2. **Navigate to VM Instances**:
    - Click on **Compute Engine** in the sidebar.
    - Select **VM instances**.
3. **Create a New Instance**:
    - Click on **Create Instance**.
    - **Configure the Instance**:
        - Choose the desired machine type.
        - Select the region and zone.
        - Configure any additional settings as needed.
    - Click **Create** to launch the VM.



Step 2: Connect to the VM

1. **Open SSH Terminal**:
    - Once the VM is running, locate your instance in the VM instances list.
    - Click on the **SSH** button next to your instance.
    - A new window will open with a terminal connected to your VM.

Step 3: Set Up the VM Environment

Install Python and the necessary libraries to interact with Google Cloud Storage (GCS).

Update Packages and Install Python:

```
sudo apt-get update
sudo apt-get install -y python3-pip
```

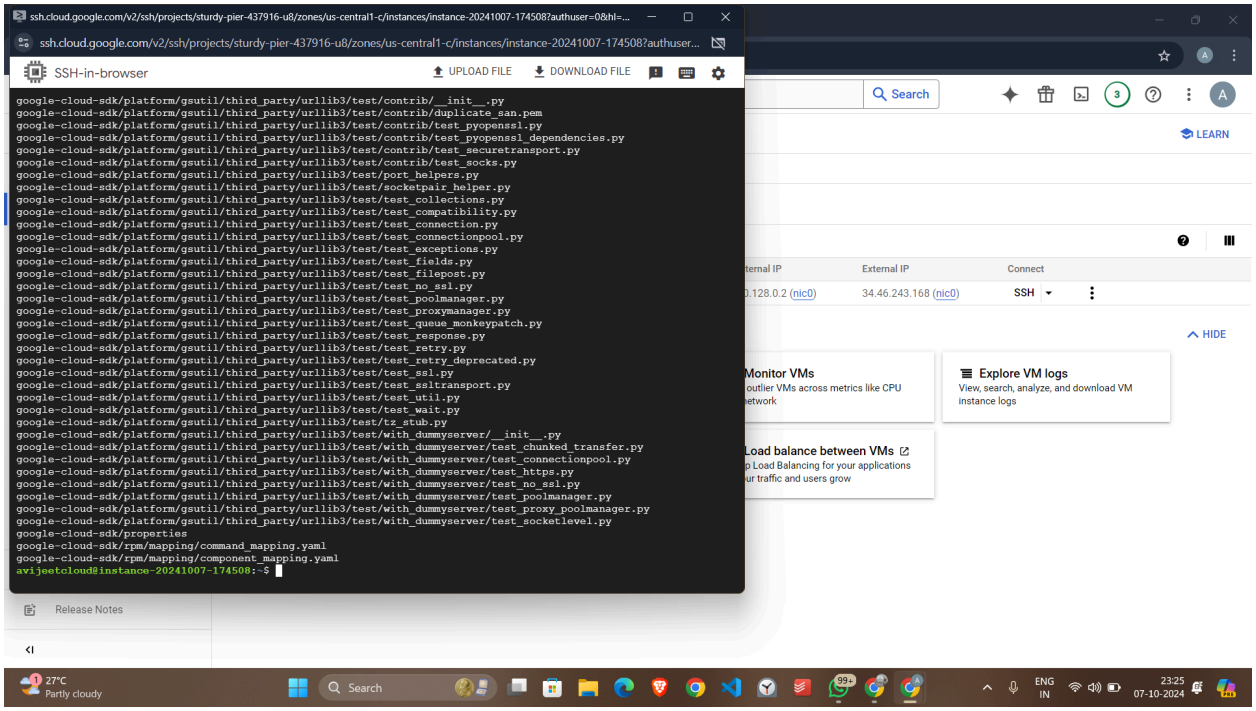
Install Google Cloud Storage Library:

```
pip3 install google-cloud-storage
```

Set Up Google Cloud SDK and Authenticate:

