Virtual Machine Setup with Linux as Operating System

Use of a virtual machine on a cloud platform:

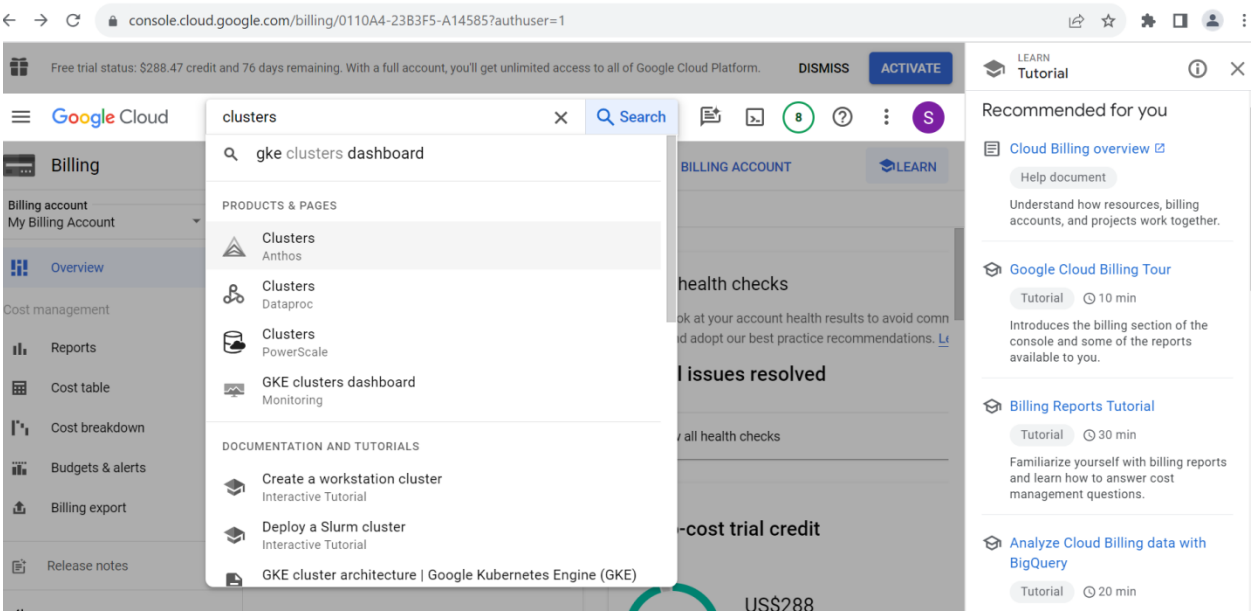
A virtual machine will be used to run applications and services on the cloud platform. It is flexible, scalable and cost effective.

Pre-requisites:

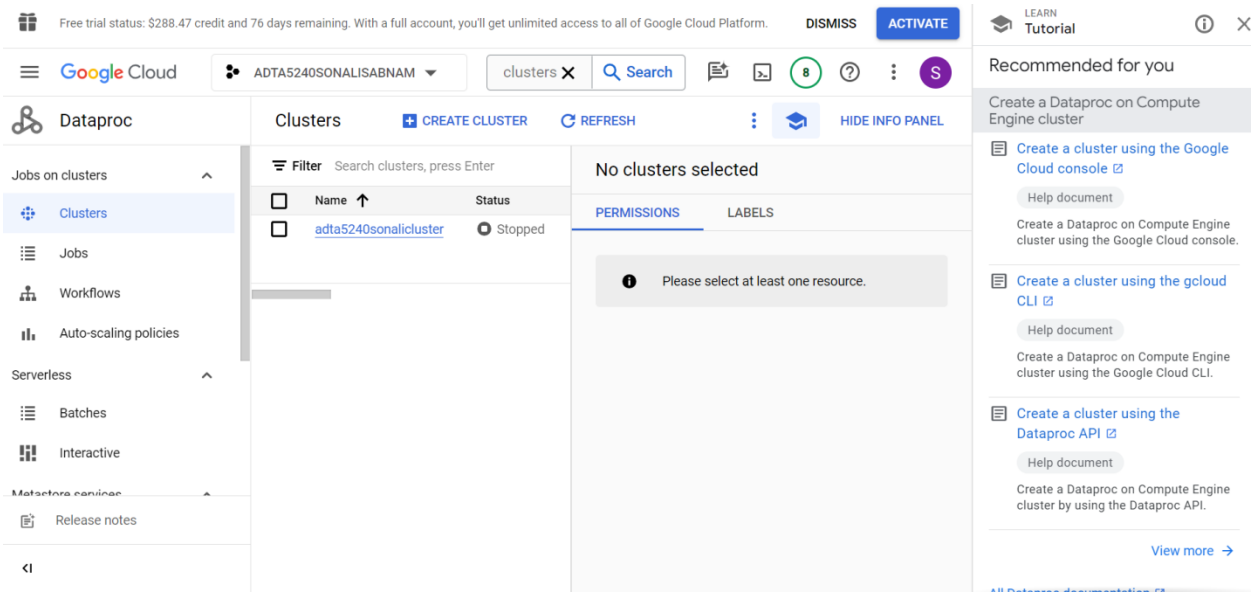
- An active Google/Gmail account
- GCP account should have been created with the same Google/Gmail account.
- A project associated with a storage bucket, folders with data files and Hadoop spark cluster.

