

Set up Google Cloud

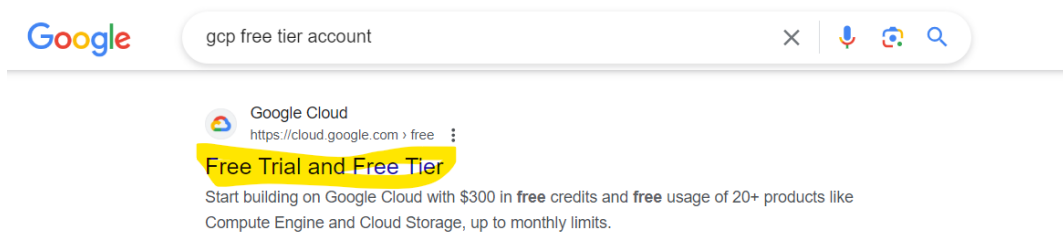
Create Hadoop Cluster

Set up GCloud CLI

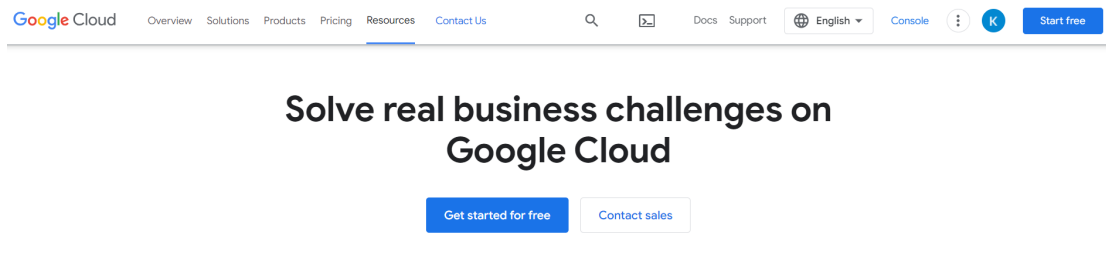
Upload sample file to HDFS

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**

 Try Google Cloud for free

Step 1 of 2 Account Information

 **Kavya Kotagiri**
kotagirikavyagcp@gmail.com [SWITCH ACCOUNT](#)

Country

By using this application, you agree to the [Google Cloud Platform](#), [Supplemental Free Trial](#), and [any applicable services and APIs Terms of Service](#).

[AGREE & CONTINUE](#)

Access to all Google Cloud products

Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.


\$300 credit for free

Put Google Cloud to work with \$300 in credit to spend over the next 90 days.

No autocharge after free trial ends



We ask you for your credit card to make sure you are not a robot. If you use a credit or debit card, you won't be charged unless you manually activate your full account.

5. Select Account Type as **Individual**

 Try Google Cloud for free

Step 2 of 2 Payment Information Verification

Your payment information helps us reduce fraud and abuse. **If using a credit or debit card, you won't be charged until you manually activate your account.**

 **Account type** 



Individual

Only Business accounts can have multiple users. You cannot change the account type after signing up. In some countries, this selection affects your tax options. If you choose Individual as your account type, you agree that use of your account is for your trade, business, craft, or profession. [Learn more](#)

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**.


Payment method

 Add credit or debit card 

#


Card number






MM


YY



Cardholder name




Address line 1





Address line 2

City





Postal code






State



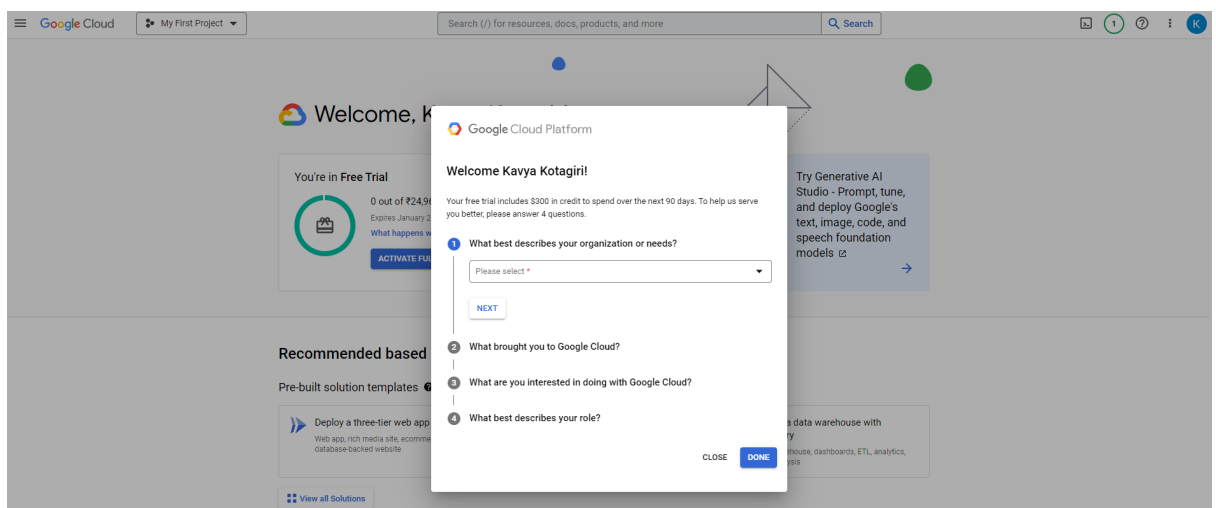


 Reserve Bank of India requires that cards support automatic payments according to RBI regulations. If your card doesn't support automatic payments, you'll need to make manual payments or use a different card. We'll check your card in the next step. [Learn more](#)

START FREE



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



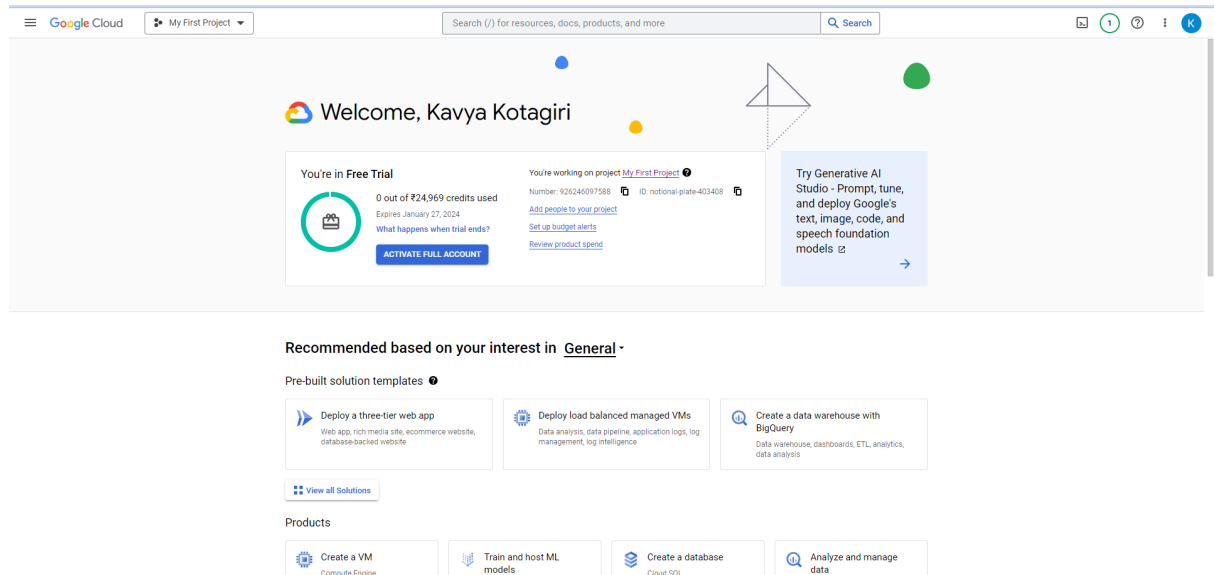
The screenshot shows the Google Cloud Platform welcome page for a user named Kavya Kotagiri. A survey overlay is displayed in the center, titled "Welcome Kavya Kotagiri!". The survey consists of four questions:

1. What best describes your organization or needs? (Please select *)
2. What brought you to Google Cloud?
3. What are you interested in doing with Google Cloud?
4. What best describes your role?