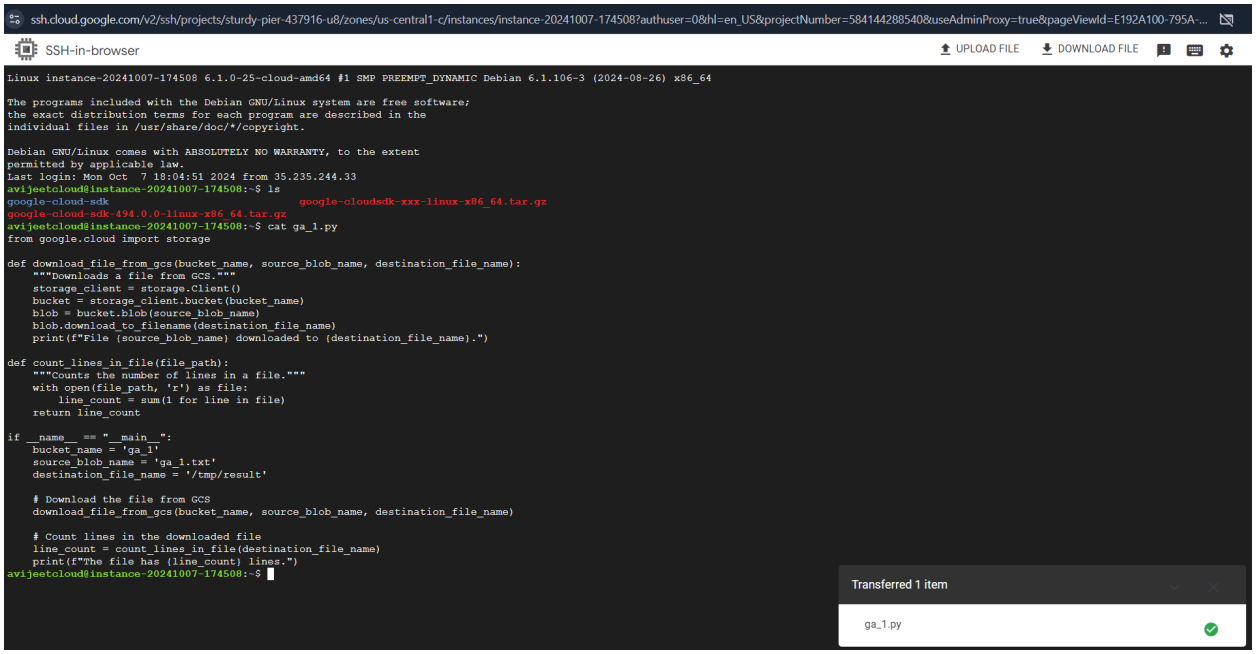
```
curl -O
https://dl.google.com/dl/cloudsdk/channels/rapid/downloads/google-cloud-sdk-XXX-linux-x86_64.tar.gz
tar -xf google-cloud-sdk-XXX-linux-x86_64.tar.gz
```

```
./google-cloud-sdk/install.sh
./google-cloud-sdk/bin/gcloud init
./google-cloud-sdk/bin/gcloud auth application-default login
```