Step 1: Login to GCP console and start the cluster

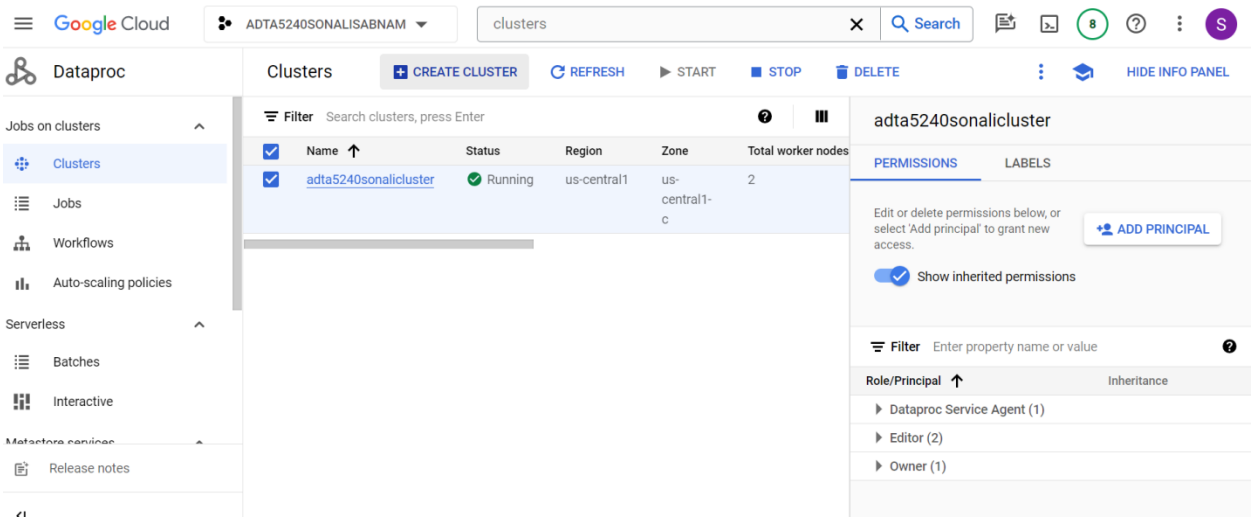
- Go to https://console.cloud.google.com and sign in using your Google/Gmail account. If you are using crome browser and have logged in to the browser using the Google/Gmail account, it will directly take you to the GCP home page (no login required).
- From the Navigation menu go to “Cluster” or type “Cluster” in the search bar and click on “Clusters Dataproc”.



- The below screen gets displayed and it shows my previously created cluster which is not running.

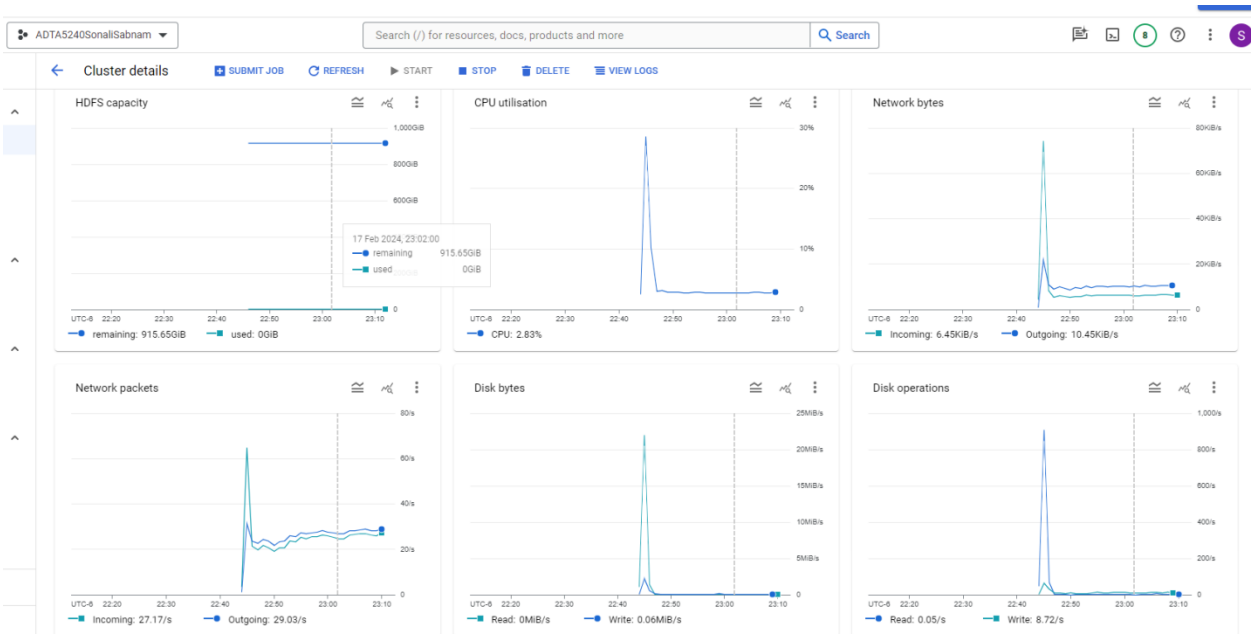


- Select the cluster and click on “RESTART/RESUME” to start the cluster. The below screenshot shows the cluster is now running.

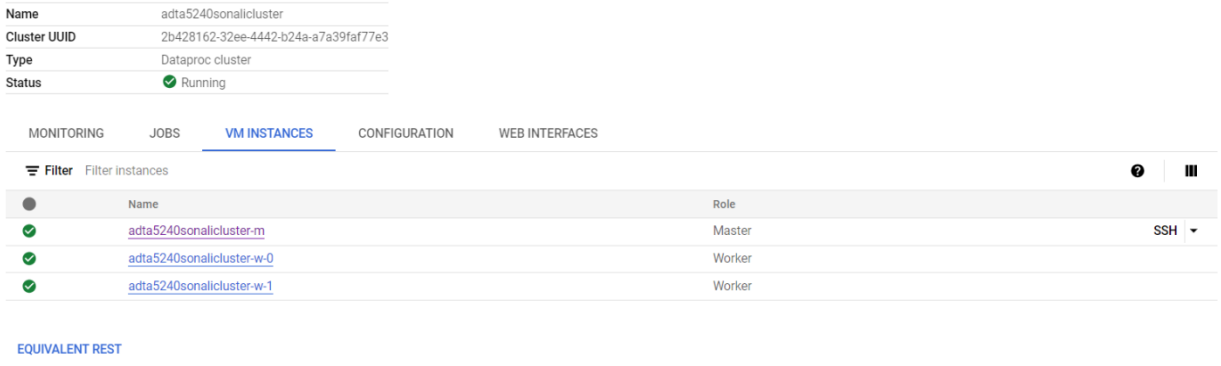


Step 2: Monitoring the cluster and SSH

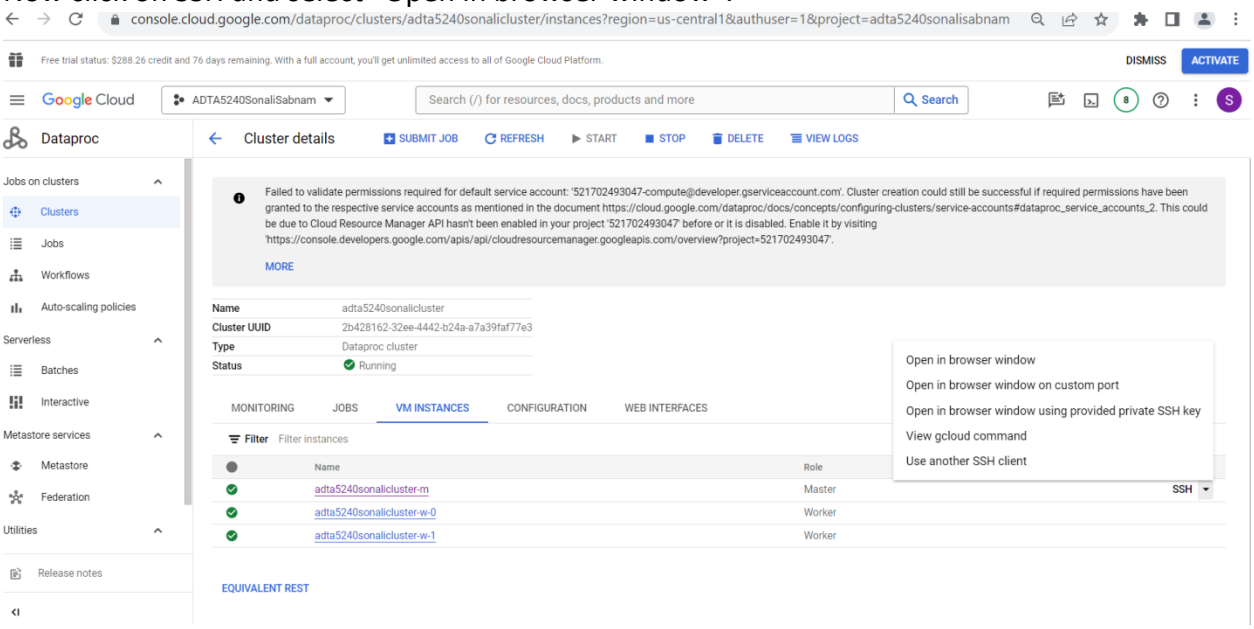
- Click on the cluster to see the usage data. In the “Monitoring” we can see graphs showing “YARN Memory”, “YARN Pending Memory” and “YARN Node Managers” and some other graphs. These graphs show the usage data.



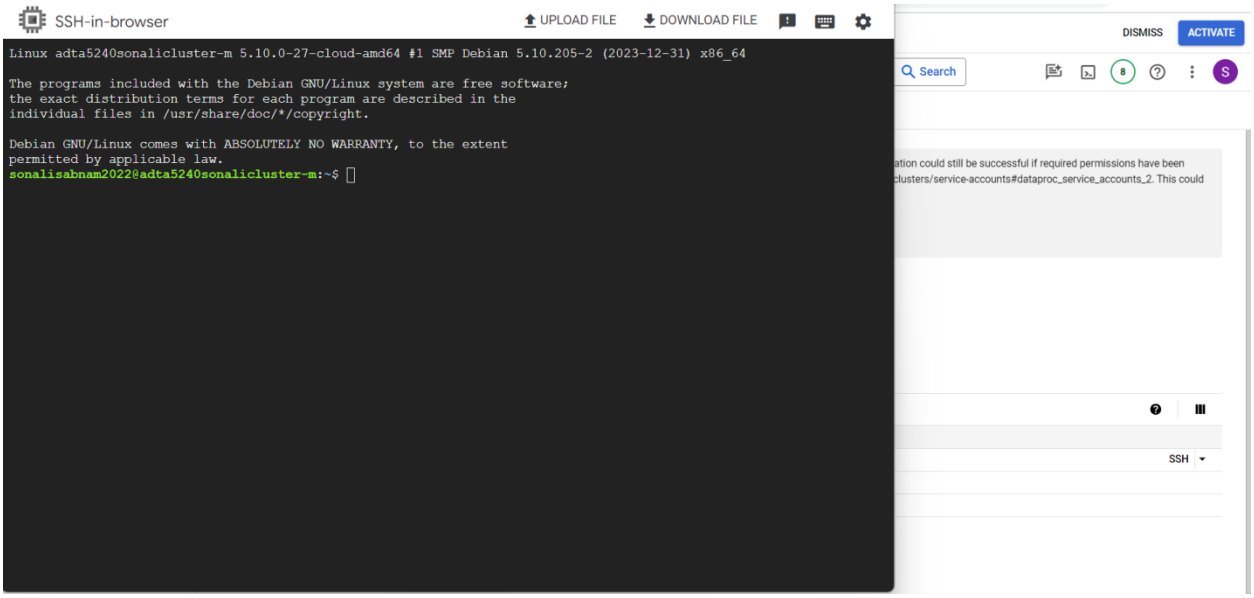
- Click on “VM Instances”. The below screenshot shows.



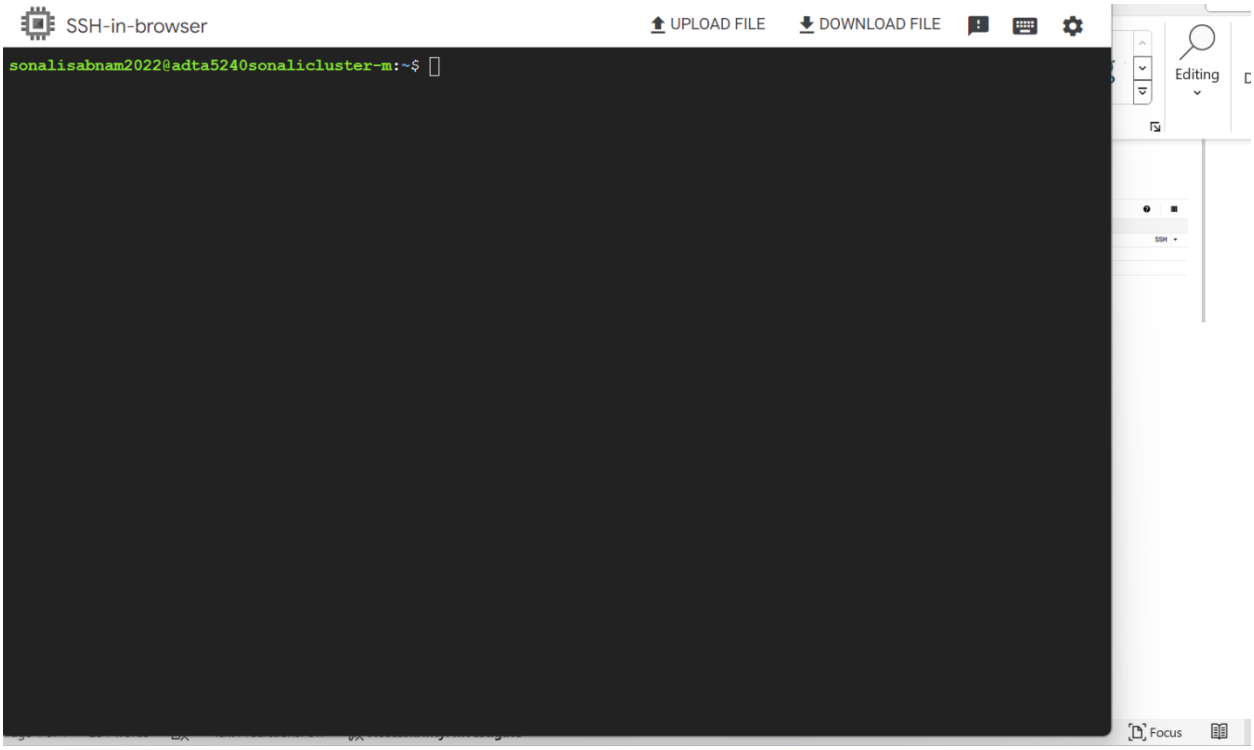
- Now click on SSH and select “Open in browser window”.



- The terminal opens in a browser window. A small popup will ask you to “Authorize” and the below screen gets displayed. This terminal will be used to connect to the VM/server and send commands.



- Type “clear” to clear the terminal.



- We will be accessing HDFS(Hadoop Distributed File System) so we will use a few Hadoop commands too(starting with hdfs).
- Type the below commands:
    - “whoami”
    - “pwd”

- “hdfs dfs -ls /”

- The below screenshot shows the commands.

```
sonalisabnam2022@adta5240sonalicluster-m:~$ whoami
sonalisabnam2022
sonalisabnam2022@adta5240sonalicluster-m:~$ pwd
/home/sonalisabnam2022
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -ls /
Found 3 items
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /tmp
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /user
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /var
sonalisabnam2022@adta5240sonalicluster-m:~$
```

- We can see the three Hadoop folders in the above screenshot. We will be working with the “/user” folder.
- Run the below command to go inside the “user” folder.
  - hdfs dfs -ls /user

```
sonalisabnam2022@adta5240sonalicluster-m:~$ whoami
sonalisabnam2022
sonalisabnam2022@adta5240sonalicluster-m:~$ pwd
/home/sonalisabnam2022
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -ls /
Found 3 items
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /tmp
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /user
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /var
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -ls /user
Found 11 items
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /user/dataproc
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /user/hbase
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /user/hdfs
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /user/hive
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /user/mapred
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /user/pig
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /user/solr
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /user/spark
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /user/yarn
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /user/zeppelin
drwxrwxrwt   - hdfs hadoop          0 2024-02-11 21:06 /user/zookeeper
sonalisabnam2022@adta5240sonalicluster-m:~$
```

- Here we can see some folders created by the system.

Step 3: Create a folder/directory in HDFS

- Write the below commands:
  - hdfs dfs -mkdir /user/sonalisabnam2022
  - hdfs dfs -ls /user

```
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -mkdir /user/sonalisabnam2022
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -ls /user
Found 12 items
drwxrwxrwt   - hdfs             hadoop          0 2024-02-11 21:06 /user/dataproc
drwxrwxrwt   - hdfs             hadoop          0 2024-02-11 21:06 /user/hbase
drwxrwxrwt   - hdfs             hadoop          0 2024-02-11 21:06 /user/hdfs
drwxrwxrwt   - hdfs             hadoop          0 2024-02-11 21:06 /user/hive
drwxrwxrwt   - hdfs             hadoop          0 2024-02-11 21:06 /user/mapred
drwxrwxrwt   - hdfs             hadoop          0 2024-02-11 21:06 /user/pig
drwxrwxrwt   - hdfs             hadoop          0 2024-02-11 21:06 /user/solr
drwxr-xr-x   - sonalisabnam2022 hadoop          0 2024-02-18 05:32 /user/sonalisabnam2022
drwxrwxrwt   - hdfs             hadoop          0 2024-02-11 21:06 /user/spark
drwxrwxrwt   - hdfs             hadoop          0 2024-02-11 21:06 /user/yarn
drwxrwxrwt   - hdfs             hadoop          0 2024-02-11 21:06 /user/zeppelin
drwxrwxrwt   - hdfs             hadoop          0 2024-02-11 21:06 /user/zookeeper
sonalisabnam2022@adta5240sonalicluster-m:~$
```

- The above screenshot shows the folders newly created.
- The subfolder "/user/sonalisabnam2022" shows read/write/execute permissions.
- Use the below command to access the contents of the subfolder.
    - hdfs dfs -ls /user/ sonalisabnam2022

```
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -ls /user/sonalisabnam2022
sonalisabnam2022@adta5240sonalicluster-m:~$
```

- As there are no files/folders here. It shows nothing.

Step 4: Create a subfolder to place data files and logs

- Type the below commands to create a subfolder in "/user/sonalisabnam2022".
    - hdfs dfs -mkdir /user/sonalisabnam2022/data
    - hdfs dfs -ls /user/sonalisabnam2022/data

```
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -ls /user/sonalisabnam2022
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -mkdir /user/sonalisabnam2022/data
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -ls /user/sonalisabnam2022/data
sonalisabnam2022@adta5240sonalicluster-m:~$
```

Step 5: Copy data files from storage bucket to "user/sonalisabnam2022/data"

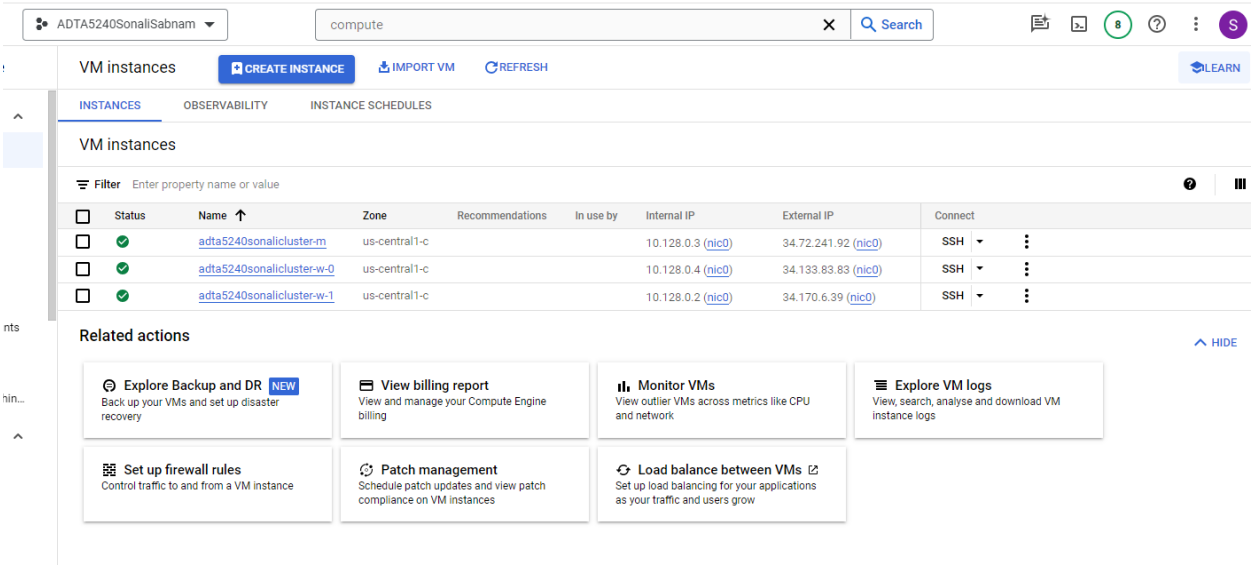
- Type the below commands to create a subfolder "userdata" and list it once created.
    - hdfs dfs -mkdir /user/sonalisabnam2022/data/userdata
    - hdfs dfs -ls /user/sonalisabnam2022/data

```
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -mkdir /user/sonalisabnam2022/data/userdata
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -ls /user/sonalisabnam2022/data
Found 1 items
drwxr-xr-x   - sonalisabnam2022 hadoop          0 2024-02-18 05:54 /user/sonalisabnam2022/data/userdata
sonalisabnam2022@adta5240sonalicluster-m:~$
```

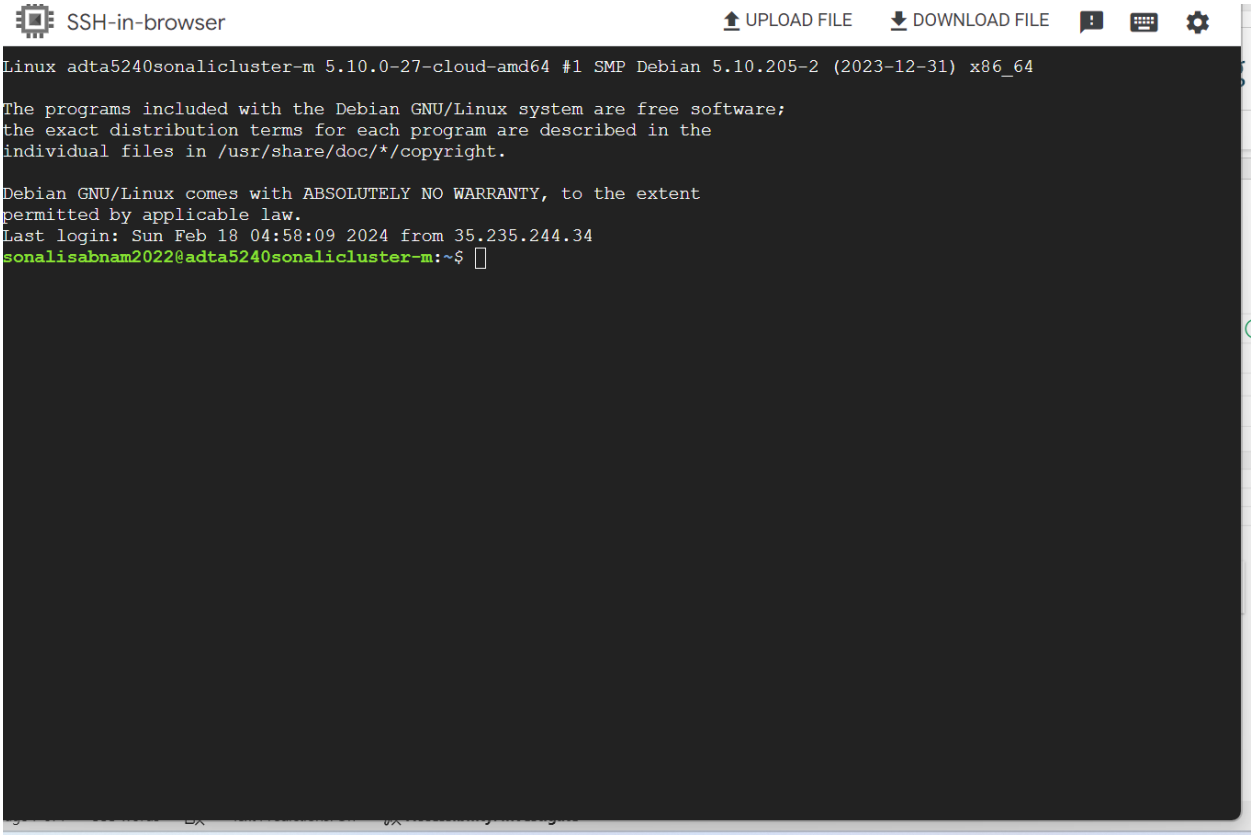
- The "userdata" folder is now created and is displayed.
- Type the below command to create "weblog" subfolder and list it once created.
    - hdfs dfs -mkdir /user/sonalisabnam2022/data/weblog
    - hdfs dfs -ls /user/sonalisabnam2022/data

```
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -ls /user/sonalisabnam2022/data
Found 1 items
drwxr-xr-x   - sonalisabnam2022 hadoop          0 2024-02-18 05:54 /user/sonalisabnam2022/data/userdata
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -mkdir /user/sonalisabnam2022/data/weblog
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -ls /user/sonalisabnam2022/data
Found 2 items
drwxr-xr-x   - sonalisabnam2022 hadoop          0 2024-02-18 05:54 /user/sonalisabnam2022/data/userdata
drwxr-xr-x   - sonalisabnam2022 hadoop          0 2024-02-18 05:59 /user/sonalisabnam2022/data/weblog
sonalisabnam2022@adta5240sonalicluster-m:~$
```

- We can see the "userdata" and "weblog" subfolders are now created and displayed.
- Go to GCP console and type "compute engine" in the search bar and select the "Compute Engine". The below screen is displayed.

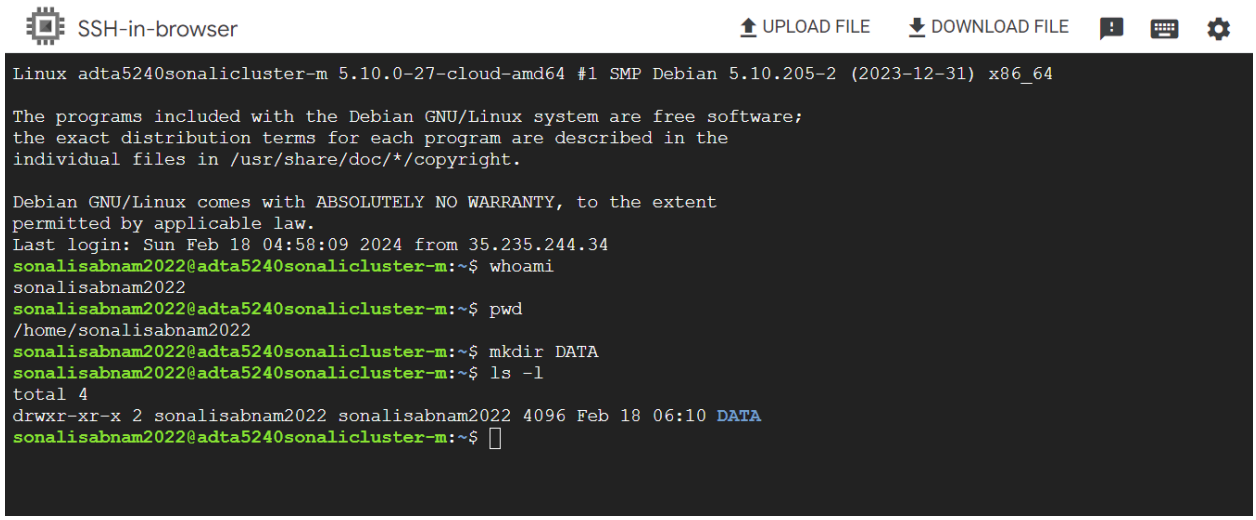


- Click on the SSH to open the SSH Browser and authorize it.
- The new terminal is now open. This is the linux machine so we do not need to use hdfs commands all the time(commands starting with hdfs). But we will still use hdfs commands to copy files in the later steps.

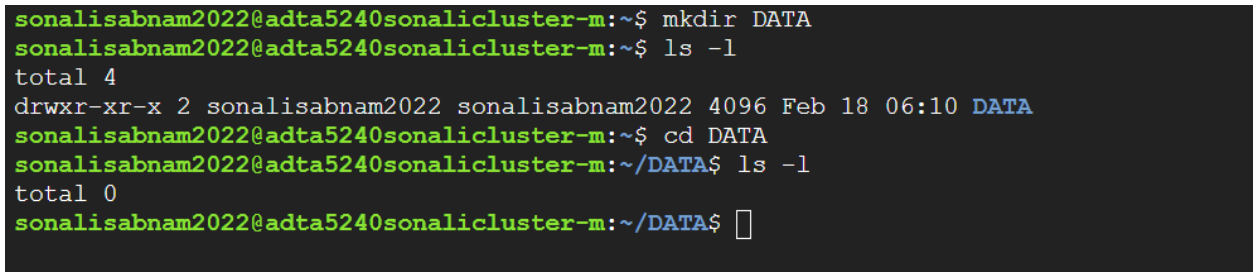


- Type the below commands to go to the required directory.
  - “whoami”
  - “pwd”

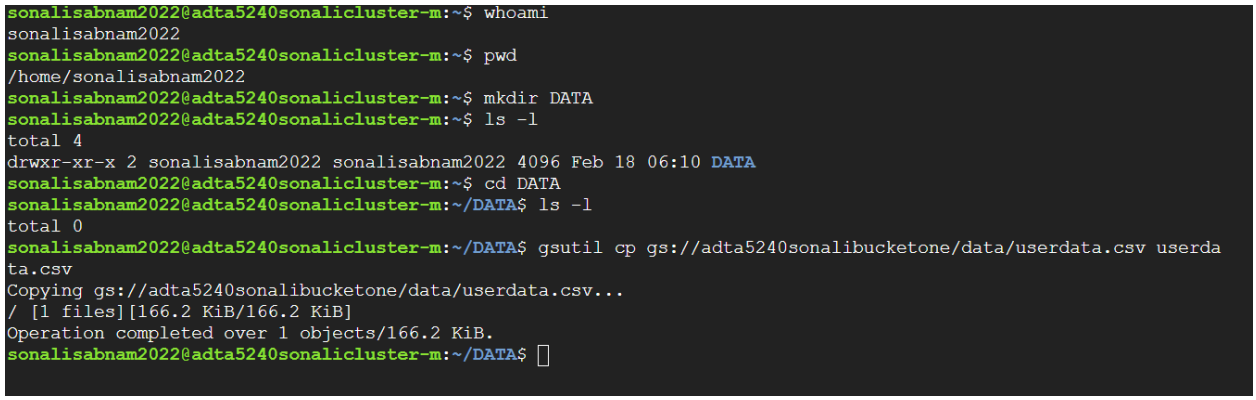
- “mkdir”
- “ls -l”



- Here we can see the “DATA” folder is created.
- Type the below commands to go inside the folder
  - “cd DATA”
  - “ls -l” to list the contents of the folder



- To copy the file “userdata” from GCP storage bucket to HDFS type the below command. “gsutil” is a strong command line utility provided by GCP to perform various operations including file copy on the GCP platform.
  - gsutil cp gs://adta5240sonalibucketone/data/userdata.csv userdata.csv



- The userdata.csv file is now copied.

- To copy the “weblog” file from GCP storage bucket to HDFS type the below command.
    - gsutil cp gs://adta5240sonalibucketone/data/weblog.csv weblog.csv

```
sonalisabnam2022@adta5240sonalicluster-m:~$ whoami
sonalisabnam2022
sonalisabnam2022@adta5240sonalicluster-m:~$ pwd
/home/sonalisabnam2022
sonalisabnam2022@adta5240sonalicluster-m:~$ mkdir DATA
sonalisabnam2022@adta5240sonalicluster-m:~$ ls -l
total 4
drwxr-xr-x 2 sonalisabnam2022 sonalisabnam2022 4096 Feb 18 06:10 DATA
sonalisabnam2022@adta5240sonalicluster-m:~$ cd DATA
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$ ls -l
total 0
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$ gsutil cp gs://adta5240sonalibucketone/data/userdata.csv userda
ta.csv
Copying gs://adta5240sonalibucketone/data/userdata.csv...
/ [1 files][166.2 KiB/166.2 KiB]
Operation completed over 1 objects/166.2 KiB.
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$ gsutil cp gs://adta5240sonalibucketone/data/weblog.csv weblog.c
sv
Copying gs://adta5240sonalibucketone/data/weblog.csv...
/ [1 files][  5.0 MiB/  5.0 MiB]
Operation completed over 1 objects/5.0 MiB.
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$
```

- Now the weblog file is also copied successfully.

```
sonalisabnam2022@adta5240sonalicluster-m:~$ whoami
sonalisabnam2022
sonalisabnam2022@adta5240sonalicluster-m:~$ pwd
/home/sonalisabnam2022
sonalisabnam2022@adta5240sonalicluster-m:~$ mkdir DATA
sonalisabnam2022@adta5240sonalicluster-m:~$ ls -l
total 4
drwxr-xr-x 2 sonalisabnam2022 sonalisabnam2022 4096 Feb 18 06:10 DATA
sonalisabnam2022@adta5240sonalicluster-m:~$ cd DATA
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$ ls -l
total 0
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$ gsutil cp gs://adta5240sonalibucketone/data/userdata.csv userda
ta.csv
Copying gs://adta5240sonalibucketone/data/userdata.csv...
/ [1 files][166.2 KiB/166.2 KiB]
Operation completed over 1 objects/166.2 KiB.
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$ gsutil cp gs://adta5240sonalibucketone/data/weblog.csv weblog.c
sv
Copying gs://adta5240sonalibucketone/data/weblog.csv...
/ [1 files][  5.0 MiB/  5.0 MiB]
Operation completed over 1 objects/5.0 MiB.
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$ ls -l
total 5320
-rw-r--r-- 1 sonalisabnam2022 sonalisabnam2022  170209 Feb 18 06:20 userdata.csv
-rw-r--r-- 1 sonalisabnam2022 sonalisabnam2022 5272996 Feb 18 06:23 weblog.csv
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$
```

- The above screenshot shows the two files in the DATA folder.

Step 6: Copy the files from master node to HDFS

- Type the below command in the Linux terminal to copy “userdata.csv” file
    - “hdfs dfs -put userdata.csv /user/sonalisabnam2022/data/userdata”

```
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$ ls -l
total 5320
-rw-r--r-- 1 sonalisabnam2022 sonalisabnam2022  170209 Feb 18 06:20 userdata.csv
-rw-r--r-- 1 sonalisabnam2022 sonalisabnam2022 5272996 Feb 18 06:23 weblog.csv
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$ hdfs dfs -put userdata.csv /user/sonalisabnam2022/data/userdata
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$ ls-l
-bash: ls-l: command not found
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$ ls -l
total 5320
-rw-r--r-- 1 sonalisabnam2022 sonalisabnam2022  170209 Feb 18 06:20 userdata.csv
-rw-r--r-- 1 sonalisabnam2022 sonalisabnam2022 5272996 Feb 18 06:23 weblog.csv
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$
```

- To verify if the file is copied successfully, type the below command in the HDFS SSH terminal (first terminal).
    - "hdfs dfs -ls /user/sonalisabnam2022/data/userdata"

```
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -ls /user/sonalisabnam2022/data/userdata
Found 1 items
-rw-r--r--   2 sonalisabnam2022 hadoop     170209 2024-02-18 06:32 /user/sonalisabnam2022/data/userdata/userdata
.csv
sonalisabnam2022@adta5240sonalicluster-m:~$
```

- The above screenshot shows the "userdata.csv" file successfully copied to HDFS.
- Type the below command in the Linux terminal to copy "weblog.csv" file.
    - "hdfs dfs -put weblog.csv /user/sonalisabnam2022/data/weblog"

```
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$ hdfs dfs -put weblog.csv /user/sonalisabnam2022/data/weblog
sonalisabnam2022@adta5240sonalicluster-m:~/DATA$
```

- To verify if the file is copied successfully, type the below command in the HDFS SSH terminal (first terminal).
    - "hdfs dfs -ls /user/sonalisabnam2022/data/weblog"

```
sonalisabnam2022@adta5240sonalicluster-m:~$ hdfs dfs -ls /user/sonalisabnam2022/data/weblog
Found 1 items
-rw-r--r--   2 sonalisabnam2022 hadoop    5272996 2024-02-18 06:43 /user/sonalisabnam2022/data/weblog/weblog.csv
sonalisabnam2022@adta5240sonalicluster-m:~$
```

- With this we have successfully copied the files from master node to the HDFS file system.

Step 7: Turn off the cluster

- Go to dashboard and type cluster in the search bar.
- Once the cluster is displayed, select the checkbox and click on STOP.
- The below screenshot shows that my cluster is now stopped.