Set up Google Cloud

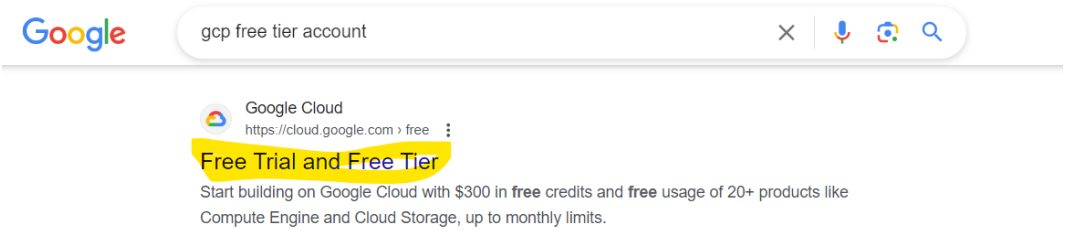
Create Hadoop Cluster

Set up GCloud CLI

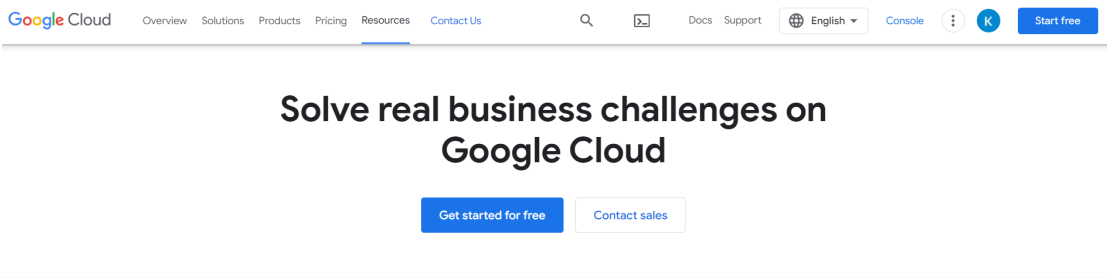
Upload sample file to HDFS

Let's get started:

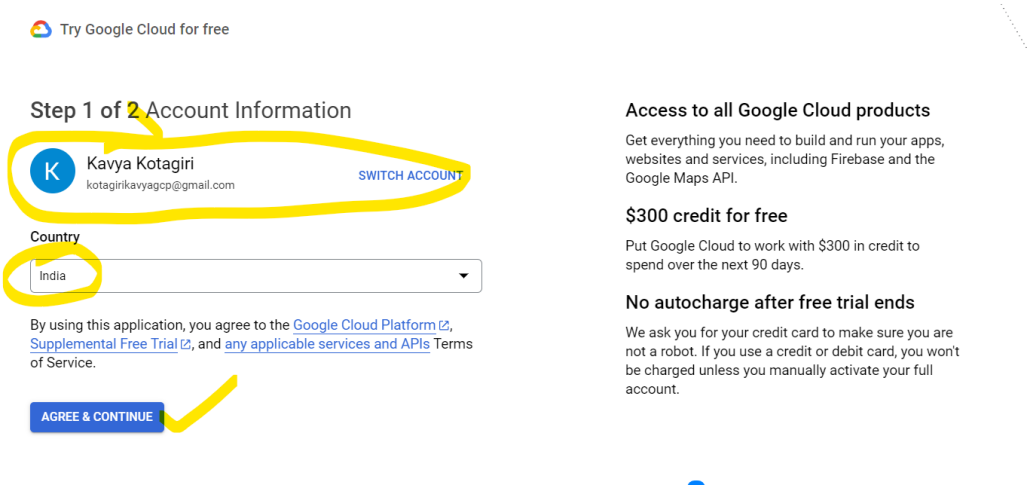
1. Open the browser and search for "gcp free tier account"
2. Click on the below link https://cloud.google.com/free highlighted in the screen snippet below:



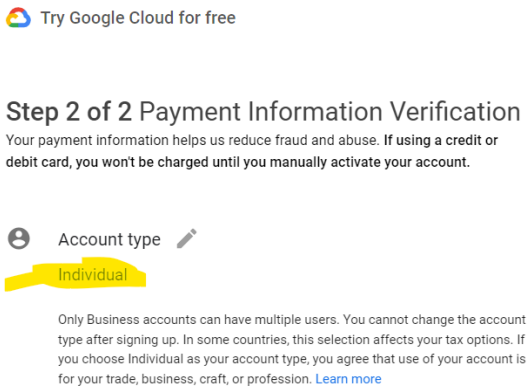
3. Click on **Get Started for Free**



4. Select the Google Account, Country and then click on **AGREE & CONTINUE**

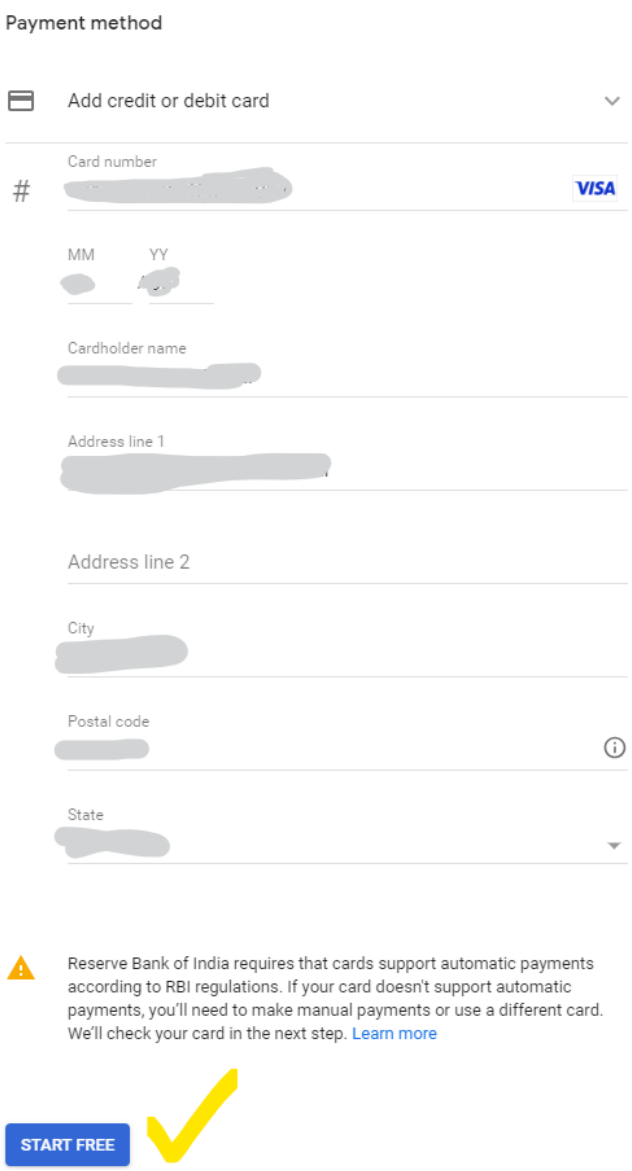


5. Select Account Type as **Individual**

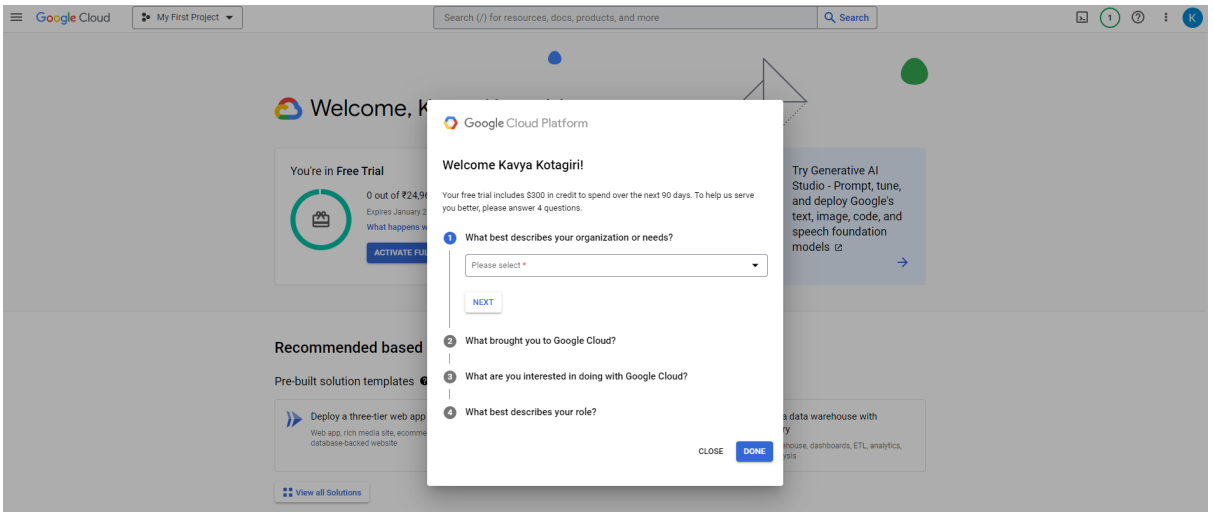


Note: Make sure the card is VISA or MASTERCARD and International Usage is enabled for your card.

6. Now enter the card details and Address then click on **START FREE**.

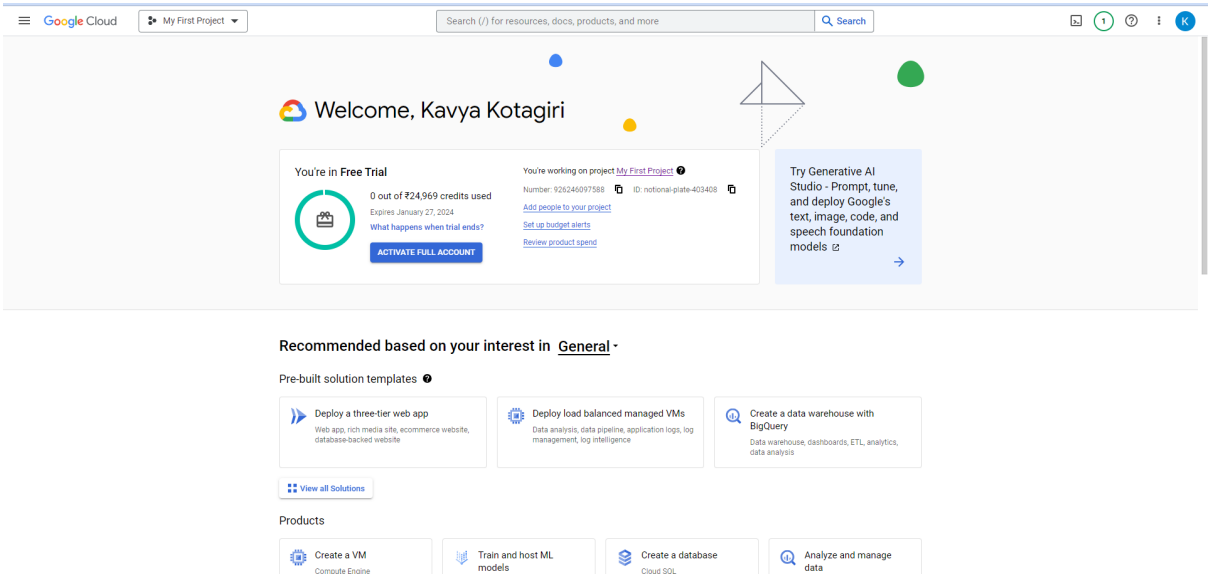


7. Once the card details are verified and amount of 2 INR will be deducted. The below screen comes up.



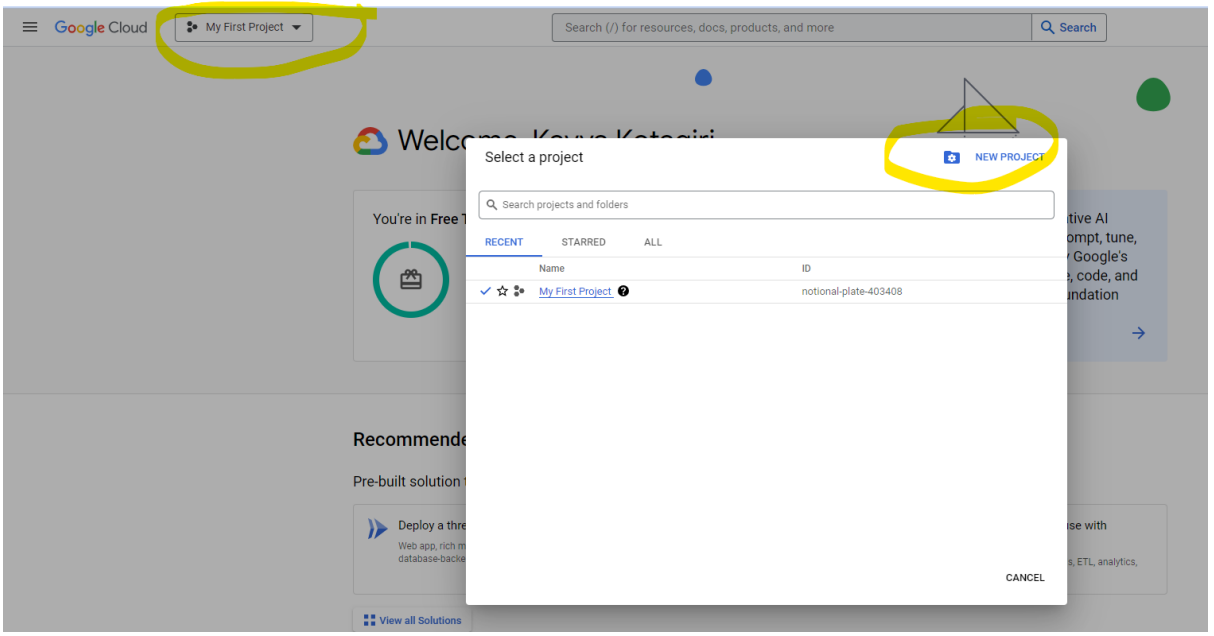
You can either provide the details or just click CLOSE.