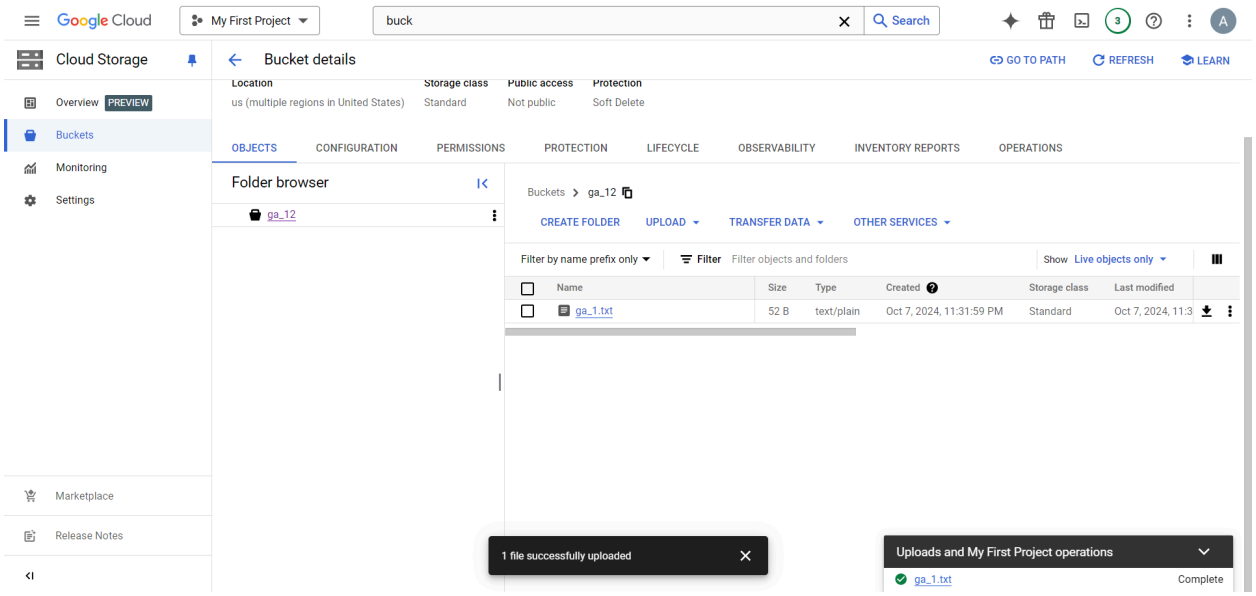


Step 4: Write the Python Script and Create the Bucket

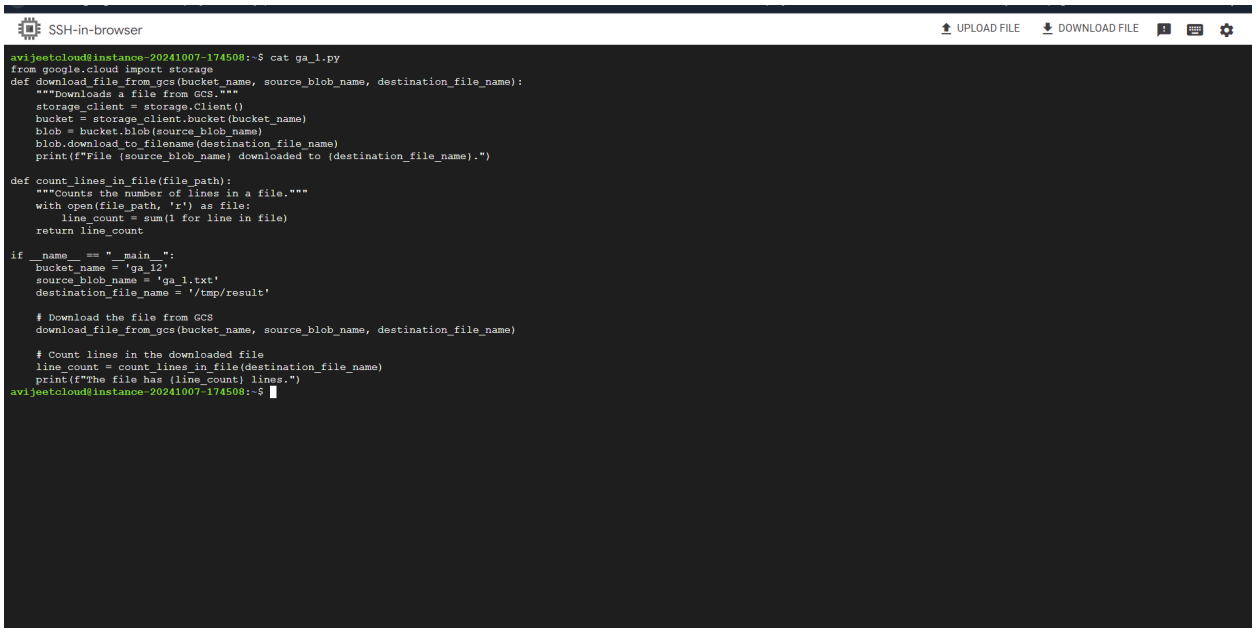
1. **Create the Python Script**:
    - On your VM, Upload a new file named `ga_1.py`.



2. **Create a Cloud Storage Bucket**:
    - In the Google Cloud Console, navigate to **Cloud Storage** > **Buckets**.
    - Click on **Create Bucket**.
    - **Name the Bucket**: Enter `ga_21` as the bucket name.
    - Configure any additional settings as needed.
    - Click **Create** to finalize the bucket creation.
3. **Upload the Text File**:
    - Within the bucket `ga_12`, click on **Upload Files**.
    - Select the `ga_1.txt` file from your local machine.
    - Upload the file, which contains the content to be read by your Python script.



Step 5: Run the Script



Execute the Python Script:

`python3 ga_1.py`

Script Functionality:

- The script will download the specified file from the GCS bucket to the VM's local filesystem (e.g., `/tmp/results`).

- It will then count the number of lines in the file and output the result.

Result: