Setting Virtual Machine through Linux commands

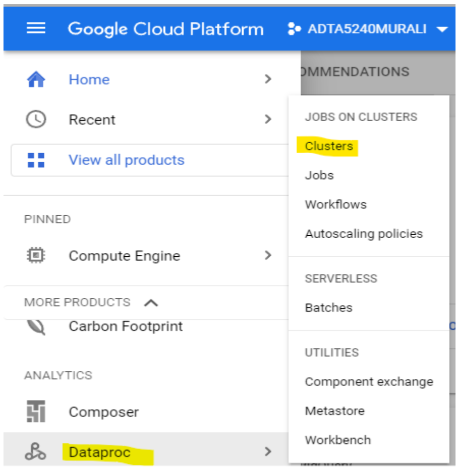
We will access the Hadoop ecosystem using Linuz commands through Dataproc on Google Cloud Platform.

You must have your GCP account set up and ready with the following:

1) Active GCP account
2) A project on Hadoop-Spark cluster
3) A storage bucket ready for use (we added user dataset, refer to previous document on how to create clusters and bucket)
4) The cluster must contain 1 master node and 2 worker nodes. (Make sure to turn on/off during and after completion of assignment-refer to previous document to start and stop clusters)

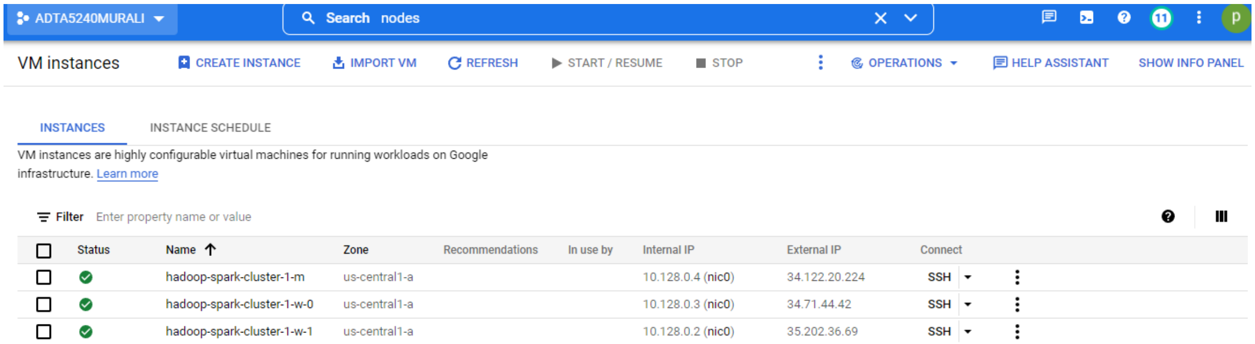
TURN ON CLUSTERS!

1) Log into your GCP account. Open the navigator (3 horizontal lines) and under "clusters", select "Clusters dataproc".



IMPORTANT! It is alright if your clusters are running but make sure to shut down nodes after working on your project.

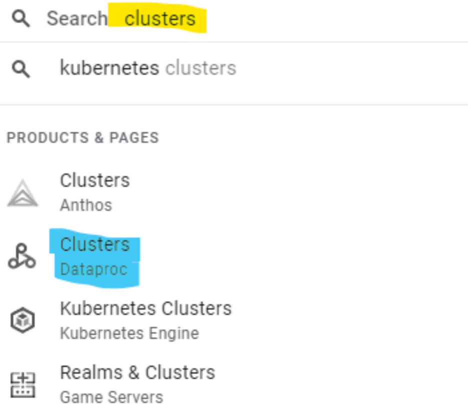
2) To turn on clusters – Under the navigation panel, select "Compute Engines" – VM instances and select nodes to turn on. This is what your panel should look like. (Refer previous documents to learn creation of nodes as well as to turn on or off nodes)

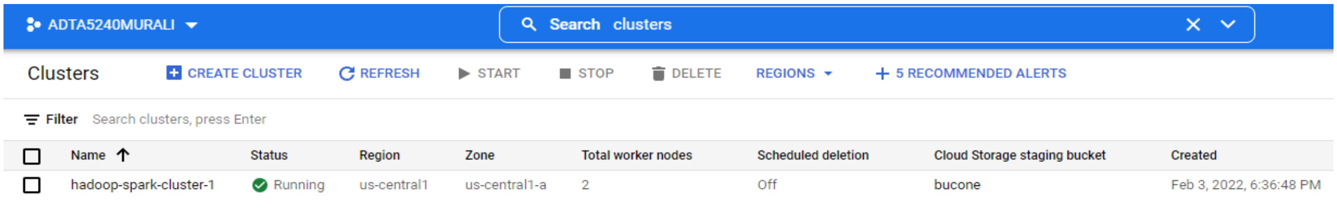


VM instances are highly configurable virtual machines for running workloads on Google infrastructure. Learn more

□	Status	Name ↑	Zone	Recommendations	In use by	Internal IP	External IP	Connect
□	✓	hadoop-spark-cluster-1-m	us-central1-a			10.128.0.4 (nic0)	34.122.20.224	SSH
□	✓	hadoop-spark-cluster-1-w-0	us-central1-a			10.128.0.3 (nic0)	34.71.44.42	SSH
□	✓	hadoop-spark-cluster-1-w-1	us-central1-a			10.128.0.2 (nic0)	35.202.36.69	SSH

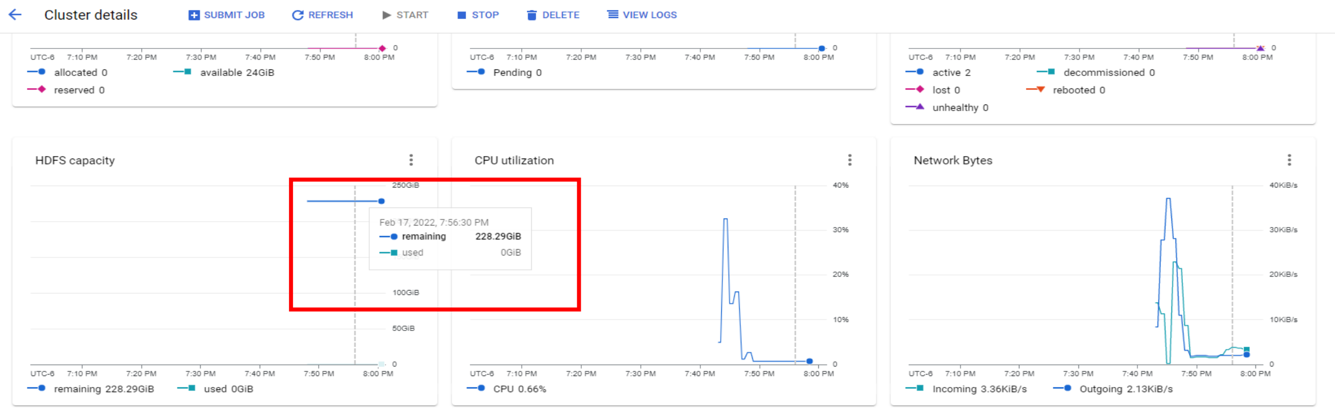
3) In the search tab type “Clusters” and select Clusters created through Dataproc.



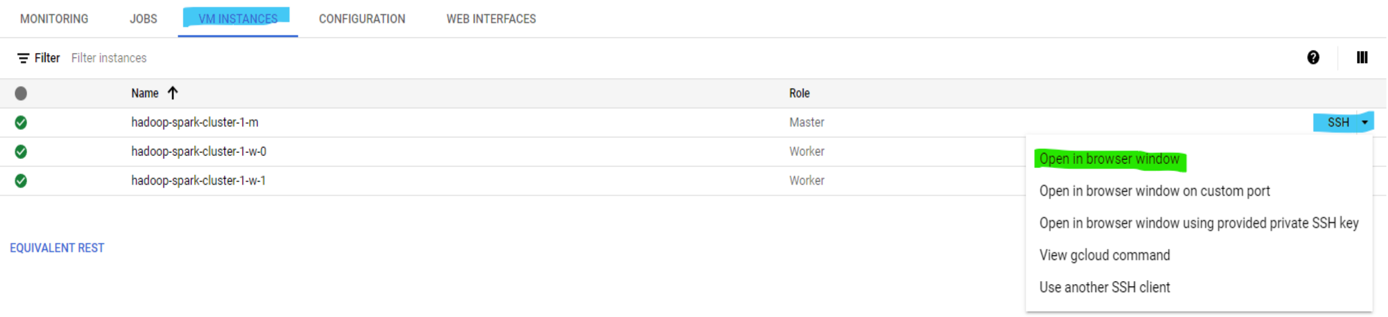


□	Name ↑	Status	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created
□	hadoop-spark-cluster-1	✓ Running	us-central1	us-central1-a	2	Off	bucone	Feb 3, 2022, 6:36:48 PM

4) Click on the cluster name and scroll down. You will be able to see your usage. There are a lot of graphs showing your usage. We are interested in in HDFS capacity/usage.



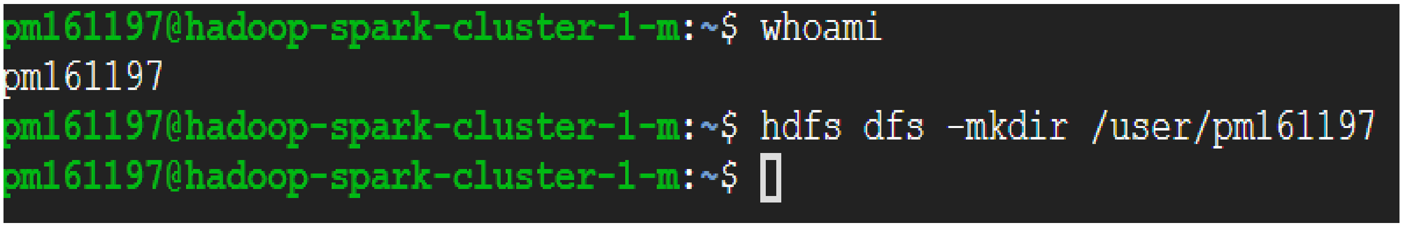
5) We will access the nodes through Linux. Scroll up and click on “VM instances”. Click on SSH drop down menu and select “Open in browser window”.



●	Name ↑	Role
✓	hadoop-spark-cluster-1-m	Master
✓	hadoop-spark-cluster-1-w-0	Worker
✓	hadoop-spark-cluster-1-w-1	Worker

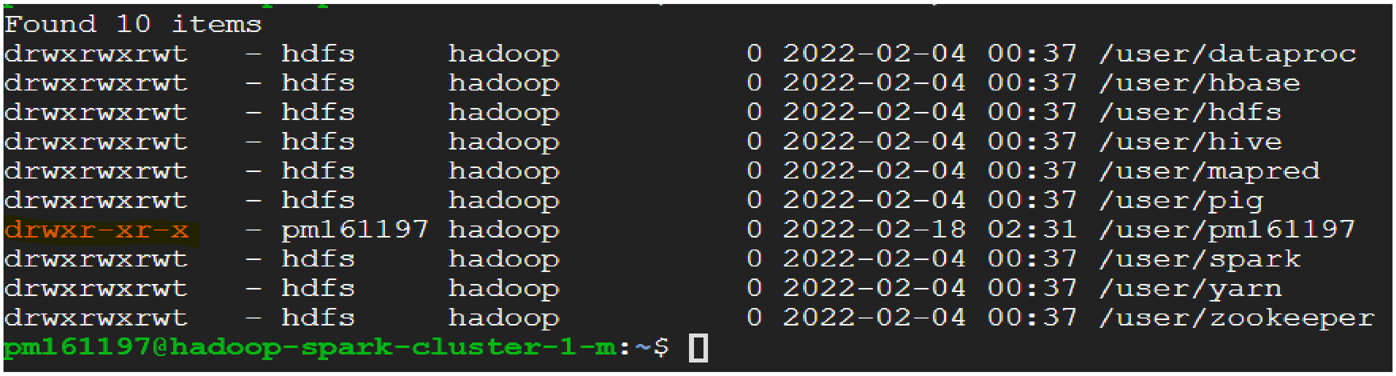
SSH
Open in browser window
Open in browser window on custom port
Open in browser window using provided private SSH key
View gcloud command
Use another SSH client

6) We will access the nodes through SSH using Linux commands. Type “clear” to start with a clean terminal.
7) We are going create a new directory in the folder. We recommend using same name for the directory and username. To find out your username, just type “whoami”.
8) Type: hdfs dfs -mkdir /user/pm161197 >>> Creating subfolder with same name as username.



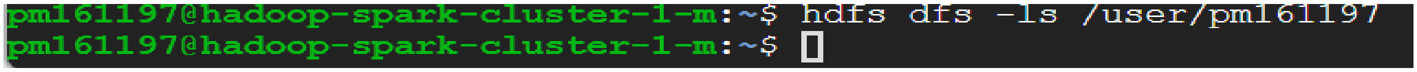
```
pm161197@hadoop-spark-cluster-1-m:~$ whoami
pm161197
pm161197@hadoop-spark-cluster-1-m:~$ hdfs dfs -mkdir /user/pm161197
pm161197@hadoop-spark-cluster-1-m:~$ 
```

9) Type: hdfs dfs -ls /user >>> It will display everything that is there in the directory.
10) drwxrwxrwt >>> “d” is for directory, “r” is for read,” w” is for write and “x” is for execute.We as owners have access to read, write and execute.



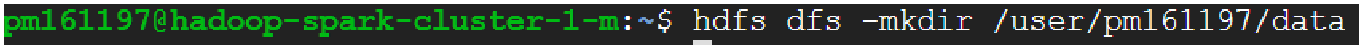
```
Found 10 items
drwxrwxrwt   - hdfs     hadoop          0 2022-02-04 00:37 /user/dataproc
drwxrwxrwt   - hdfs     hadoop          0 2022-02-04 00:37 /user/hbase
drwxrwxrwt   - hdfs     hadoop          0 2022-02-04 00:37 /user/hdfs
drwxrwxrwt   - hdfs     hadoop          0 2022-02-04 00:37 /user/hive
drwxrwxrwt   - hdfs     hadoop          0 2022-02-04 00:37 /user/mapred
drwxrwxrwt   - hdfs     hadoop          0 2022-02-04 00:37 /user/pig
drwxr-xr-x   - pm161197 hadoop          0 2022-02-18 02:31 /user/pm161197
drwxrwxrwt   - hdfs     hadoop          0 2022-02-04 00:37 /user/spark
drwxrwxrwt   - hdfs     hadoop          0 2022-02-04 00:37 /user/yarn
drwxrwxrwt   - hdfs     hadoop          0 2022-02-04 00:37 /user/zookeeper
pm161197@hadoop-spark-cluster-1-m:~$ 
```

11) Type: hdfs dfs -ls /user/pm161197 >>> It will not return anything because the file is empty.



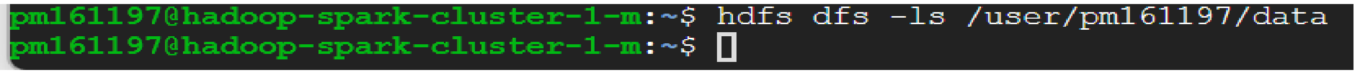
```
pm161197@hadoop-spark-cluster-1-m:~$ hdfs dfs -ls /user/pm161197
pm161197@hadoop-spark-cluster-1-m:~$ 
```

12) Create a subfolder “data” which will contain the data that we uploaded to GCP.
13) Type: hdfs dfs -mkdir /user/pm161197/data >>> Creating a subfolder “data” under folder “pm161197”.



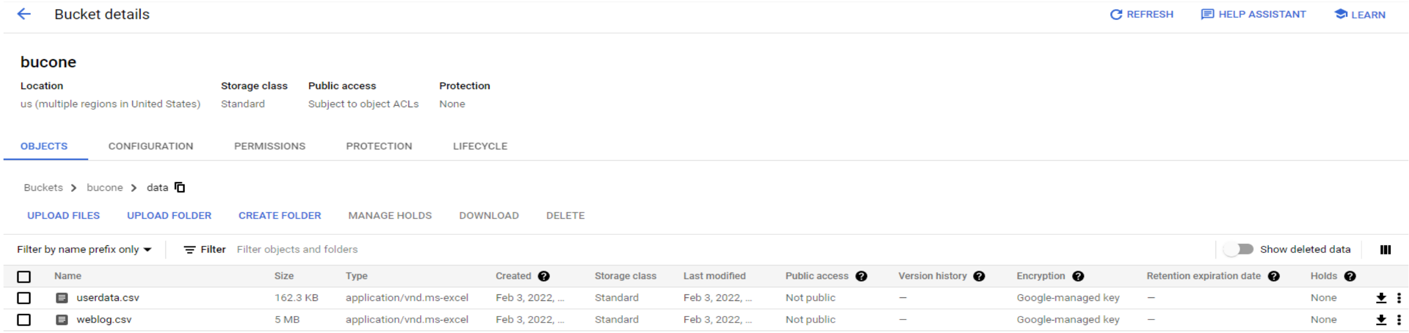
```
pm161197@hadoop-spark-cluster-1-m:~$ hdfs dfs -mkdir /user/pm161197/data
```

14) Type: hdfs dfs -ls /user/pm161197/data >>> Used to list everything in the data folder. In this case it will not return anything as we did not add anything to it.



```
pm161197@hadoop-spark-cluster-1-m:~$ hdfs dfs -ls /user/pm161197/data
pm161197@hadoop-spark-cluster-1-m:~$ 
```

15) While working on GCP window, we uploaded two datasets to the bucket. We will copy the two datasets namely “userdata.csv” and “weblog.csv”.



bucone

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Subject to object ACLs	None

☐	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Retention expiration date	Holds
☐	userdata.csv	162.3 KB	application/vnd.ms-excel	Feb 3, 2022, ...	Standard	Feb 3, 2022, ...	Not public	–	Google-managed key	–	None
☐	weblog.csv	5 MB	application/vnd.ms-excel	Feb 3, 2022, ...	Standard	Feb 3, 2022, ...	Not public	–	Google-managed key	–	None

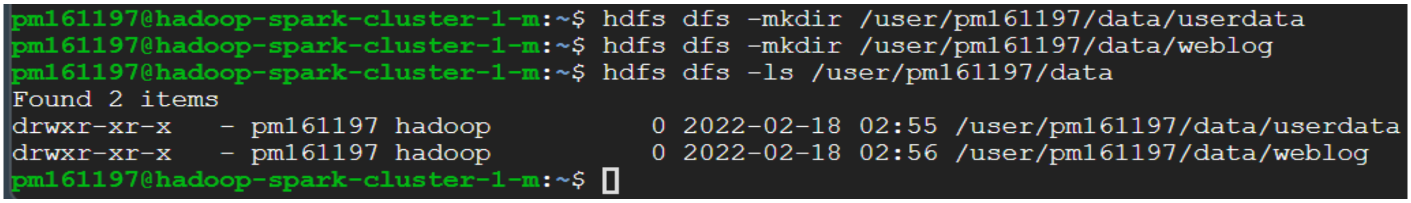
16) Follow the steps to create subfolder inside “data” to dump the two datasets.

Type: hdfs dfs -mkdir /user/pm161197/data/userdata

Type: hdfs dfs -ls /user/pm161197/data/userdata

Type: hdfs dfs -mkdir /user/pm161197/data/weblog

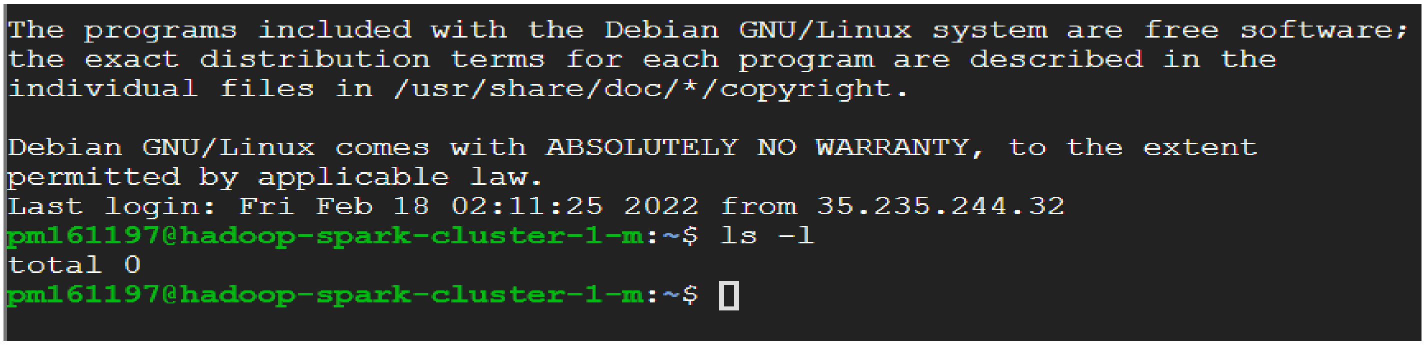
Type: hdfs dfs -ls /user/pm161197/data



```
pm161197@hadoop-spark-cluster-1-m:~$ hdfs dfs -mkdir /user/pm161197/data/userdata
pm161197@hadoop-spark-cluster-1-m:~$ hdfs dfs -mkdir /user/pm161197/data/weblog
pm161197@hadoop-spark-cluster-1-m:~$ hdfs dfs -ls /user/pm161197/data
Found 2 items
drwxr-xr-x   - pm161197 hadoop          0 2022-02-18 02:55 /user/pm161197/data/userdata
drwxr-xr-x   - pm161197 hadoop          0 2022-02-18 02:56 /user/pm161197/data/weblog
pm161197@hadoop-spark-cluster-1-m:~$ 
```

17) To be done on GCP: Open another SSH terminal from the VM instances (Navigation >> Compute Engine >> VM Instances >> SSH >> Open in browser window.

18) Type : ls -l >>> It will not display anything as the folder is empty.



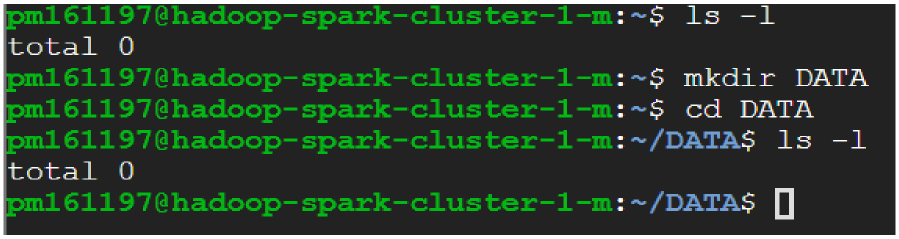
```
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Fri Feb 18 02:11:25 2022 from 35.235.244.32
pm161197@hadoop-spark-cluster-1-m:~$ ls -l
total 0
pm161197@hadoop-spark-cluster-1-m:~$ 
```

19) Creation of directory “DATA”. Type: mkdir DATA

20) To work on “DATA”. Type: cd DATA

21) To see contents of “DATA”. Type: ls -l



```
pm161197@hadoop-spark-cluster-1-m:~$ ls -l
total 0
pm161197@hadoop-spark-cluster-1-m:~$ mkdir DATA
pm161197@hadoop-spark-cluster-1-m:~$ cd DATA
pm161197@hadoop-spark-cluster-1-m:~/DATA$ ls -l
total 0
pm161197@hadoop-spark-cluster-1-m:~/DATA$ ▯
```

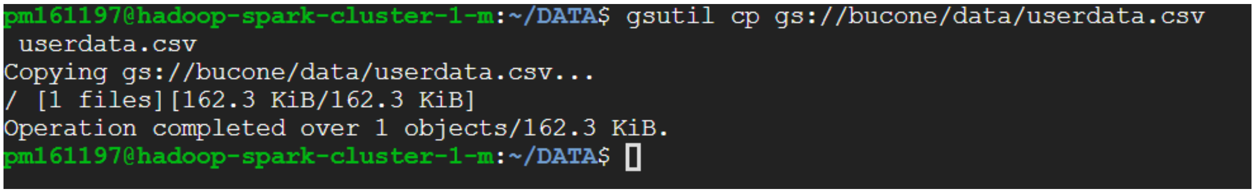
22) Copy the dataset from GCP bucket into HDFS

➔ Copy the GCP bucket to the master node of cluster.

➔ Type : gsutil cp gs://bucone/data/userdata.csv userdata.csv

➔Name of my bucket is bucone.

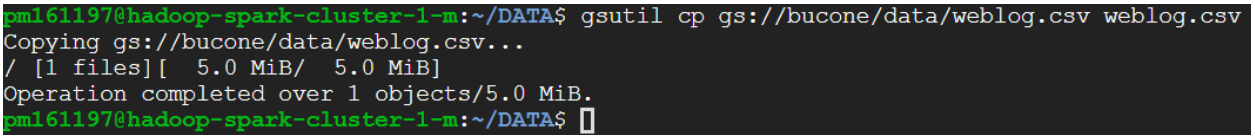
➔We are basically copying the data in bucone into the userdata subfolder we created through previous commands.



```
pm161197@hadoop-spark-cluster-1-m:~/DATA$ gsutil cp gs://bucone/data/userdata.csv
 userdata.csv
Copying gs://bucone/data/userdata.csv...
/ [1 files][162.3 KiB/162.3 KiB]
Operation completed over 1 objects/162.3 KiB.
pm161197@hadoop-spark-cluster-1-m:~/DATA$ ▯
```

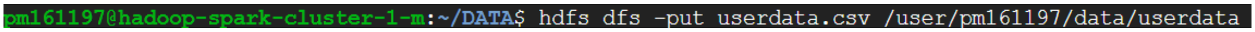
➔We will do the same for the other dataset

➔Type: gsutil cp gs://bucone/data/weblog.csv weblog.csv



```
pm161197@hadoop-spark-cluster-1-m:~/DATA$ gsutil cp gs://bucone/data/weblog.csv weblog.csv
Copying gs://bucone/data/weblog.csv...
/ [1 files][  5.0 MiB/  5.0 MiB]
Operation completed over 1 objects/5.0 MiB.
pm161197@hadoop-spark-cluster-1-m:~/DATA$ ▯
```

23) Move the files from the master note into the HDFS. Type: hdfs dfs -put userdata.csv /user/pm161197/data/userdata

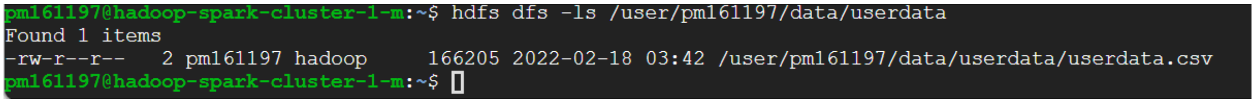


```
pm161197@hadoop-spark-cluster-1-m:~/DATA$ hdfs dfs -put userdata.csv /user/pm161197/data/userdata
```

24) Check if the data has been moved. - You must type the following in the first terminal if you lost connection with it.

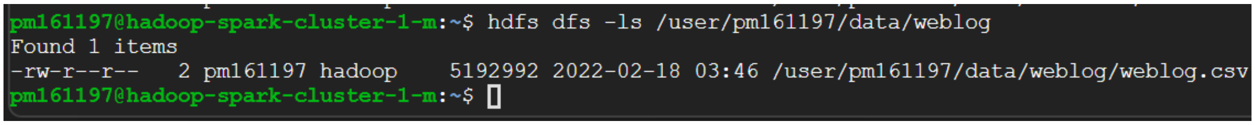
Hdfs dfs -ls ‘user/pm161197/data

25) Type the following in the first terminal: hdfs dfs -ls /user/pm161197/data/userdata



```
pm161197@hadoop-spark-cluster-1-m:~$ hdfs dfs -ls /user/pm161197/data/userdata
Found 1 items
-rw-r--r--   2 pm161197 hadoop     166205 2022-02-18 03:42 /user/pm161197/data/userdata/userdata.csv
pm161197@hadoop-spark-cluster-1-m:~$ ▯
```

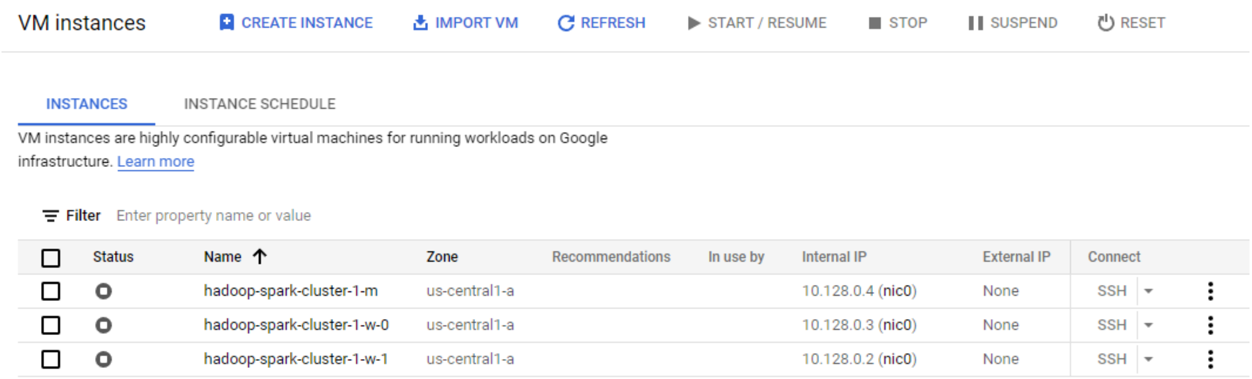
26) Type : hdfs dfs -ls user/pm161197/data/weblog



```
pm161197@hadoop-spark-cluster-1-m:~$ hdfs dfs -ls /user/pm161197/data/weblog
Found 1 items
-rw-r--r--   2 pm161197 hadoop    5192992 2022-02-18 03:46 /user/pm161197/data/weblog/weblog.csv
pm161197@hadoop-spark-cluster-1-m:~$ ▯
```

All the data has been moved into the master node.
That brings us to end of setting up VM through linux.

TURN OF YOUR CLUSTERS!



VM instances are highly configurable virtual machines for running workloads on Google infrastructure. Learn more

□	Status	Name ↑	Zone	Recommendations	In use by	Internal IP	External IP	Connect	
□	⊙	hadoop-spark-cluster-1-m	us-central1-a			10.128.0.4 (nic0)	None	SSH ▾	⋮
□	⊙	hadoop-spark-cluster-1-w-0	us-central1-a			10.128.0.3 (nic0)	None	SSH ▾	⋮
□	⊙	hadoop-spark-cluster-1-w-1	us-central1-a			10.128.0.2 (nic0)	None	SSH ▾	⋮