8. You will see the below screen which has your free credit details and used credits and other details.

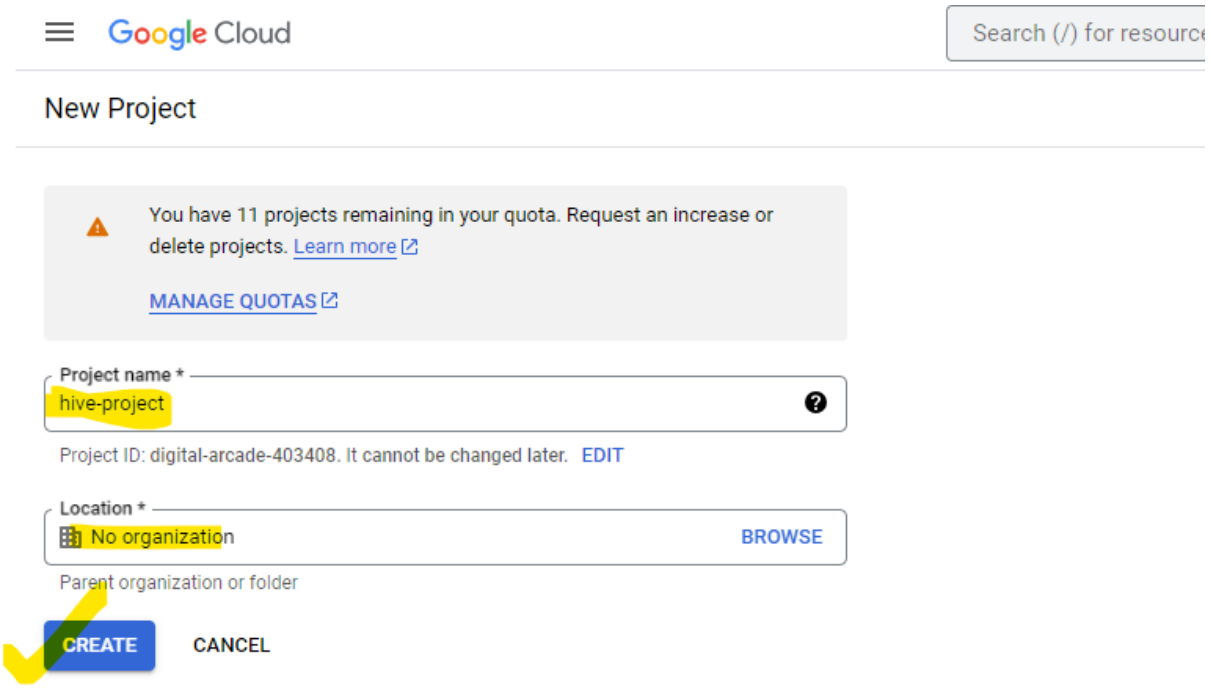


Now let's create a new PROJECT.

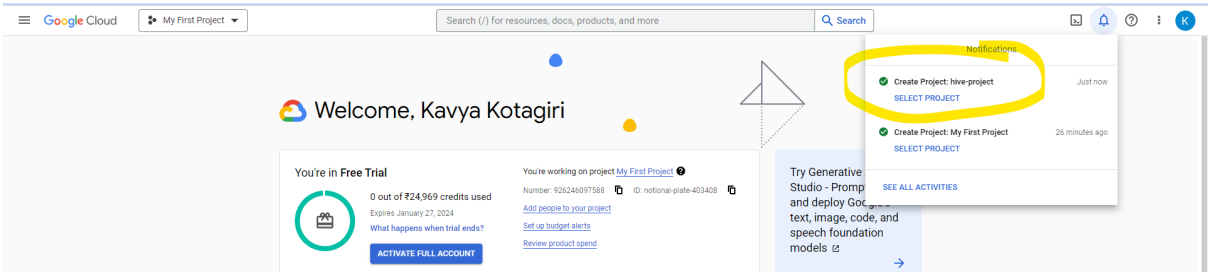
9. Click on the **My First Project** drop down on the top and click on **NEW PROJECT**.



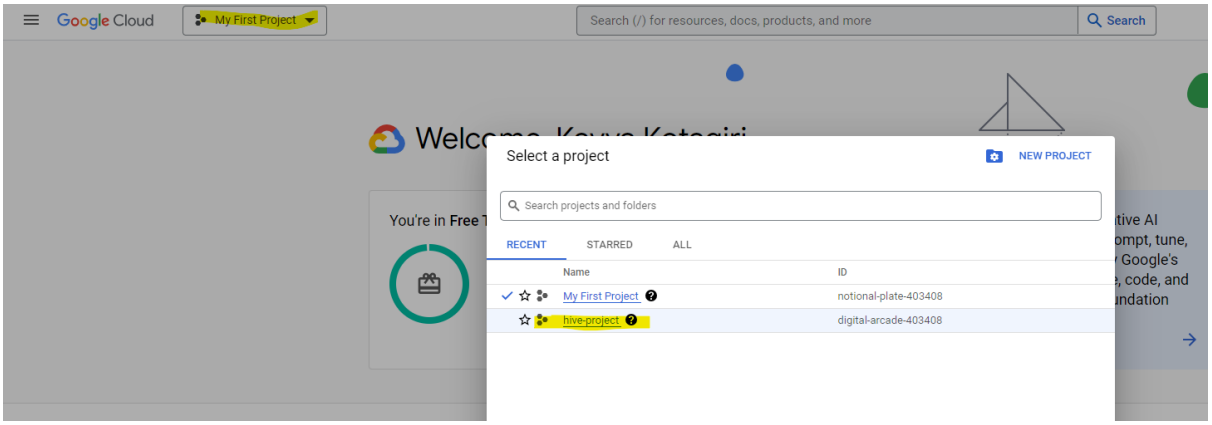
10. Provide some name to your project and do not change anything in Location let it be No Organization and on click on **CREATE**



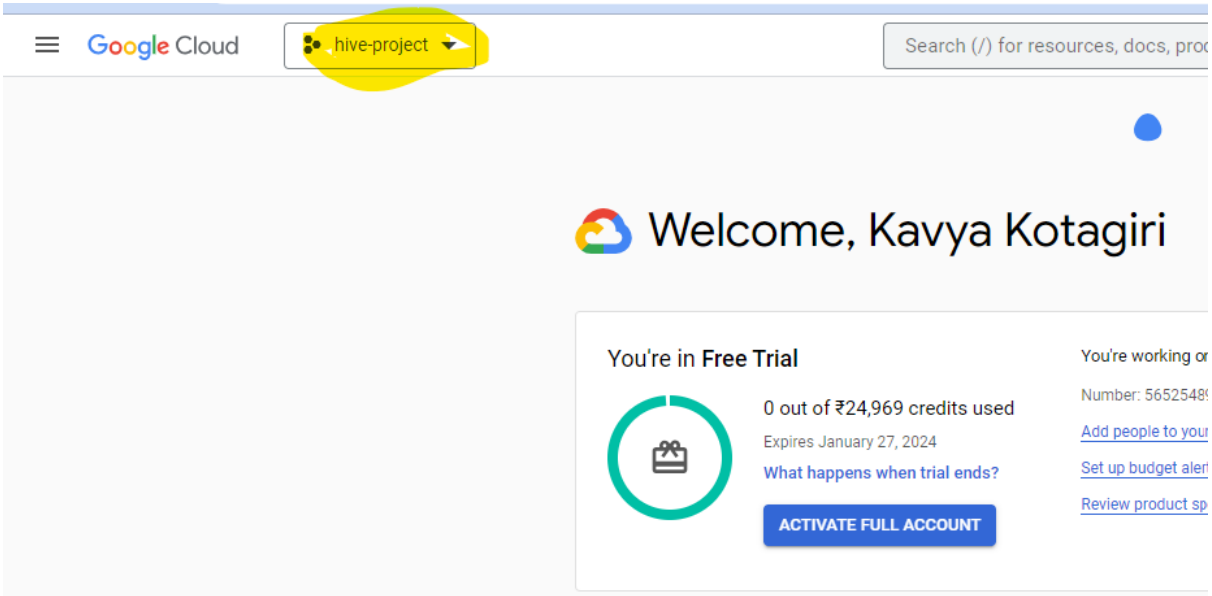
11. You will be redirected1 to home page and a notification with green tick will be displayed once the new project is created.



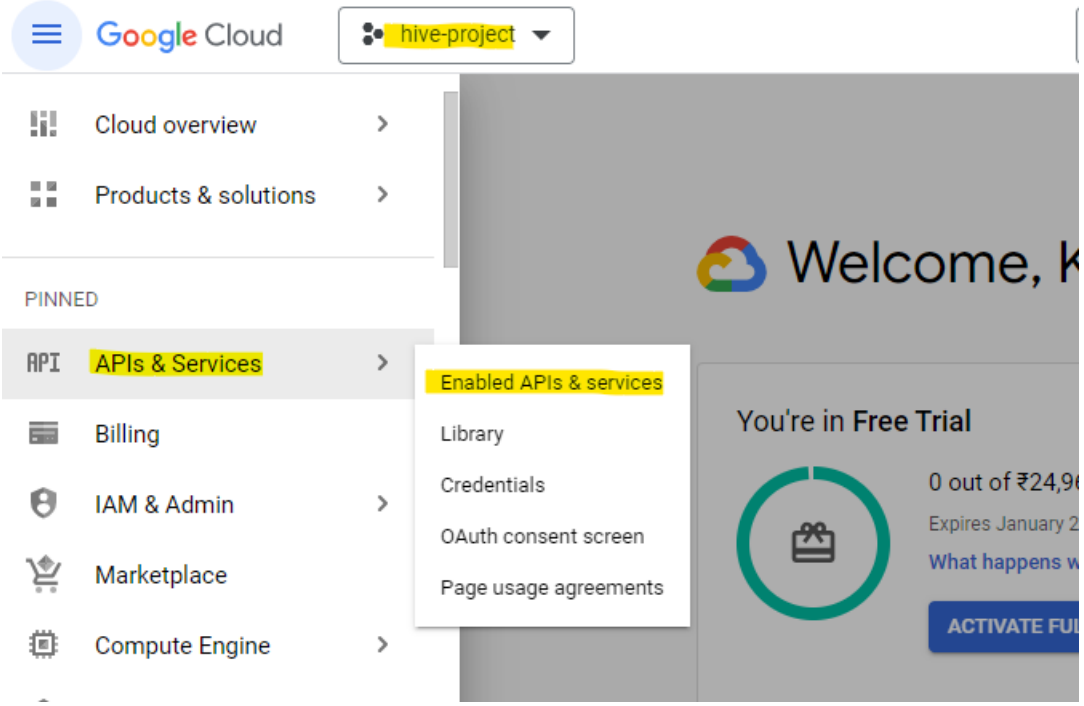
12. Now click on the My First Project drop down and select the project you selected.



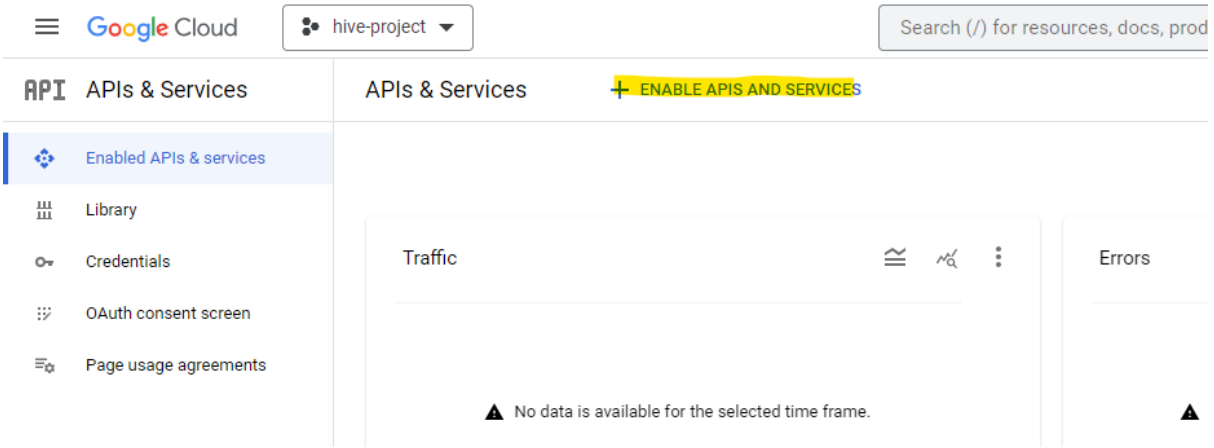
13. Once the project is selected it will start showing the project selected on the top left of your screen.



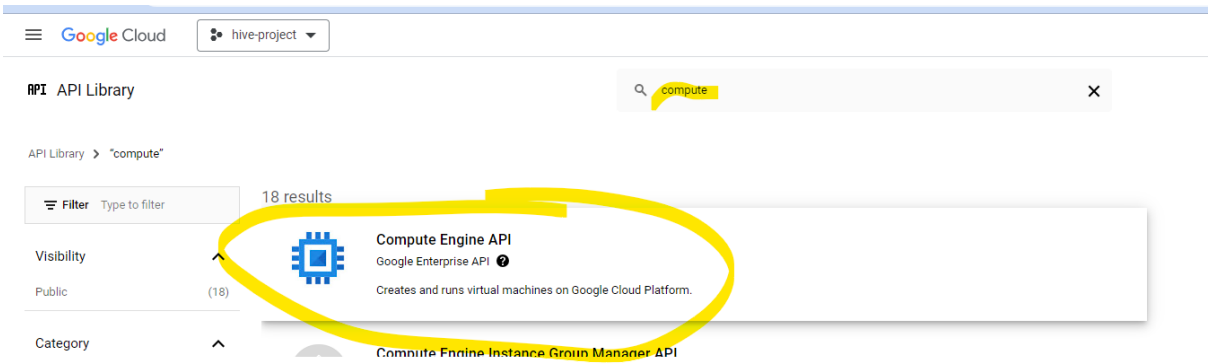
14. Now we need to enable 3 APIs.
    Click on again 3 lines and now select **APIs & Services**.

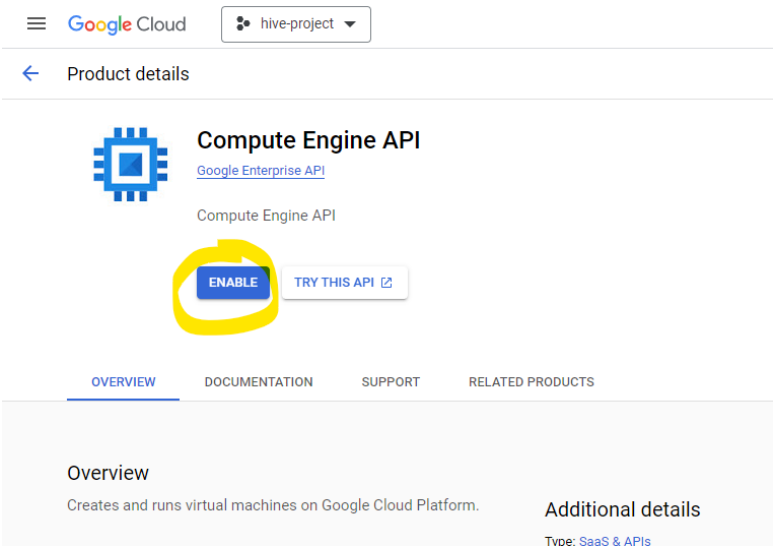


Click on **+ ENABLE APIS AND SERVICES**

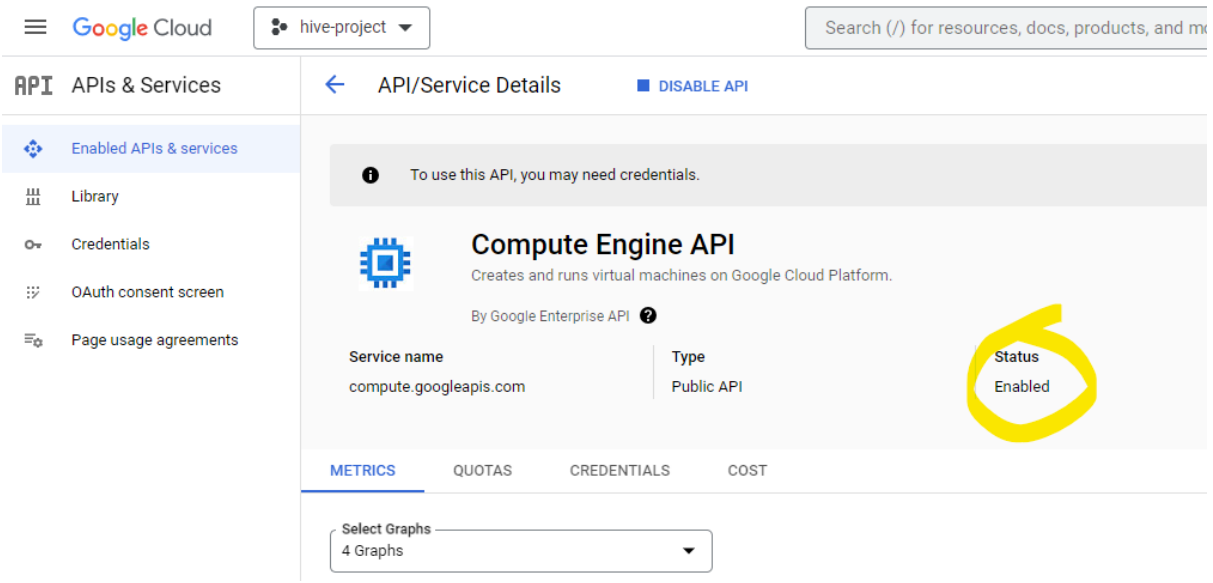


Search for compute and select **Compute Engine API**



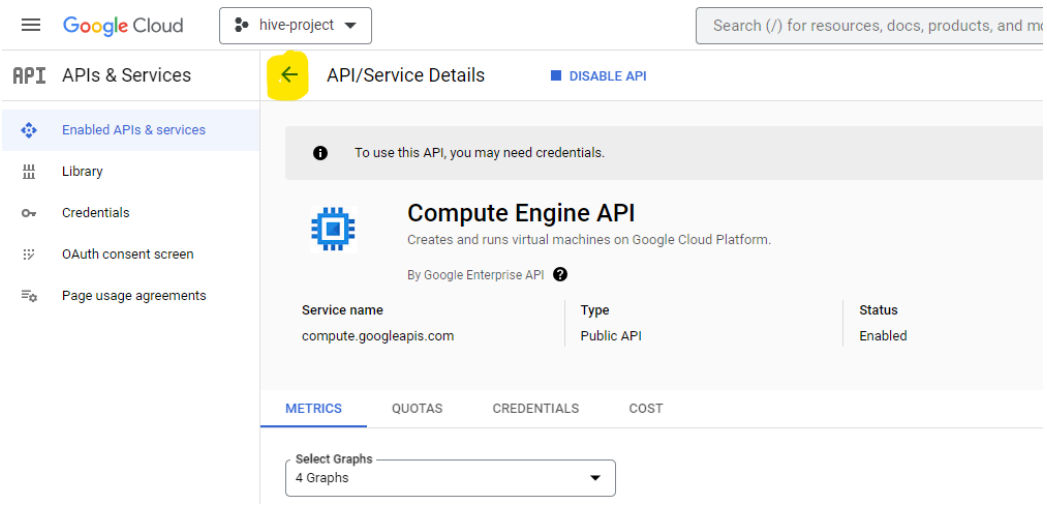


Once its enabled, you will see status as Enabled.

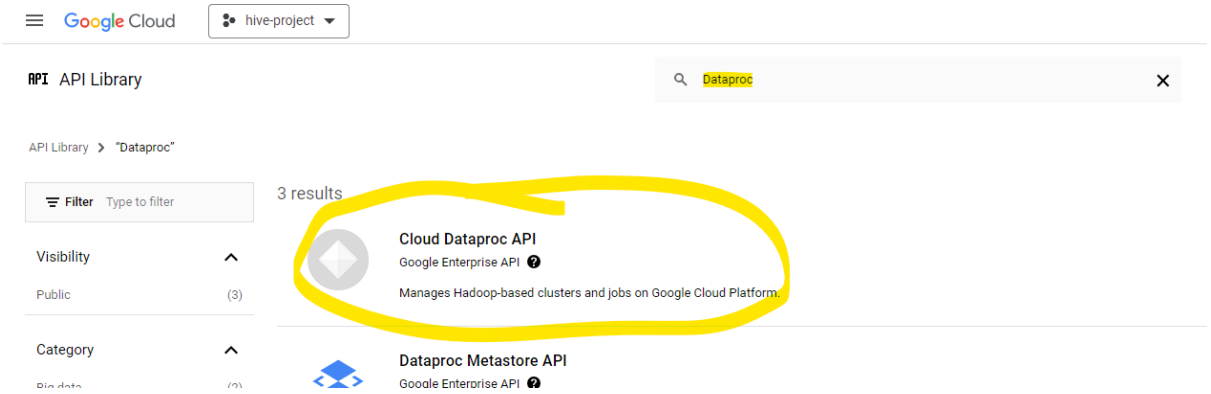


Now let's enable other 2 APIs as well.

Go back by clicking on back arrow and click on **+ ENABLE APIS AND SERVICES**

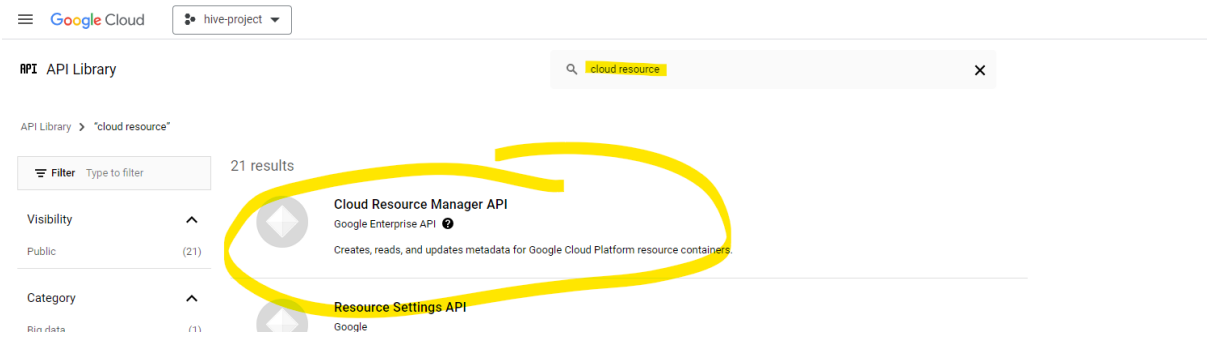


Search for Dataproc and select **Cloud Dataproc API**



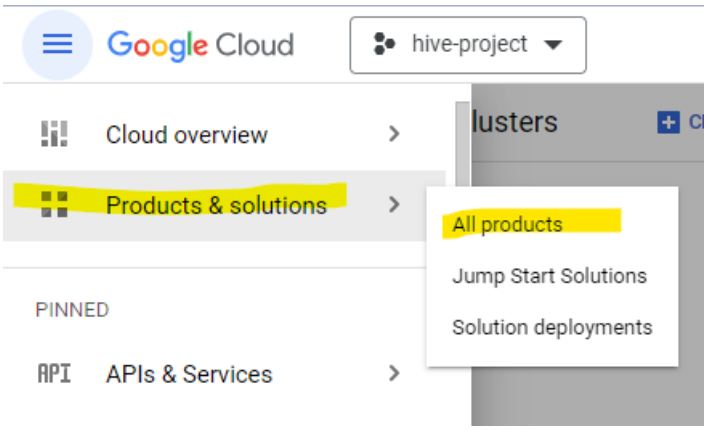
Once its enabled, go back.

Search for cloud resource and select **Cloud Resource Manager API**.

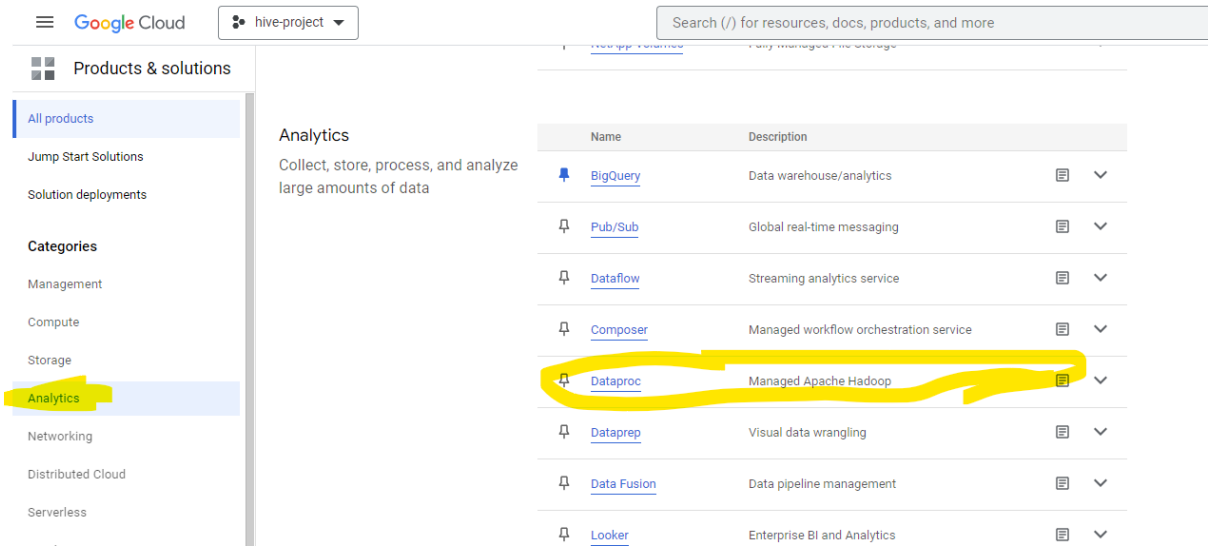


We are now done with enabled the required APIs.

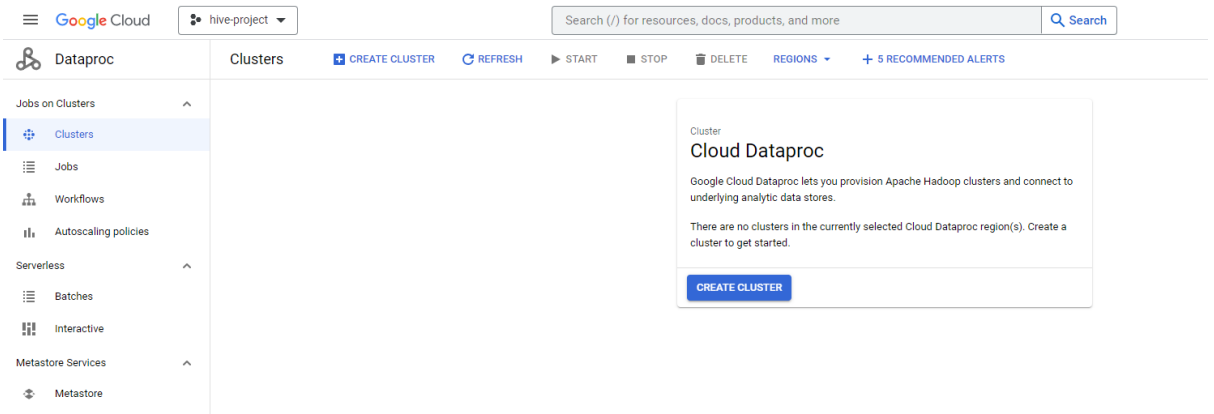
15. Now let's start creating Hadoop cluster.
    Click on 3 lines on top left, click Products & solutions then All products



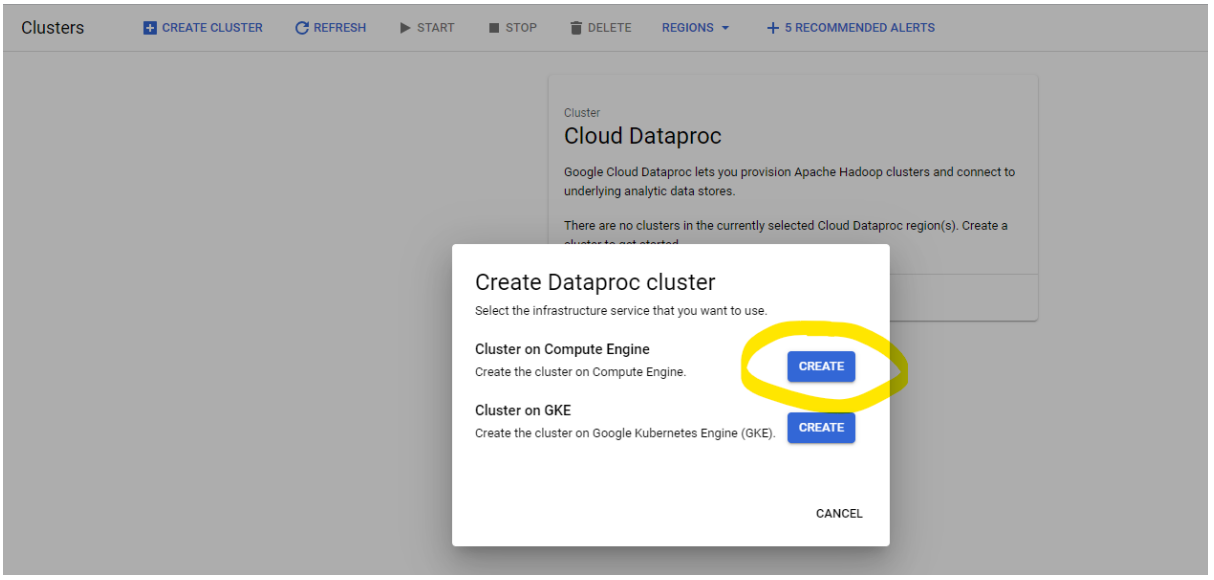
16. Scroll down to Analytics Category and you will find **Dataproc.**
    Click on Dataproc.



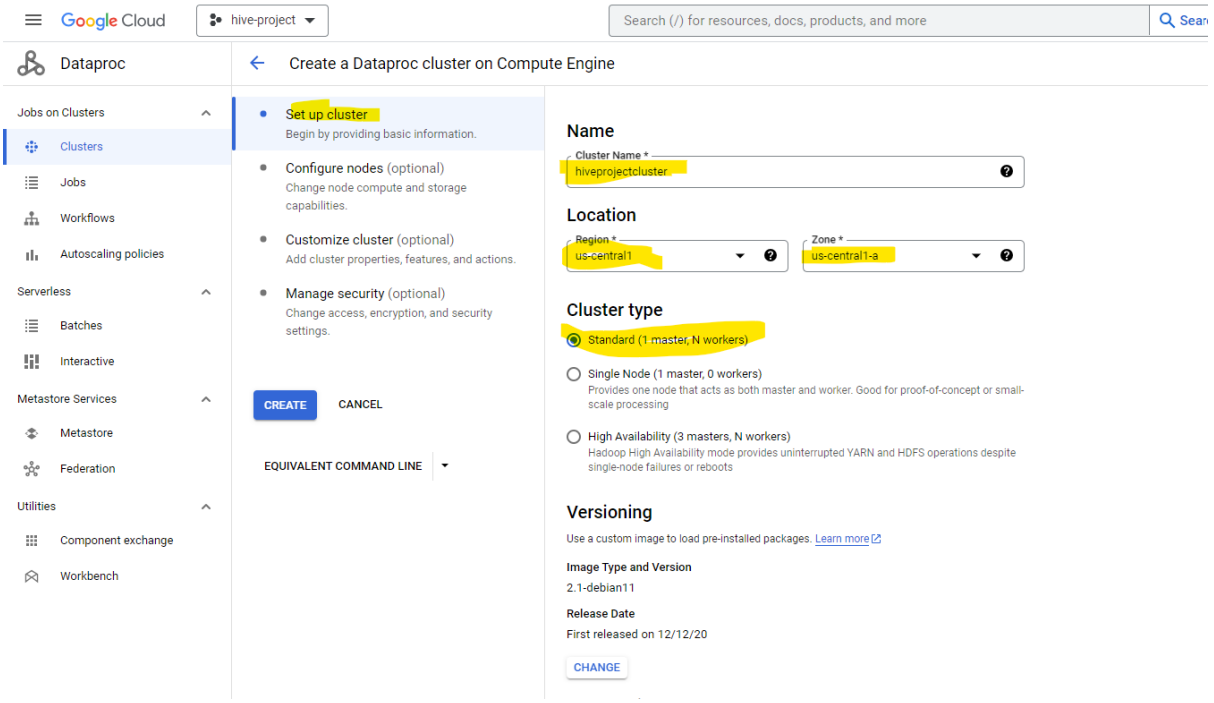
17. You will see the below screen.
    If you face any error, just refresh the page or trying re-logging in.



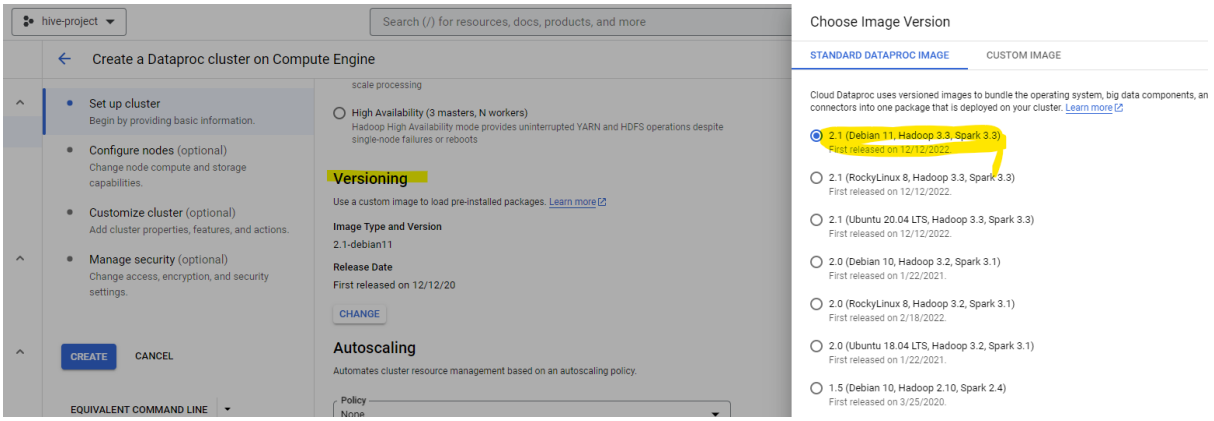
18. Now click on **CREATE CLUSTER**
Click **CREATE** option of **Cluster on Compute Engine**



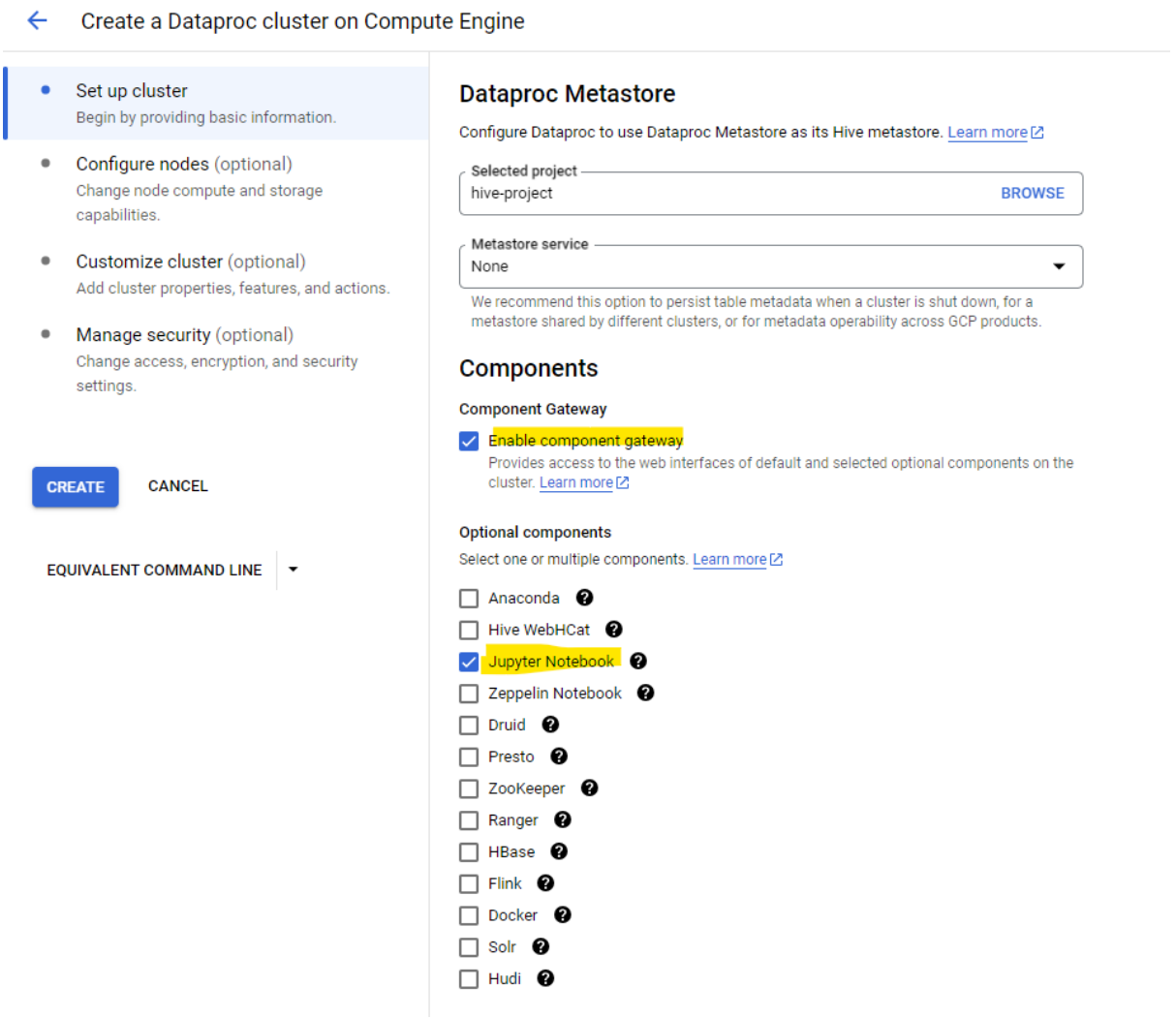
19. Now in **Set up cluster,** give Cluster Name, Region and select zone as us-central1-a and Cluster type as Standard(1 master, N workers).



Don't change anything in Versioning, let it be **2.1 (Debian 11, Hadoop 3.3, Spark 3.3)**

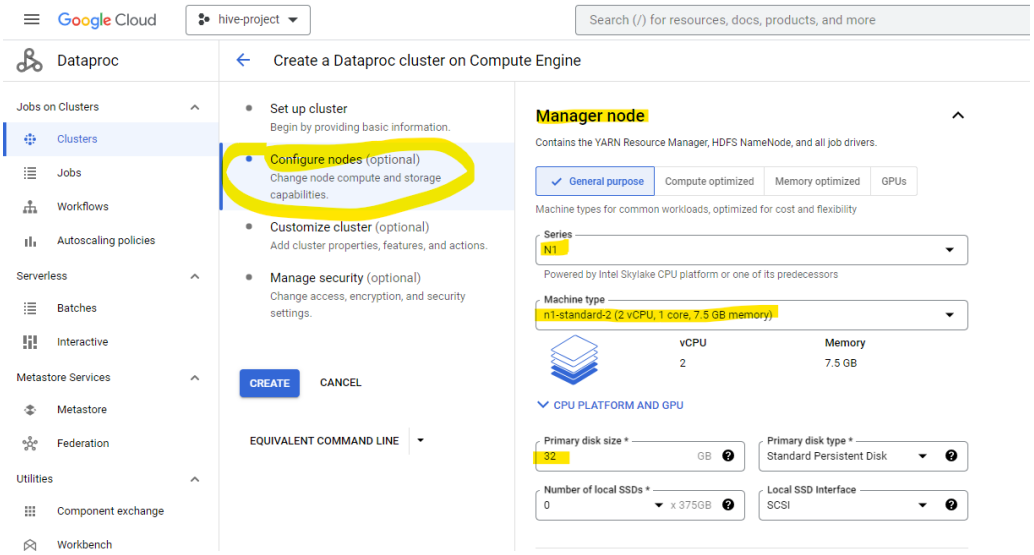


20. In **Components**, check the option **Enable component gateway**
In **optional components** check **Jupyter Notebook**

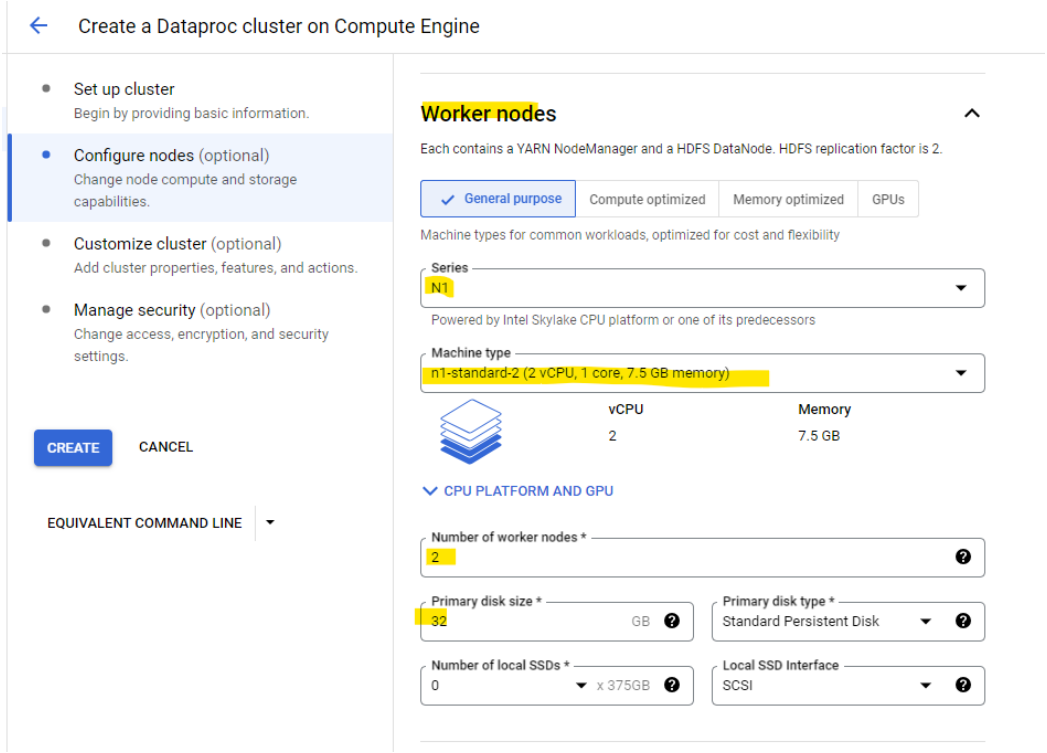


21. Now click on **Configure nodes**

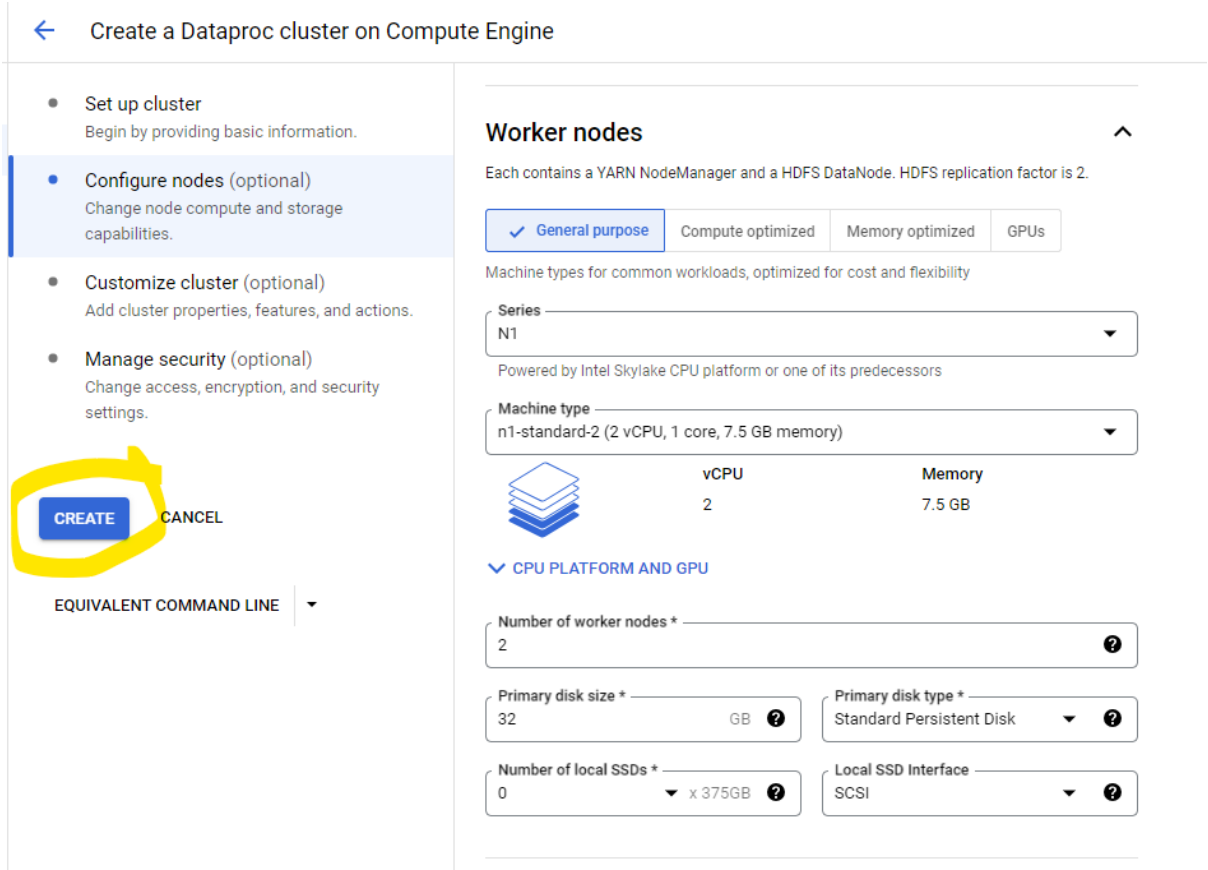
Under **Manager node**
select **Series** as **N1**
Machine type as **n1-standard-2**
Primary disk size as **32** GB



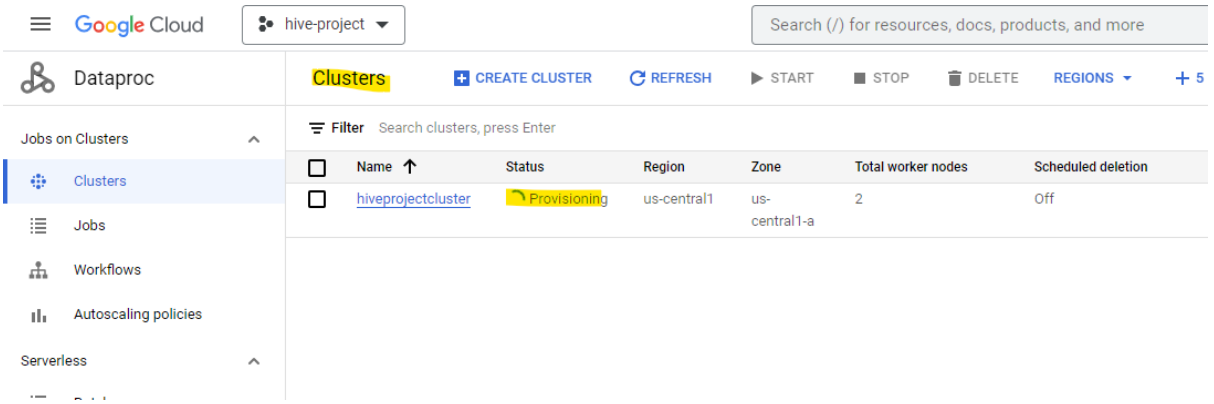
Now scroll down to Worker node
Under **Worker nodes**
select **Series** as **N1**
Machine type as **n1-standard-2**
Number of worker nodes as **2**
Primary disk size as **32** GB



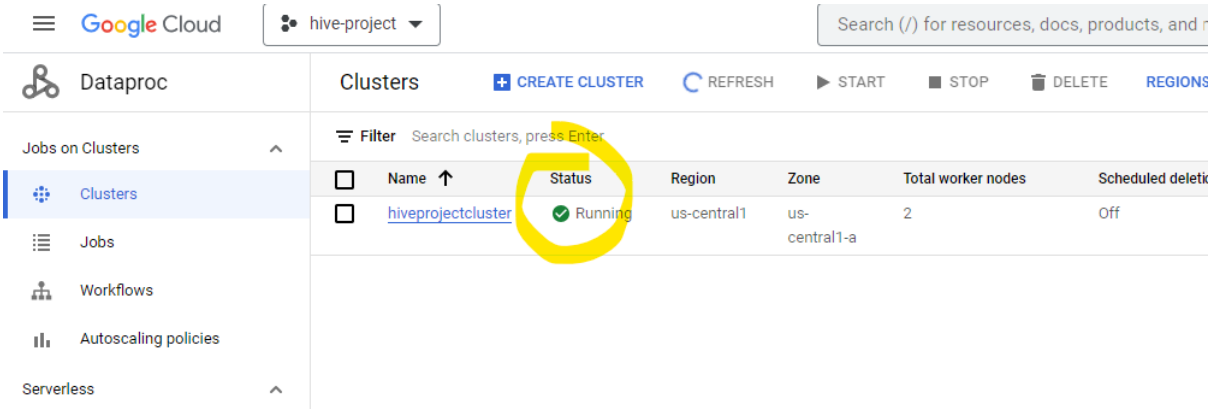
22. That's all, now click on **CREATE**



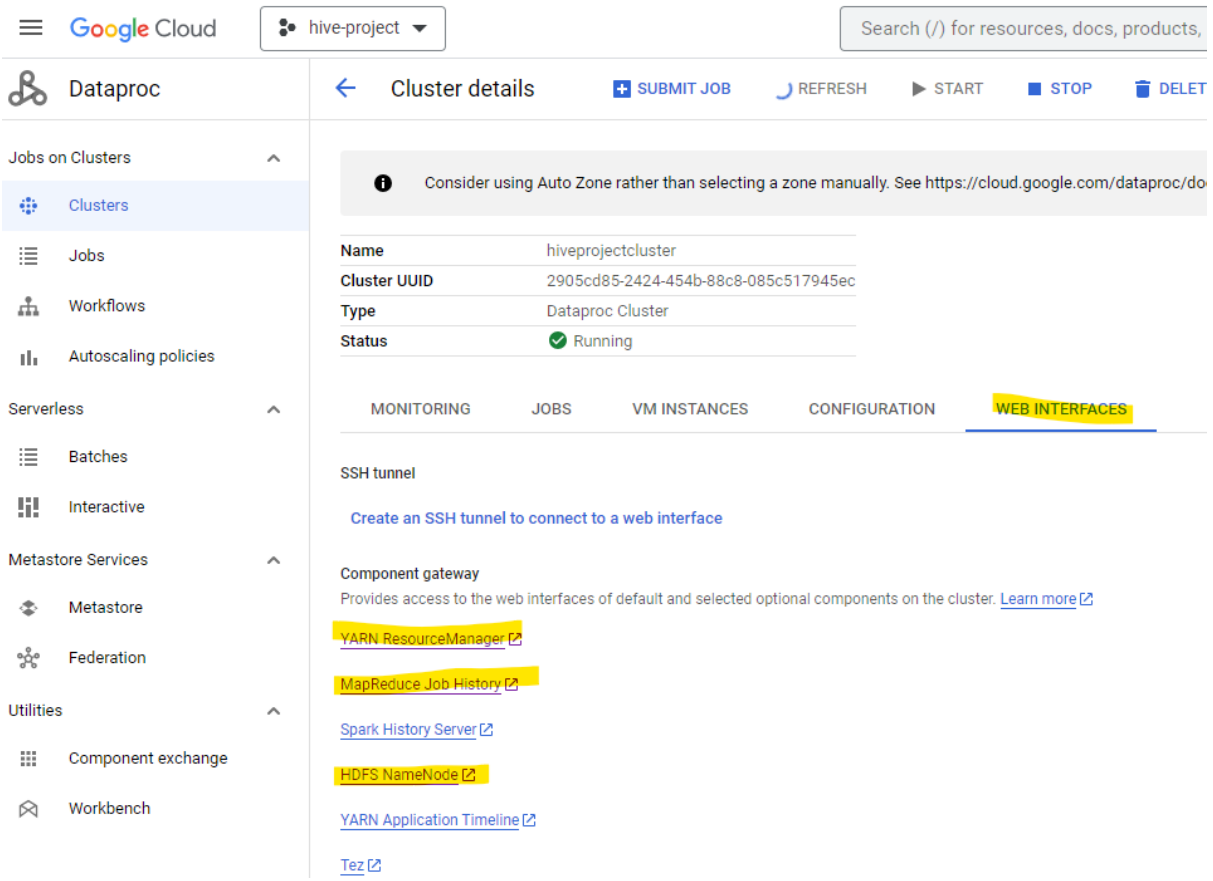
23. 2A new window will open as below and the cluster creation is in progress and you will see the status as provisioning.



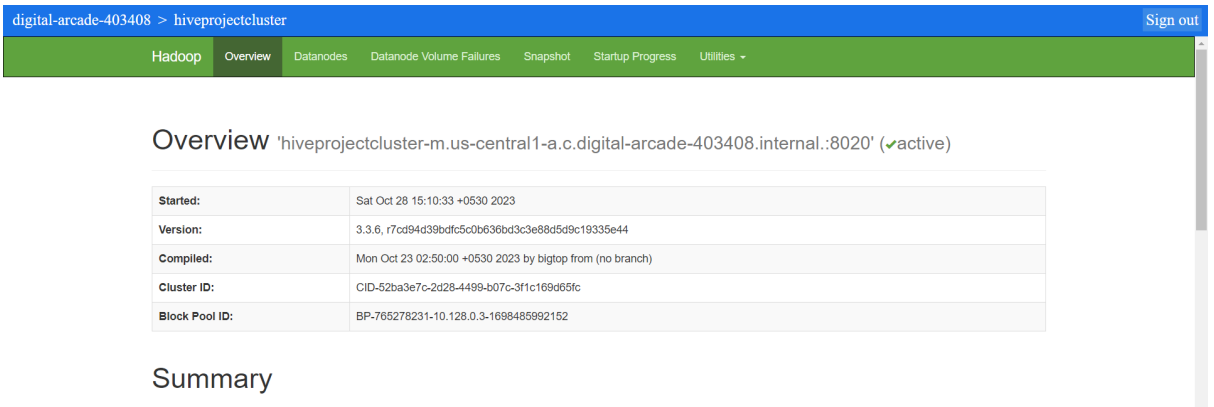
This will take some time around 5 mins for cluster to be up and Running.
Once the cluster is created you will see status changed to Running.



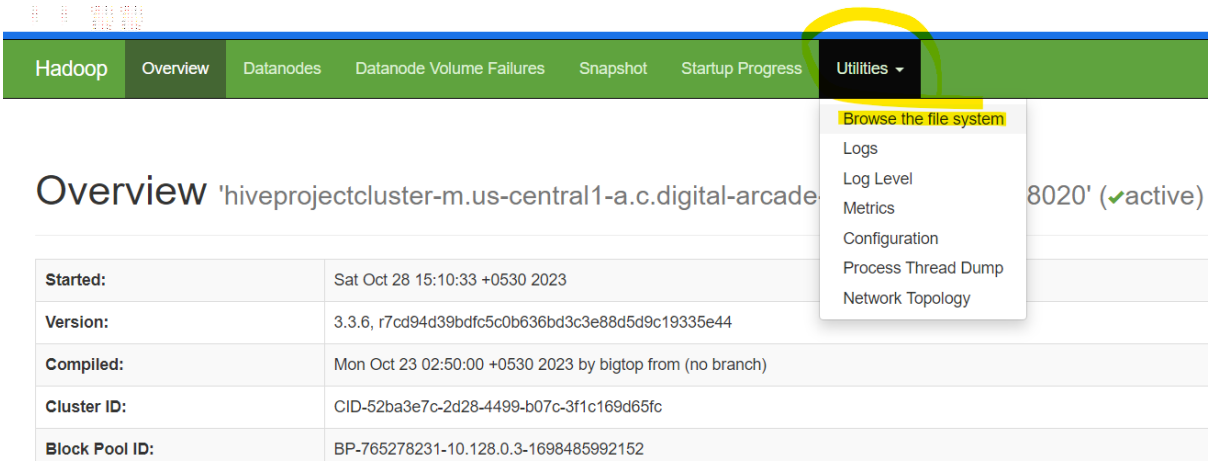
24. Now click on the cluster and go to **WEB INTERFACES**.
    You can see YARN, MapReduce and HDFS interfaces.
    You can open them and see.



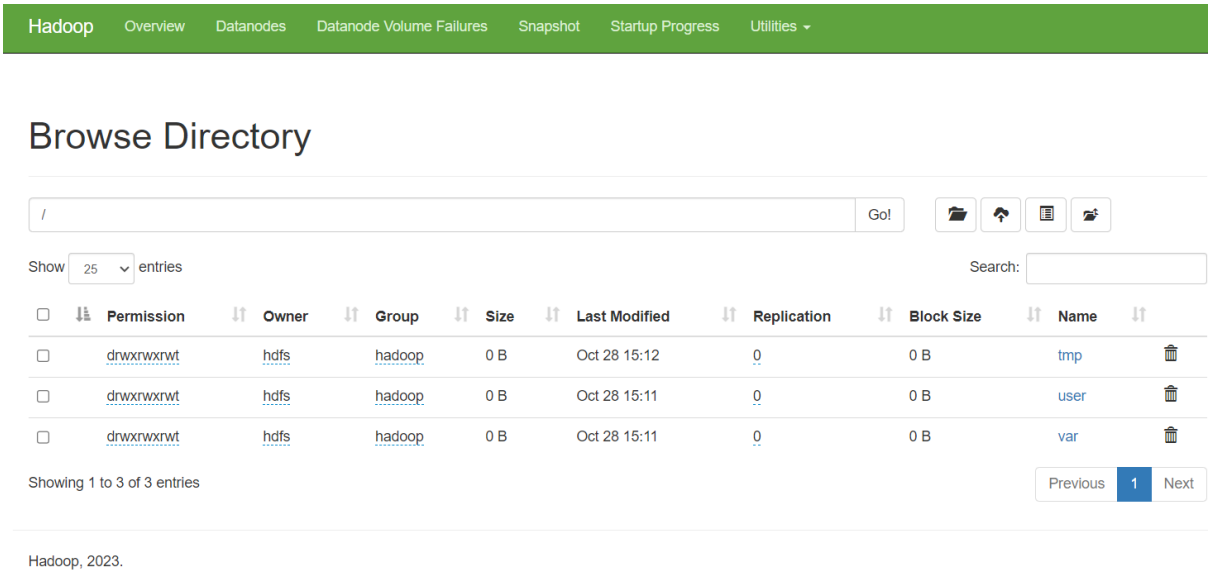
When you open HDFS, the screen will look as below:



25. To see the HDFS File system: Click on Utilities -> Browse the file system



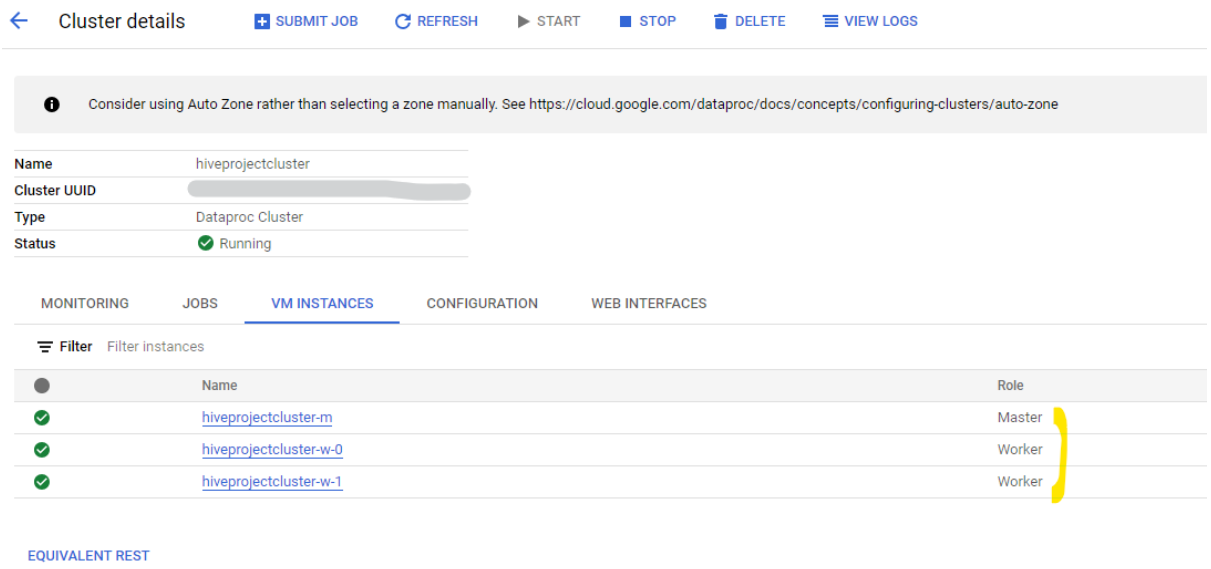
Below are the three you will see by default.



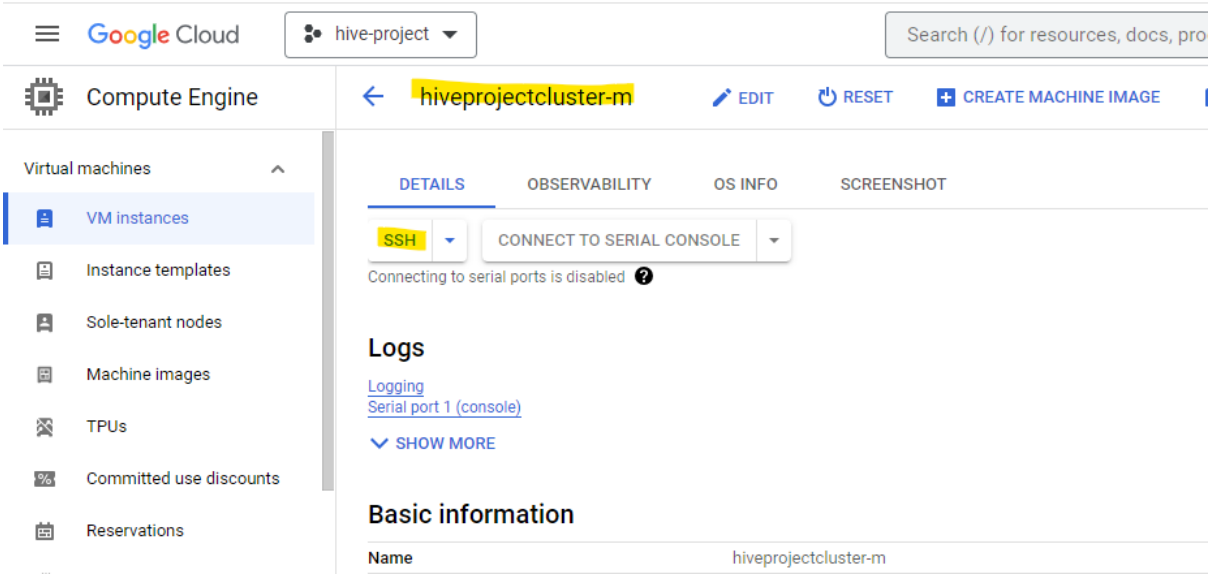
26. You can also check this from the terminal.

    In the Cluster details, click on VM Instances.

    You will see 1 Master node and 2 Worker nodes created.



Now click on Master node. Click SSH.



If you get any pop-ups, please Authorize.

New window SSH-in-browser will open. Here you can write your hdfs and hive commands.

SSH-in-browser UPLOAD FILE DOWNLOAD FILE

```
Linux hiveprojectcluster-m 5.10.0-26-cloud-amd64 #1 SMP Debian 5.10.197-1 (2023-09-29) x86_64

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.
kotagirikavyagcp@hiveprojectcluster-m:~$
```

To see the file in the terminal, use the below command

hdfs dfs -ls /

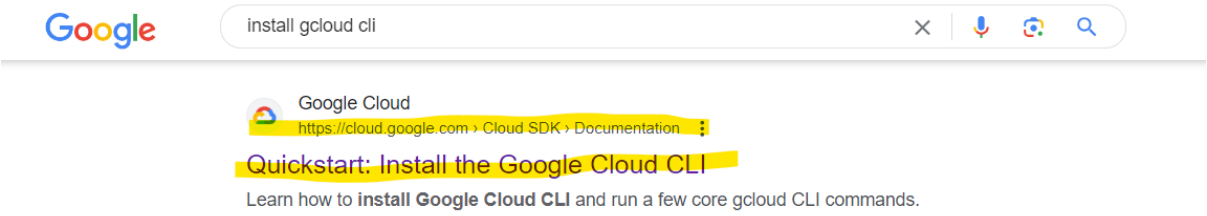
SSH-in-browser

```
Linux hiveprojectcluster-m 5.10.0-26-cloud-amd64 #1 SMP Debian 5.10.197-1 (2023-09-29) x86_64

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: Sat Oct 28 09:50:38 2023 from 35.235.240.1
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -ls /
Found 3 items
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:42 /tmp
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:41 /user
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:41 /var
kotagirikavyagcp@hiveprojectcluster-m:~$
```

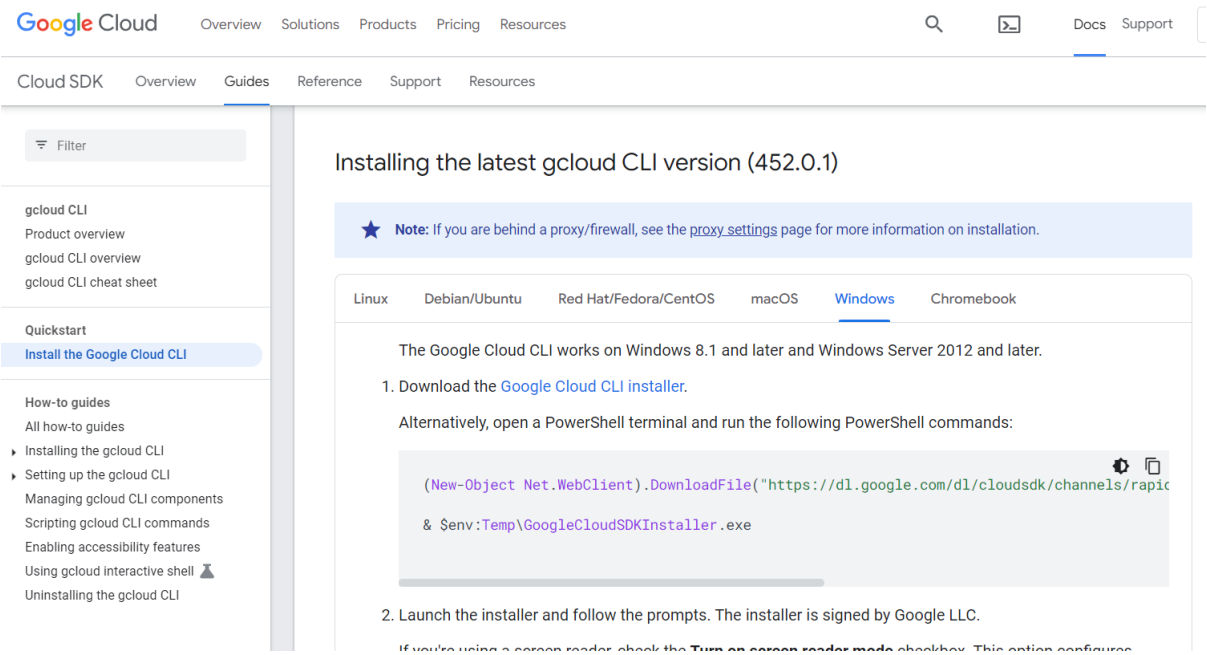
That’s all.

Now let's install GCLoud CLI.

27. First let's go to the installation documentation
28. Open the browser and search for install gcloud cli.
29. Open the below highlighted link.

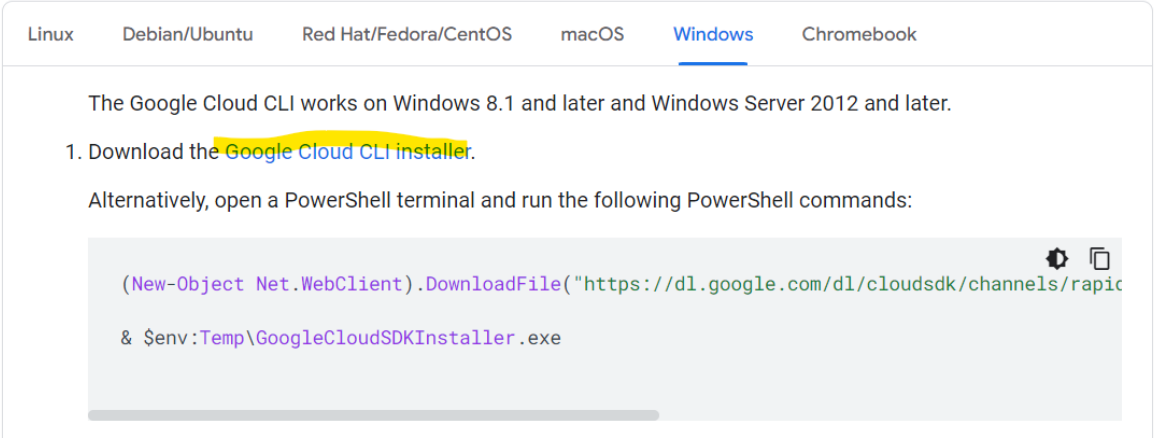


30. Steps are provided for different OS. Follow the steps based on your system OS.

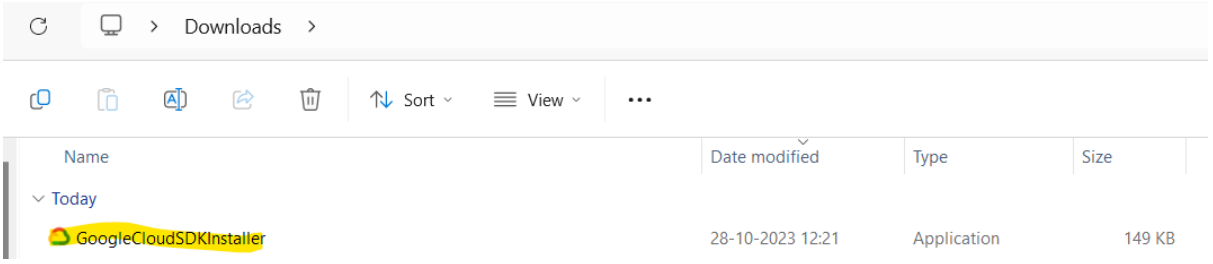


I will go through Windows Installation steps.

31. Click on the Google Cloud CLI installer.



32. An .exe file will be downloaded. Run the .exe file.



33. If you need screen reader mode, check the option else directly click on Next>

34.

Click I Agree

Select Install Type as Single user or All users.

If you want to change destination folder for this install, please change or leave it as is.

Click Next

Click Install

Once the installation is completed, below Terminal will be opened Welcome text.

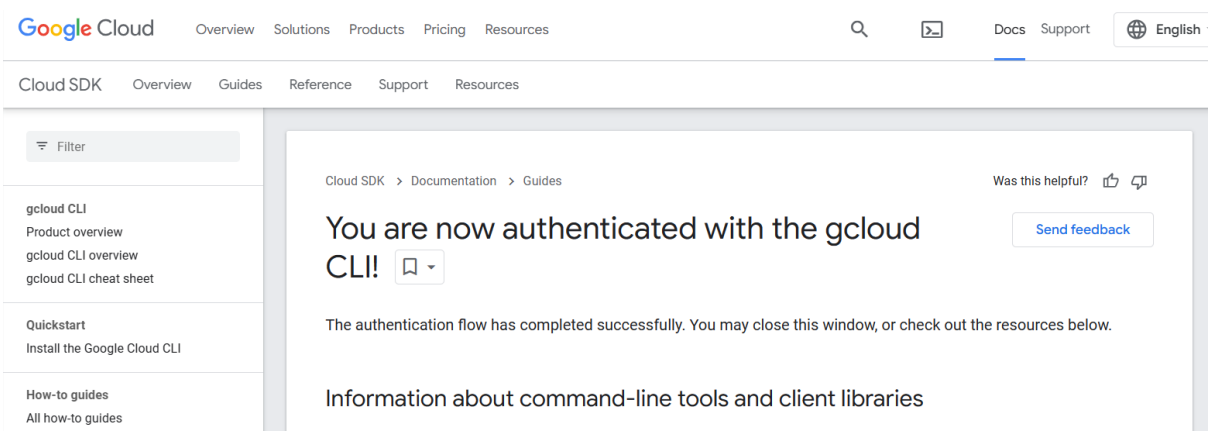
```
C:\WINDOWS\SYSTEM32\cmc

Welcome to the Google Cloud CLI! Run "gcloud -h" to get the list of available commands.
---
Welcome! This command will take you through the configuration of gcloud.

Your current configuration has been set to: [default]

You can skip diagnostics next time by using the following flag:
  gcloud init --skip-diagnostics

Network diagnostic detects and fixes local network connection issues.
Checking network connection...done.
Reachability Check passed.
Network diagnostic passed (1/1 checks passed).

You must log in to continue. Would you like to log in (Y/n)?  Y

Your browser has been opened to visit:

    https://accounts.google.com/o/oauth2/auth?response_type=code&client_id=32555940559.apps.goo
%2F&scope=openid+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fuserinfo.email+https%3A%2F%2Fwww.goo
pis.com%2Fauth%2Fappengine.admin+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fsqlservice.login+htt
www.googleapis.com%2Fauth%2Faccounts.reauth&state=rW67tyTyi28DQ4puKyTC3yFMlXSezA&access_type=of
Cug&code_challenge_method=S256
```

35. You must log in to continue. Would you like to log in (Y/n)?

Type **Y** and hit Enter.

Now your browser automatically opens for authentication, just provide the email you used while creating free Google Cloud Account.

Once your authorization is completed. You will see as below:



36. Now come back to Terminal.

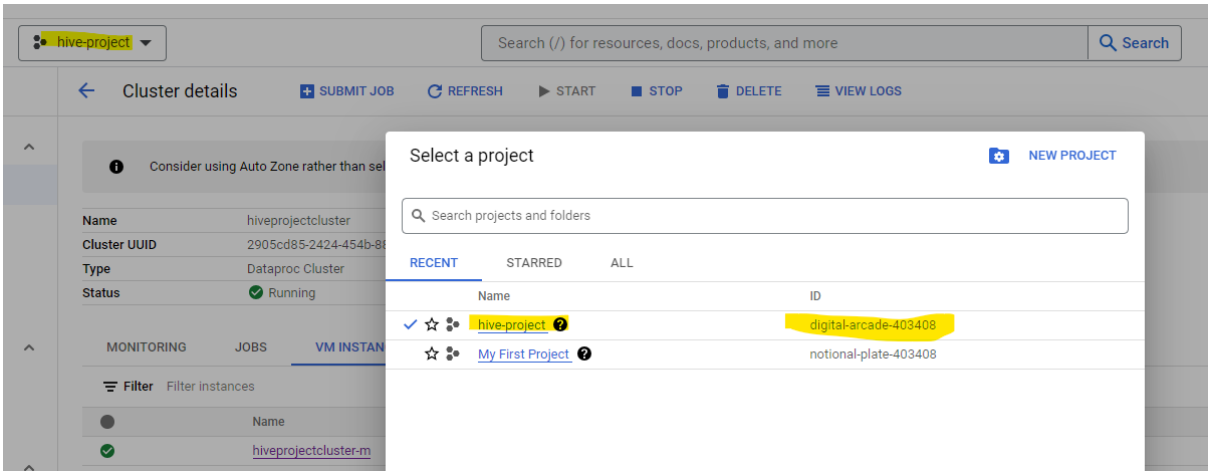
You will see logged in sucesfully with the account you provided.

```
Your browser has been opened to visit:

    https://accounts.google.com/o/oauth2/auth?response_type=code&client_id=32555940559.apps.googleuserc
%2F&scope=openid+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fuserinfo.email+https%3A%2F%2Fwww.googleapis.
pis.com%2Fauth%2Fappengine.admin+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fsqlservice.login+https%3A%2F
www.googleapis.com%2Fauth%2Faccounts.reauth&state=SJjnb8Ay2NrMIJ226nHWUgDLGAWvCG&access_type=offline&co
h9E&code_challenge_method=S256


You are now logged in as [kotagirikavyagcp@gmail.com].
Your current project is [sound-proposal-403405].  You can change this setting by running:
  $ gcloud config set project PROJECT_ID
```

By default other project have set up.

We need to change this project set up to the project we created.

You can find the Project ID on the UI top left drop down.



37. Now enter the below command in the gcloud cli terminal. Replace with your project ID in below command.

gcloud config set project **digital-arcade-403408**

```
You are now logged in as [kotagirikavyagcp@gmail.com].
Your current project is [sound-proposal-403405].  You can change this setting by running:
  $ gcloud config set project PROJECT_ID

C:\Users\sysga\Downloads\Reviews.csv>gcloud config set project digital-arcade-403408
Updated property [core/project].
```

Or you will get an option to pick the project from the list as below:

```
Pick cloud project to use:
 [1] avian-cable-403405
 [2] eternal-outlook-351617
 [3] kkproject-403405
 [4] sound-proposal-403405
 [5] Enter a project ID
 [6] Create a new project
Please enter numeric choice or text value (must exactly match list item):  5
```

38. Now Let's upload the same file into cluster master node and then to hdfs.

If you need you can download sample file with around 300 MB from my google drive link:

https://drive.google.com/file/d/10-zKUd05BLBK9ECyvUGGSTXLxHMU01qZ/view?usp=sharing

39. Now, in the terminal navigate to the path where we have this file downloaded. Using cd command.

    I have the file in my downloads so I am navigating to downloads folder using below command.

    cd C:\Users\sysga\Downloads\Reviews.csv

40. Now enter the below command that copies file from local to master node and hit Enter.

Replace with your gcloud account and master node name. You can find master node details in the SSH or on the UI.

gcloud compute scp Reviews.csv kotagirikavyagcp@hiveprojectcluster-m:/home/kotagirikavyagcp

```
C:\Users\sysga\Downloads\Reviews.csv>cd C:\Program Files (x86)\Google\Cloud SDK

C:\Program Files (x86)\Google\Cloud SDK>cd C:\Users\sysga\Downloads\Reviews.csv

C:\Users\sysga\Downloads\Reviews.csv>gcloud compute scp Reviews.csv kotagirikavyagcp@hiveprojectcluster-m:/home/kotagirikavyagcp
```

SSH-in-browser

```
Linux hiveprojectcluster-m 5.10.0-26-cloud-amd64 #1 SMP Debian 5.10.197-1 (2023-09-29) x86_64

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: Sat Oct 28 09:50:38 2023 from 35.235.240.1
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -ls /
Found 3 items
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:42 /tmp
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:41 /user
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:41 /var
kotagirikavyagcp@hiveprojectcluster-m:~$
```

You will see the file is uploading

```
C:\Users\sysga\Downloads\Reviews.csv>cd C:\Program Files (x86)\Google\Cloud SDK

C:\Program Files (x86)\Google\Cloud SDK>cd C:\Users\sysga\Downloads\Reviews.csv

C:\Users\sysga\Downloads\Reviews.csv>gcloud compute scp Reviews.csv kotagirikavyagcp@hiveprojectcluster-m:/home/kotagirikavyagcp
No zone specified. Using zone [us-central1-a] for instance: [hiveprojectcluster-m].
Updating project ssh metadata.../Updated [https://www.googleapis.com/compute/v1/projects/digital-arcade-403408].
Updating project ssh metadata...done.
Waiting for SSH key to propagate.
The server's host key is not cached. You have no guarantee
that the server is the computer you think it is.
The server's ssh-ed25519 key fingerprint is:
ssh-ed25519 255 SHA256:SuFc/LD7WNXbFGRsO/RUx/QNr93maBCAM8CIJhc+REA
If you trust this host, enter "y" to add the key to
PuTTY's cache and carry on connecting.
If you want to carry on connecting just once, without
adding the key to the cache, enter "n".
If you do not trust this host, press Return to abandon the
connection.
Reviews.csv               | 26916 kB | 2070.5 kB/s | ETA: 00:02:08 |   9%
```

File upload is completed.

```
Reviews.csv                | 293852 kB | 3093.2 kB/s | ETA: 00:00:00 | 100%

C:\Users\sysga\Downloads\Reviews.csv>
```

41. Now let's upload this file from master node home to new folder in hdfs.

This can be done using SSH.

Let's first create directory/folder using below command:

hdfs dfs -mkdir /input_file

SSH-in-browser

```
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -ls /
Found 3 items
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:42 /tmp
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:41 /user
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:41 /var
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -mkdir /input_file
kotagirikavyagcp@hiveprojectcluster-m:~$
```

Let’s use below command to check if directory created.

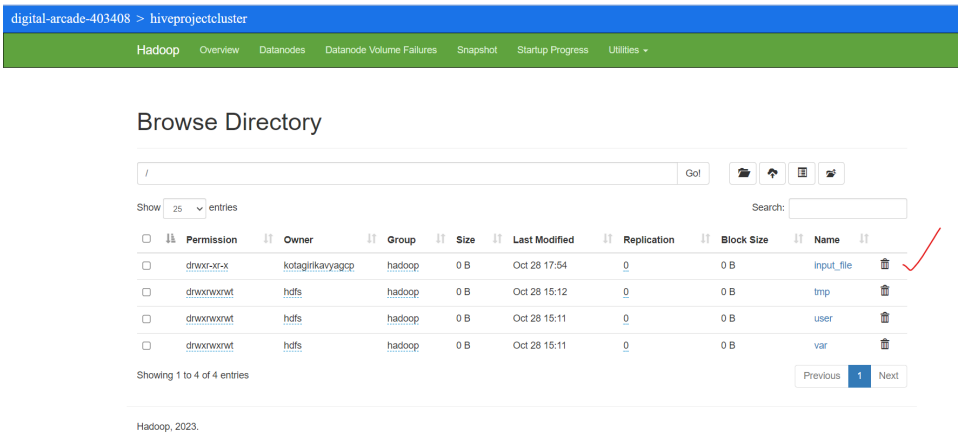
hdfs dfs -ls /

SSH-in-browser

```
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -ls /
Found 3 items
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:42 /tmp
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:41 /user
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:41 /var
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -mkdir /input_file
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -ls /
Found 4 items
drwxr-xr-x   - kotagirikavyagcp hadoop          0 2023-10-28 12:24 /input_file
drwxrwxrwt   - hdfs             hadoop          0 2023-10-28 09:42 /tmp
drwxrwxrwt   - hdfs             hadoop          0 2023-10-28 09:41 /user
drwxrwxrwt   - hdfs             hadoop          0 2023-10-28 09:41 /var
kotagirikavyagcp@hiveprojectcluster-m:~$
```

Let’s also check from UI

Refresh the page and you can see the newly created directory.



42. Now Let’s move the file from master node home to this new directl=ory using below command.

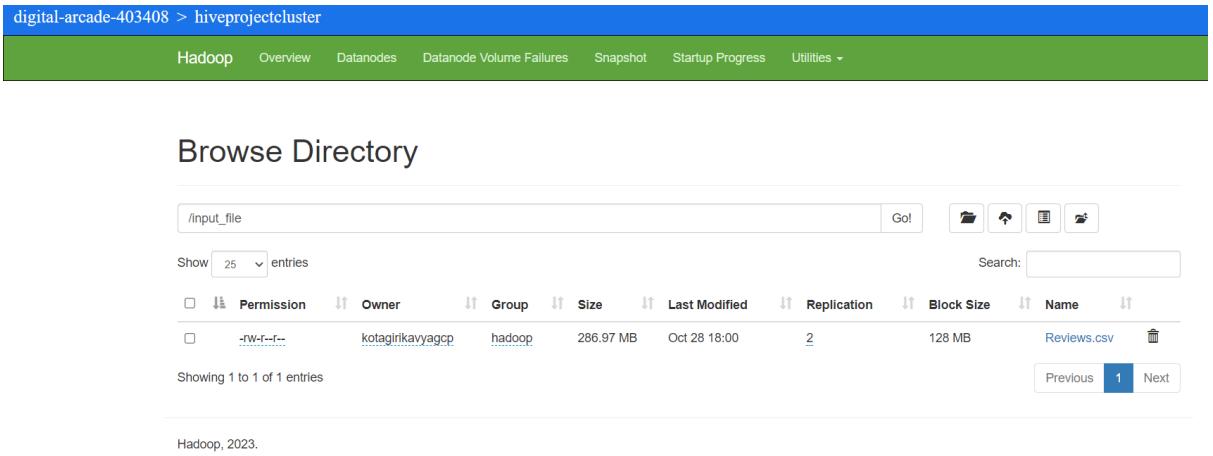
hdfs dfs -put /home/kotagirikavyagcp/Reviews.csv /input_file

```
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -put /home/kotagirikavyagcp/Reviews.csv /input_file
kotagirikavyagcp@hiveprojectcluster-m:~$
```

Let’s refresh the UI and see if the file is uploaded in directory.

Go to directory.

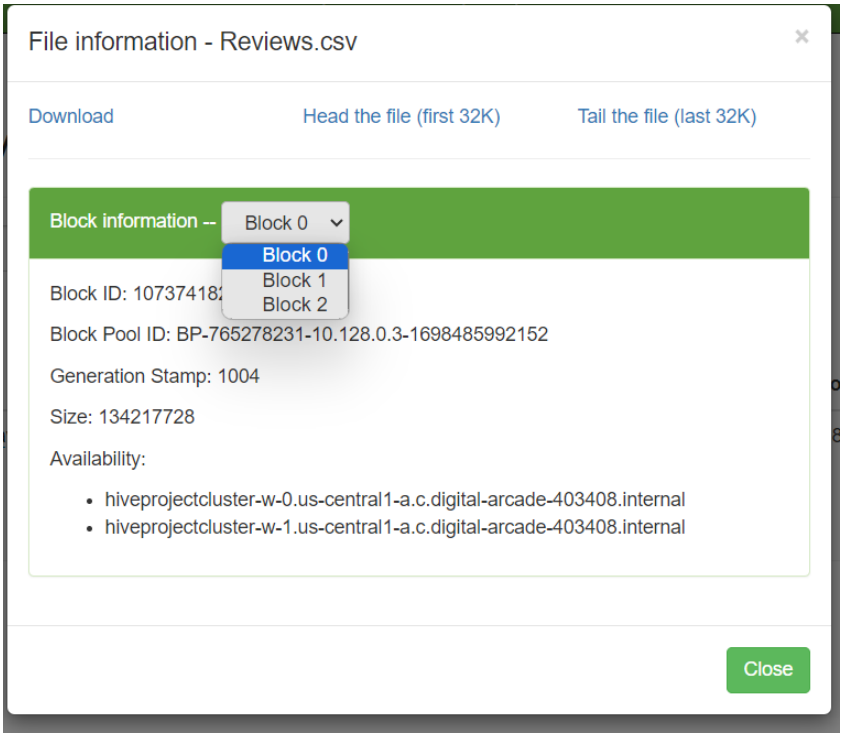
Yes, we can see the file uploaded.



43. You will see block size as 128 MB.

Let’s see how many blocks created.

Click on reviews.csv link and click on the top dropdown.



It created 3 blocks as the file size is 286 MB and each block size is 128 MB. So, it created 3 blocks.

You can upload the file using Upload File option in SSH, to directly upload file from local.

That’s all we have successfully uploaded the file form our local to hdfs.

44. To execute hive commands in SSH

Just type hive and hit enter.

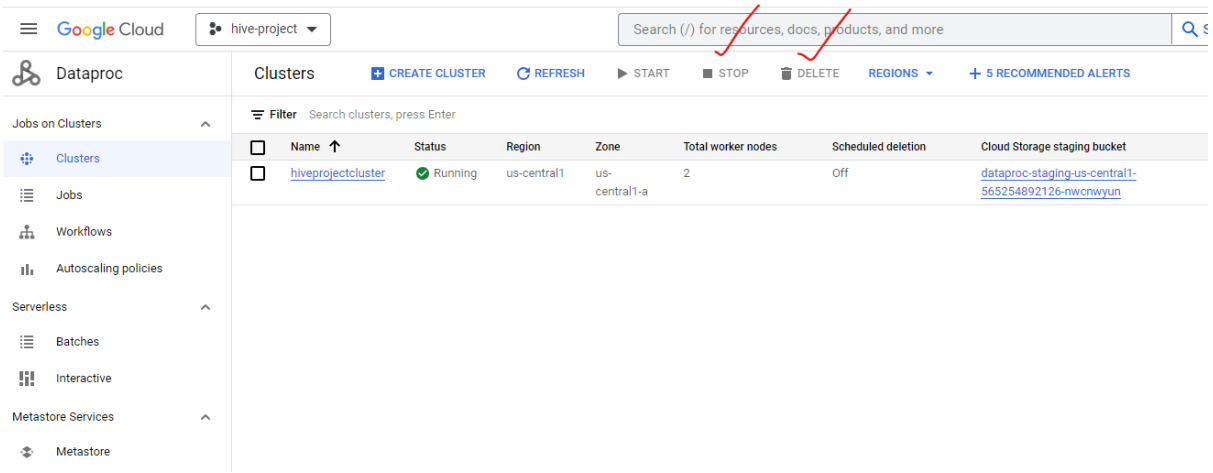
```
SSH-in-browser                                                                    UPLOAD FILE   DOWN
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -ls /
Found 3 items
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:42 /tmp
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:41 /user
drwxrwxrwt   - hdfs hadoop          0 2023-10-28 09:41 /var
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -mkdir /input_file
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -ls /
Found 4 items
drwxr-xr-x   - kotagirikavyagcp hadoop          0 2023-10-28 12:24 /input_file
drwxrwxrwt   - hdfs             hadoop          0 2023-10-28 09:42 /tmp
drwxrwxrwt   - hdfs             hadoop          0 2023-10-28 09:41 /user
drwxrwxrwt   - hdfs             hadoop          0 2023-10-28 09:41 /var
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -put /home/kotagirikavyagcp/Reviews.csv /input_files
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -put /home/kotagirikavyagcp/Reviews.csv /input_file
kotagirikavyagcp@hiveprojectcluster-m:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/tez/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/tez/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]
Hive Session ID = cbc3145c-c50e-4508-a76c-dcc6db2de01a

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.hive.common.StringInternUtils (file:/usr/lib/hive/lib/hive-common-3.1.3.jar) to field java.net.URI.string
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.hive.common.StringInternUtils
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Hive Session ID = 767ee821-924f-4ab2-b829-85c8c8d40a97
hive>
```

You can start writing Hive queries now.

45.

Once you are done with your work, make you stop/delete the cluster. If you keep it running the credits will be used. So, make sure, whenever you are not using the cluster just stop/delete it.

You can do so by going to cluster screen select the cluster and click on Stop/delete.



Once the cluster is stopped, you will see status as Stopped.