The survey has "NEXT" and "DONE" buttons. The background shows the Google Cloud welcome page with a "You're in Free Trial" banner and various service recommendations.

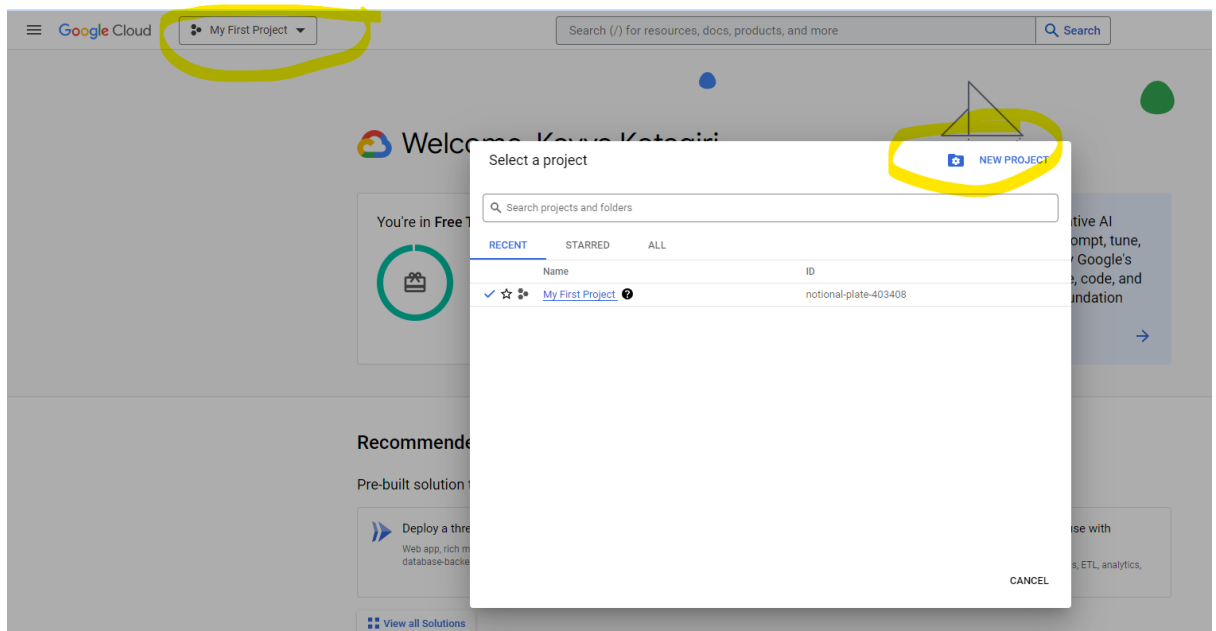
You can either provide the details or just click CLOSE.

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**

Google Cloud

Search (/) for resource

New Project

You have 11 projects remaining in your quota. Request an increase or delete projects. [Learn more](#)

[MANAGE QUOTAS](#)

Project name *
hive-project

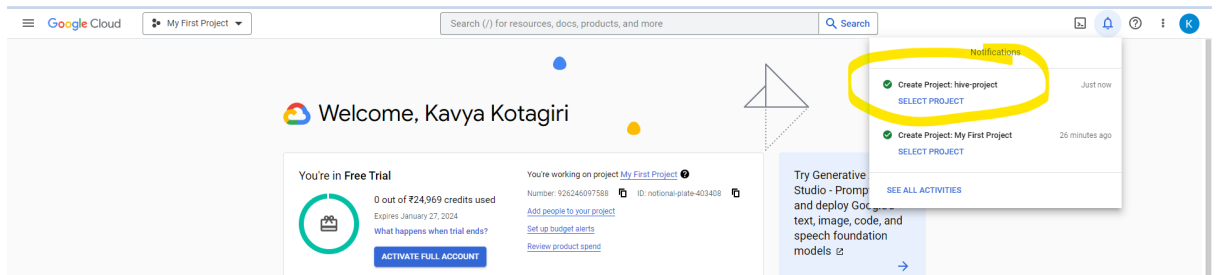
Project ID: digital-arcade-403408. It cannot be changed later. [EDIT](#)

Location *
No organization [BROWSE](#)

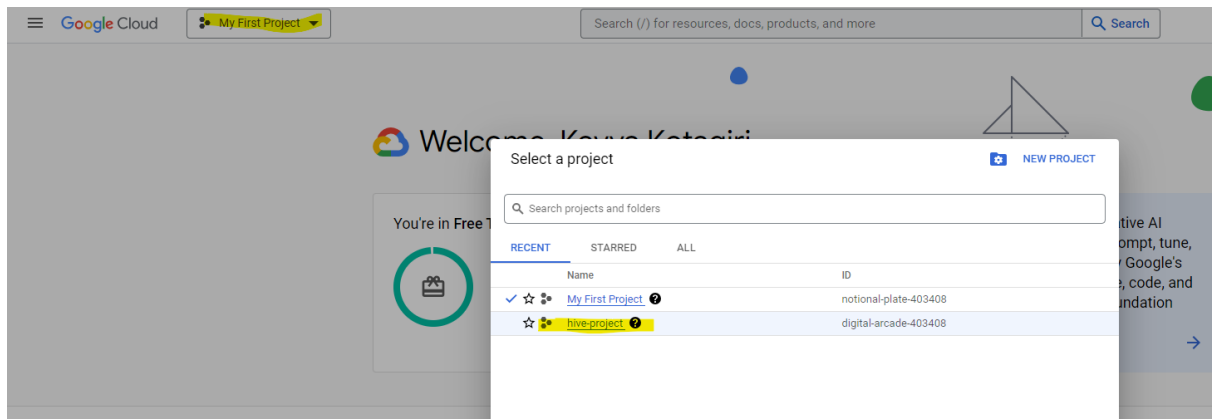
Parent organization or folder

CREATE CANCEL

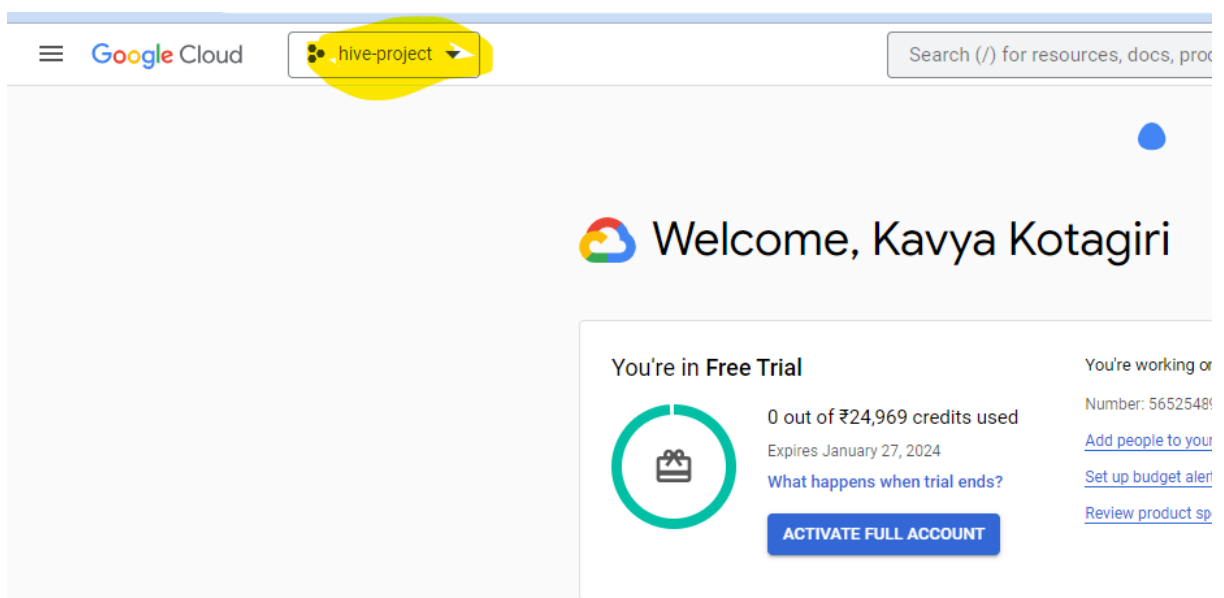
11. You will be redirected to home page and a notification with green tick will be displayed once the new project is created.



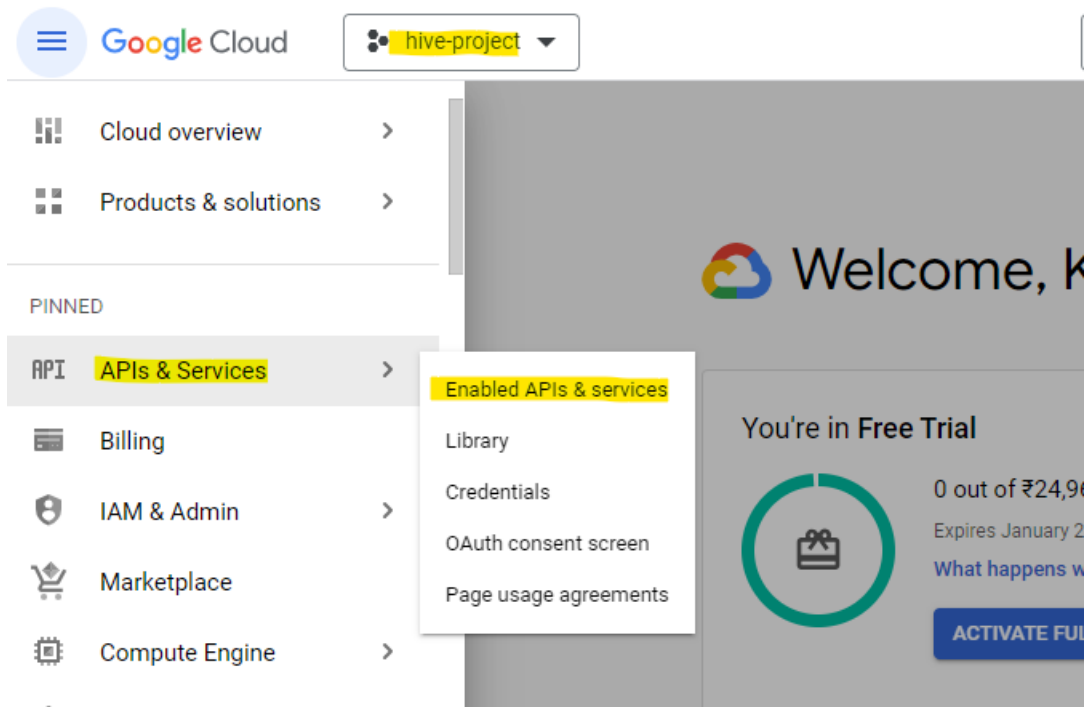
12. Now click on the My First Project drop down and select the project you selected.



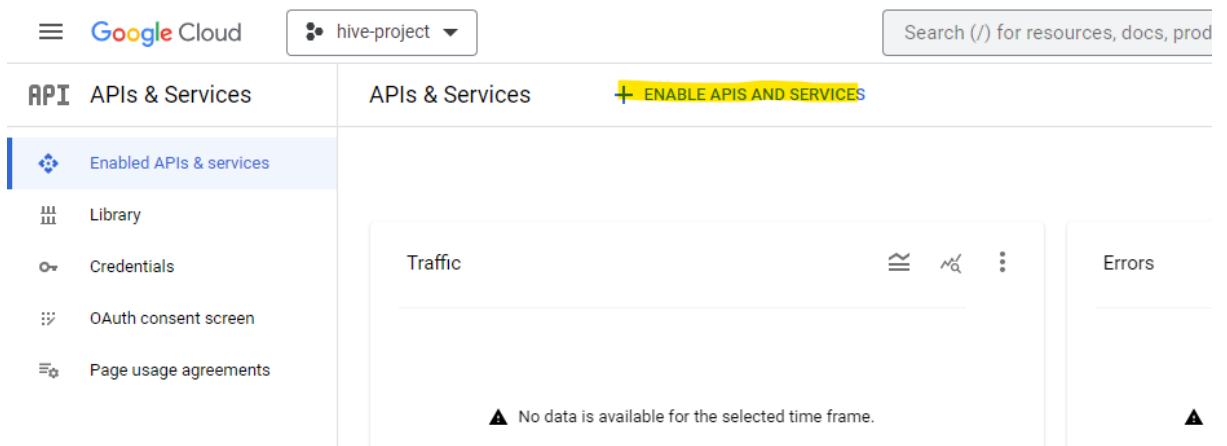
13. Once the project is selected it will start showing the project selected on the top left of your screen.



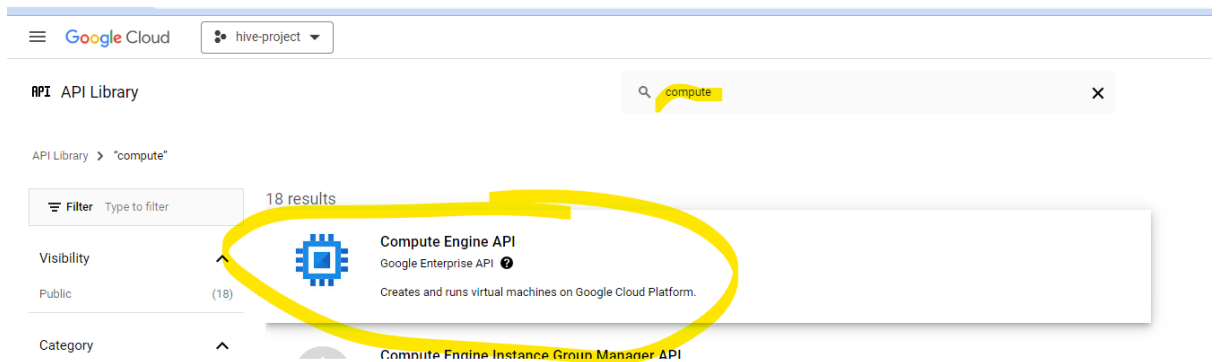
14. Now we need to enable 3 APIs.
Click on again 3 lines and now select **APIs & Services**.



Click on **+ ENABLE APIS AND SERVICES**




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



Google Cloud

hive-project

[Product details](#)



Compute Engine API

[Google Enterprise API](#)

Compute Engine API

ENABLE

TRY THIS API

OVERVIEW

DOCUMENTATION

SUPPORT

RELATED PRODUCTS

Overview

Creates and runs virtual machines on Google Cloud Platform.

Additional details

Type: [SaaS & APIs](#)

Once its enabled, you will see status as Enabled.

Google Cloud

hive-project

Search (/) for resources, docs, products, and more

API APIs & Services

Enabled APIs & services

Library


Credentials

OAuth consent screen

Page usage agreements

[API/Service Details](#)

DISABLE API



Compute Engine API

Creates and runs virtual machines on Google Cloud Platform.

By Google Enterprise API

Service name	Type
compute.googleapis.com	Public API

Status

Enabled

METRICS

QUOTAS

CREDENTIALS

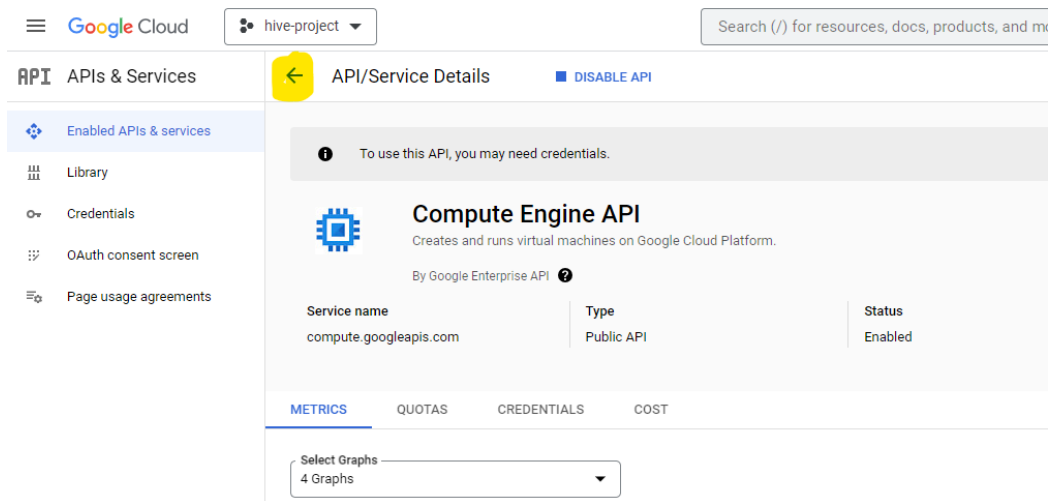
COST

Select Graphs

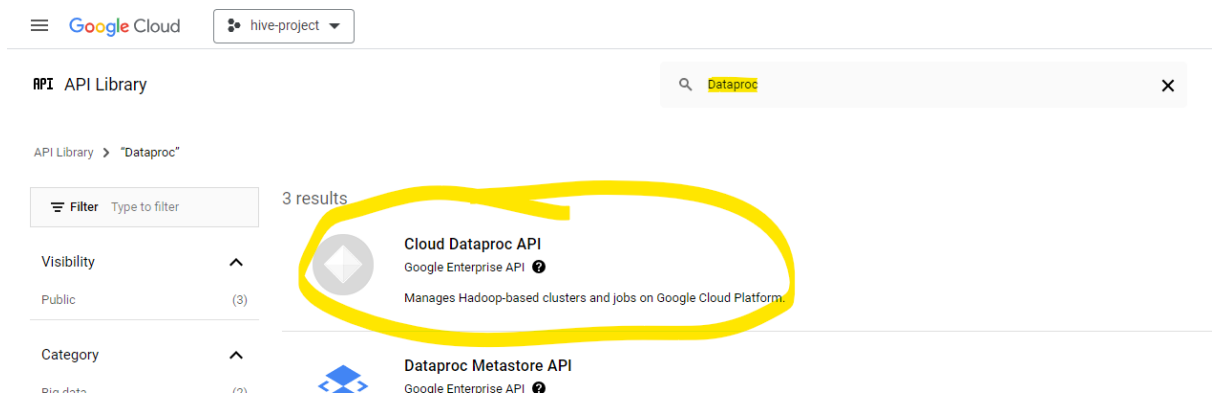
4 Graphs

Now let's enable other 2 APIs as well.

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

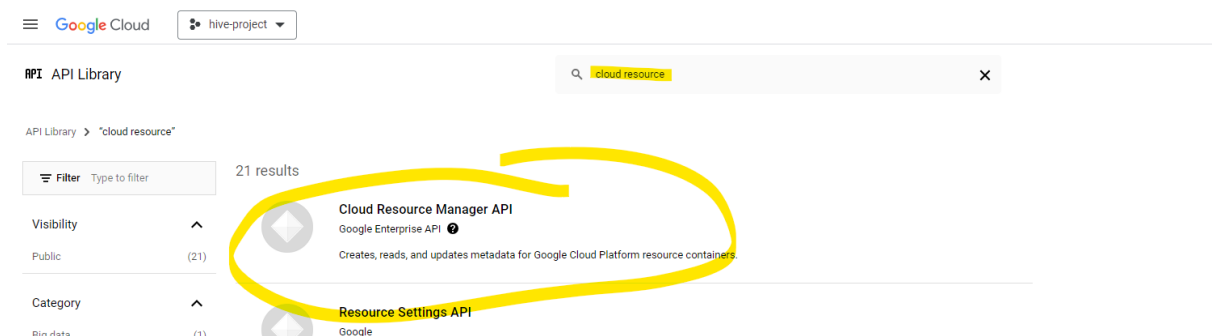


Search for Dataproc and select **Cloud Dataproc API**



Once its enabled, go back.

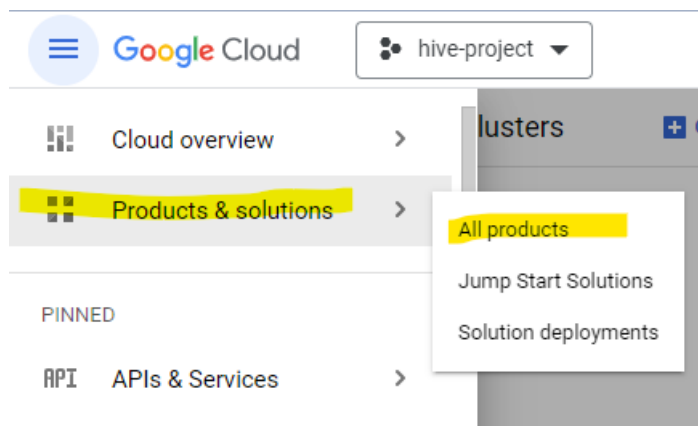
Search for cloud resource and select **Cloud Resource Manager API**.



We are now done with enabled the required APIs.

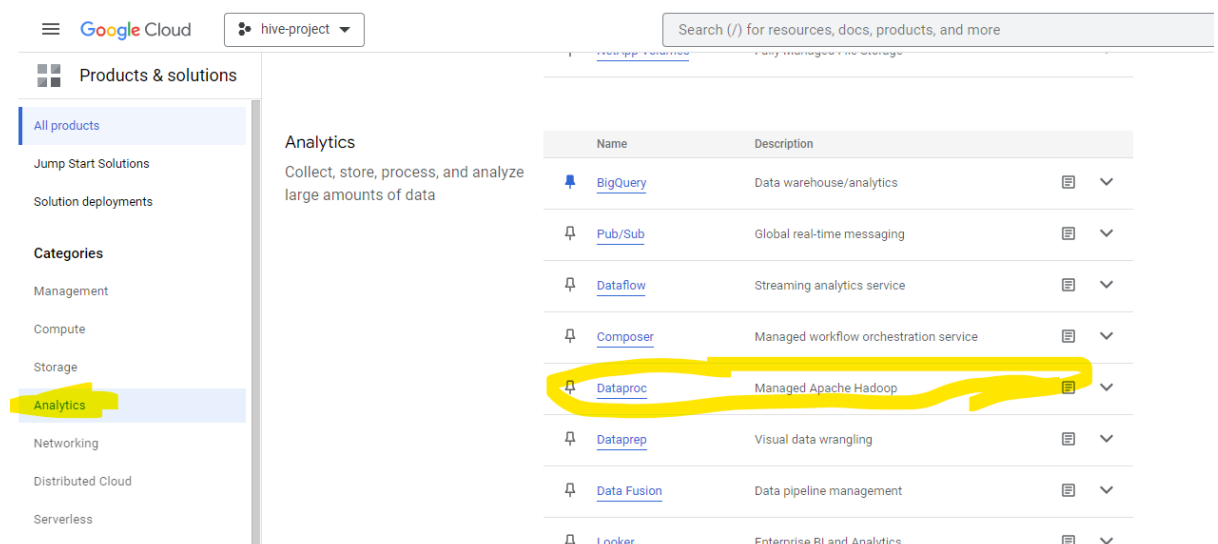
15. Now let's start creating Hadoop cluster.

Click on 3 lines on top left, click Products & solutions then All products



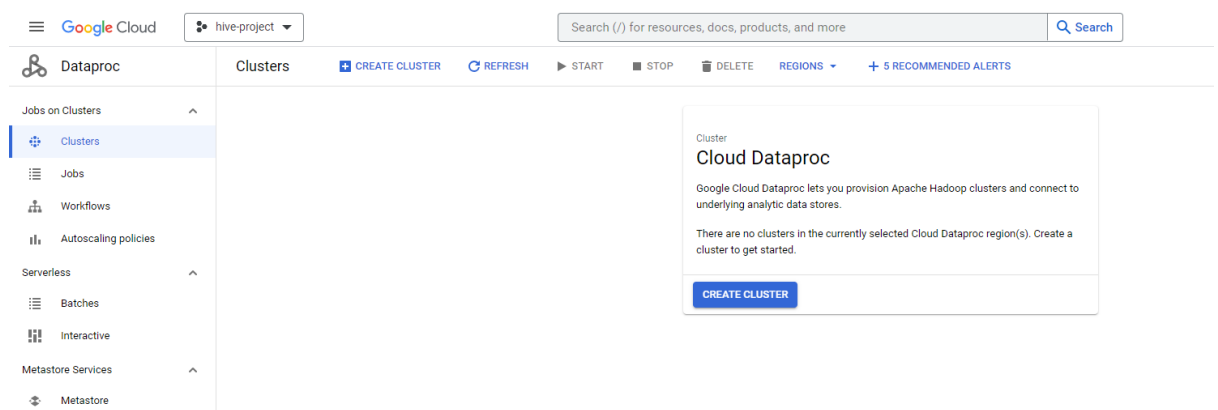
16. Scroll down to Analytics Category and you will find **Dataproc**.

Click on Dataproc.



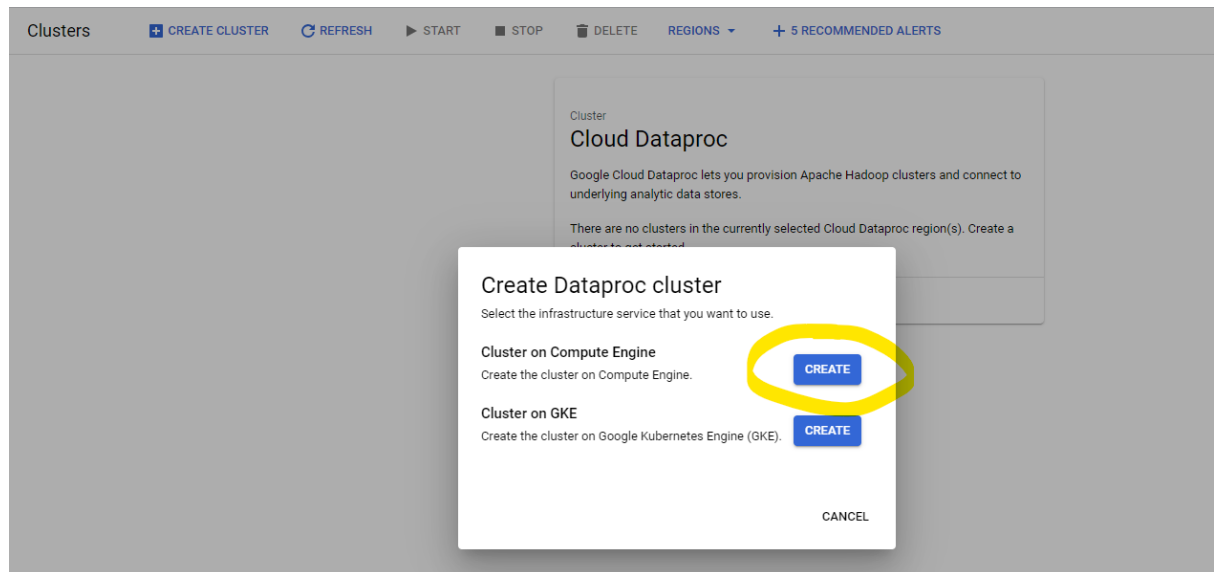
17. You will see the below screen.

If you face any error, just refresh the page or trying re-logging in.

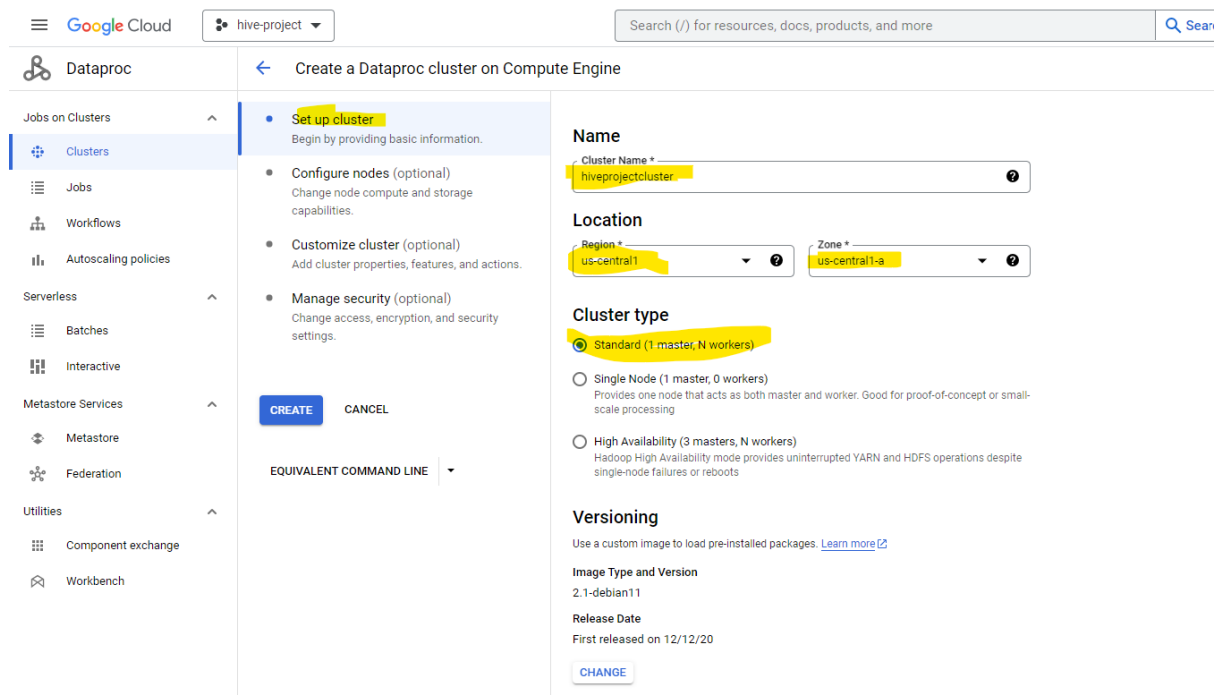


18. Now click on **CREATE CLUSTER**

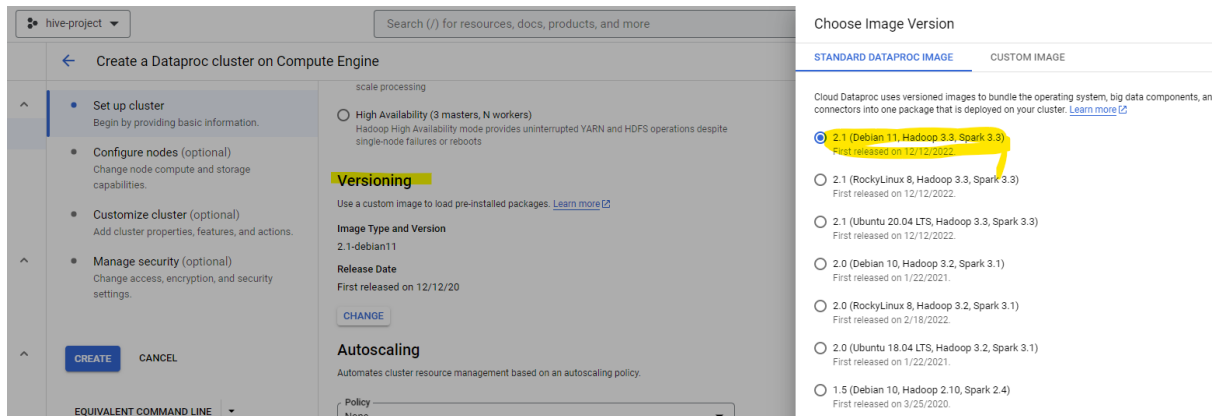
Click **CREATE** option of **Cluster on Compute Engine**



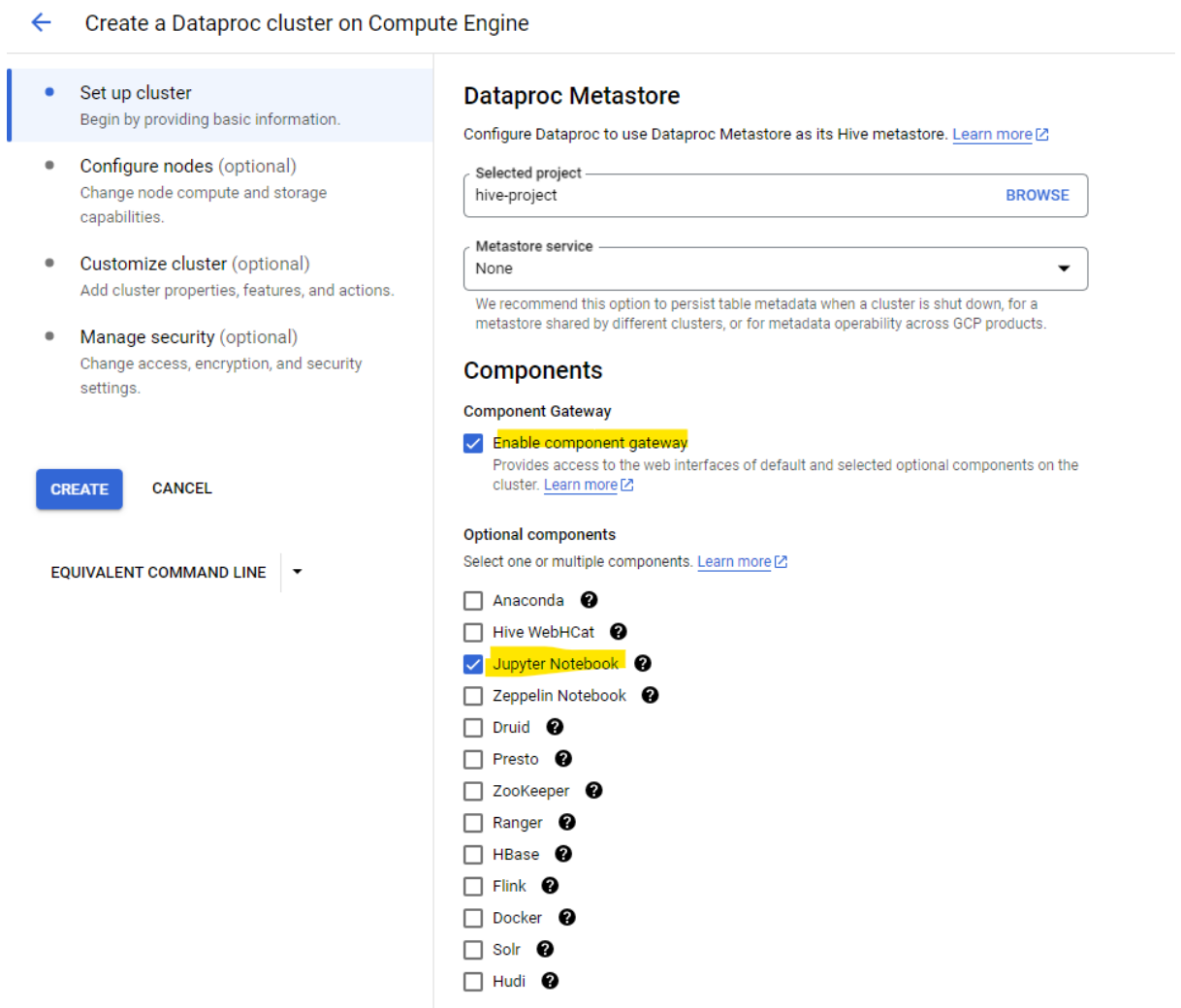
19. Now in **Set up cluster**, give Cluster Name, Region and select zone as us-central1-a and Cluster type as Standard(1 master, N workers).



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**

The screenshot shows the Google Cloud console interface for creating a Dataproc cluster. The left sidebar contains navigation links for Jobs on Clusters, Clusters, Jobs, Workflows, Autoscaling policies, Serverless, Batches, Interactive, Metastore Services, Metastore, Federation, Utilities, Component exchange, and Workbench. The main content area is titled 'Create a Dataproc cluster on Compute Engine'. It features a progress bar with four steps: 'Set up cluster', 'Configure nodes (optional)', 'Customize cluster (optional)', and 'Manage security (optional)'. The 'Configure nodes (optional)' step is highlighted with a yellow circle. Below the progress bar are 'CREATE' and 'CANCEL' buttons, and a section for 'EQUIVALENT COMMAND LINE'. The 'Manager node' configuration is shown on the right, with a description: 'Contains the YARN Resource Manager, HDFS NameNode, and all job drivers.' The configuration includes a 'General purpose' tab, a 'Series' dropdown set to 'N1', a 'Machine type' dropdown set to 'n1-standard-2 (2 vCPU, 1 core, 7.5 GB memory)', and a 'Primary disk size' dropdown set to '32 GB'. Other options include 'Primary disk type' (Standard Persistent Disk), 'Number of local SSDs' (0), and 'Local SSD Interface' (SCSI).

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**

← Create a Dataproc cluster on Compute Engine

- Set up cluster
Begin by providing basic information.
- Configure nodes (optional)**
Change node compute and storage capabilities.
- Customize cluster (optional)
Add cluster properties, features, and actions.
- Manage security (optional)
Change access, encryption, and security settings.

CREATE CANCEL

EQUIVALENT COMMAND LINE ▾

Worker nodes

Each contains a YARN NodeManager and a HDFS DataNode. HDFS replication factor is 2.

☒ General purpose ☐ Compute optimized ☐ Memory optimized ☐ GPUs

Machine types for common workloads, optimized for cost and flexibility

Series
N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type
n1-standard-2 (2 vCPU, 1 core, 7.5 GB memory)

	vCPU	Memory
	2	7.5 GB

✓ CPU PLATFORM AND GPU

Number of worker nodes * 2 ?

Primary disk size * 32 GB ?

Primary disk type * Standard Persistent Disk ?

Number of local SSDs * 0 x 375GB ?

Local SSD Interface SCSI ?

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

← Create a Dataproc cluster on Compute Engine

- Set up cluster
Begin by providing basic information.
- Configure nodes (optional)**
Change node compute and storage capabilities.
- Customize cluster (optional)
Add cluster properties, features, and actions.
- Manage security (optional)
Change access, encryption, and security settings.

CREATE CANCEL

EQUIVALENT COMMAND LINE ▾

Worker nodes

Each contains a YARN NodeManager and a HDFS DataNode. HDFS replication factor is 2.

☒ General purpose ☐ Compute optimized ☐ Memory optimized ☐ GPUs

Machine types for common workloads, optimized for cost and flexibility

Series
N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type
n1-standard-2 (2 vCPU, 1 core, 7.5 GB memory)

	vCPU	Memory
	2	7.5 GB

✓ CPU PLATFORM AND GPU

Number of worker nodes * 2 ?

Primary disk size * 32 GB ?

Primary disk type * Standard Persistent Disk ?

Number of local SSDs * 0 x 375GB ?

Local SSD Interface SCSI ?

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

Google Cloud hive-project Search (/) for resources, docs, products, and more

Dataproc **Clusters** [+ CREATE CLUSTER](#) [REFRESH](#) [START](#) [STOP](#) [DELETE](#) [REGIONS](#) [+ 5](#)

Jobs on Clusters Filter Search clusters, press Enter

<input type="checkbox"/>	Name ↑	Status	Region	Zone	Total worker nodes	Scheduled deletion
<input type="checkbox"/>	hiveprojectcluster	Provisioning	us-central1	us-central1-a	2	Off

Jobs on Clusters

- Clusters
- Jobs
- Workflows
- Autoscaling policies
- Serverless

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.

Google Cloud hive-project Search (/) for resources, docs, products, and more

Dataproc **Clusters** [+ CREATE CLUSTER](#) [REFRESH](#) [START](#) [STOP](#) [DELETE](#) [REGIONS](#)

Jobs on Clusters Filter Search clusters, press Enter

<input type="checkbox"/>	Name ↑	Status	Region	Zone	Total worker nodes	Scheduled deletion
<input type="checkbox"/>	hiveprojectcluster	Running	us-central1	us-central1-a	2	Off

Jobs on Clusters

- Clusters
- Jobs
- Workflows
- Autoscaling policies
- Serverless

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.

Google Cloud

hive-project

Search (/) for resources, docs, products,

Dataproc

Jobs on Clusters

Clusters

Jobs

Workflows

Autoscaling policies

Serverless

Batches

Interactive

Metastore Services

Metastore

Federation

Utilities

Component exchange

Workbench

Cluster details

SUBMIT JOB REFRESH START STOP DELETE

Consider using Auto Zone rather than selecting a zone manually. See https://cloud.google.com/dataproc/do

Name	hiveprojectcluster
Cluster UUID	2905cd85-2424-454b-88c8-085c517945ec
Type	Dataproc Cluster
Status	Running

MONITORING

JOBS

VM INSTANCES

CONFIGURATION

WEB INTERFACES

SSH tunnel

Create an SSH tunnel to connect to a web interface

Component gateway

Provides access to the web interfaces of default and selected optional components on the cluster. [Learn more](#)

YARN ResourceManager

MapReduce Job History

Spark History Server

HDFS NameNode

YARN Application Timeline

Tez

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

digital-arcade-403408 > hiveprojectcluster

Sign out

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

Overview

'hiveprojectcluster-m.us-central1-a.c.digital-arcade-403408.internal.:8020' (✓active)

Started:	Sat Oct 28 15:10:33 +0530 2023
Version:	3.3.6, r7cd94d39bdfc5c0b636bd3c3e88d5d9c19335e44
Compiled:	Mon Oct 23 02:50:00 +0530 2023 by bigtop from (no branch)
Cluster ID:	CID-52ba3e7c-2d28-4499-b07c-3f1c169d65fc
Block Pool ID:	BP-765278231-10.128.0.3-1698485992152

Summary

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

The screenshot shows the Hadoop Overview page for a cluster named 'hiveprojectcluster-m.us-central1-a.c.digital-arcade'. The Utilities menu is open, showing options: 'Browse the file system' (highlighted with a yellow circle), Logs, Log Level, Metrics, Configuration, Process Thread Dump, and Network Topology. Below the menu is a table with cluster details.

Started:	Sat Oct 28 15:10:33 +0530 2023
Version:	3.3.6, r7cd94d39bdfc5c0b636bd3c3e88d5d9c19335e44
Compiled:	Mon Oct 23 02:50:00 +0530 2023 by bigtop from (no branch)
Cluster ID:	CID-52ba3e7c-2d28-4499-b07c-3f1c169d65fc
Block Pool ID:	BP-765278231-10.128.0.3-1698485992152

Below are the three you will see by default.

The screenshot shows the Hadoop Browse Directory page. It features a search bar with a 'Go!' button and icons for file operations. Below the search bar is a table listing directory entries. The table has columns for Permission, Owner, Group, Size, Last Modified, Replication, Block Size, Name, and a delete icon. Three entries are shown: 'tmp', 'user', and 'var', all with permissions 'drwxrwxrwt', owner 'hdfs', and group 'hadoop'. The page also includes a 'Showing 1 to 3 of 3 entries' message and pagination controls (Previous, 1, Next).

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
drwxrwxrwt	hdfs	hadoop	0 B	Oct 28 15:12	0	0 B	tmp	
drwxrwxrwt	hdfs	hadoop	0 B	Oct 28 15:11	0	0 B	user	
drwxrwxrwt	hdfs	hadoop	0 B	Oct 28 15:11	0	0 B	var	

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.

Cluster details

[SUBMIT JOB](#) [REFRESH](#) [START](#) [STOP](#) [DELETE](#) [VIEW LOGS](#)

i Consider using Auto Zone rather than selecting a zone manually. See <https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone>

Name	hiveprojectcluster
Cluster UUID	
Type	Dataproc Cluster
Status	Running

MONITORING JOBS **VM INSTANCES** CONFIGURATION WEB INTERFACES

Filter Filter instances

	Name	Role
✓	hiveprojectcluster-m	Master
✓	hiveprojectcluster-w-0	Worker
✓	hiveprojectcluster-w-1	Worker

[EQUIVALENT REST](#)

Now click on Master node. Click SSH.

Google Cloud

hive-project

Search (/) for resources, docs, pro

Compute Engine

Virtual machines

- VM instances
- Instance templates
- Sole-tenant nodes
- Machine images
- TPUs
- Committed use discounts
- Reservations

hiveprojectcluster-m [EDIT](#) [RESET](#) [CREATE MACHINE IMAGE](#)

DETAILS OBSERVABILITY OS INFO SCREENSHOT

SSH [CONNECT TO SERIAL CONSOLE](#)

Connecting to serial ports is disabled

Logs

[Logging](#)
[Serial port 1 \(console\)](#)
[SHOW MORE](#)

Basic information

Name	hiveprojectcluster-m
------	----------------------

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

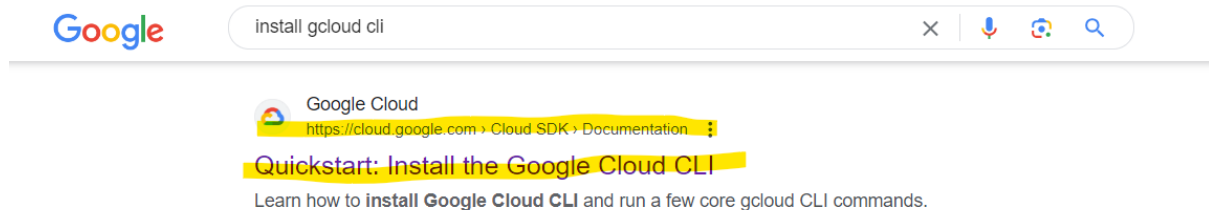
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

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:~$
```

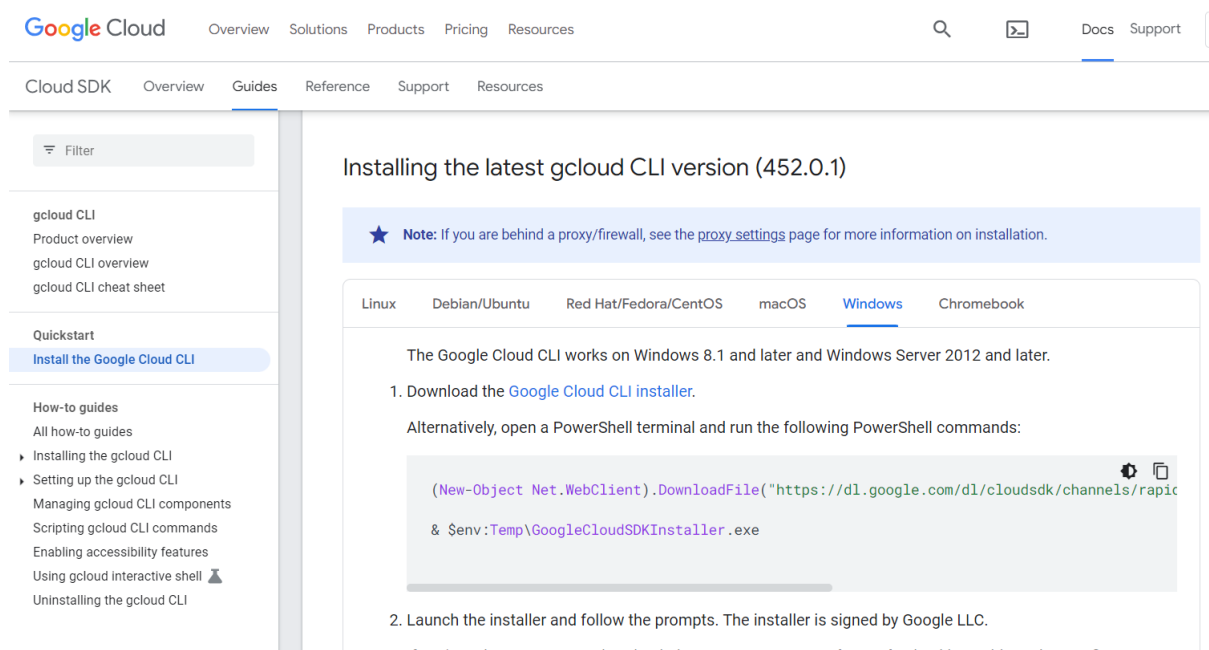
That's all.

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.

Linux Debian/Ubuntu Red Hat/Fedora/CentOS macOS **Windows** Chromebook

The Google Cloud CLI works on Windows 8.1 and later and Windows Server 2012 and later.

1. Download the [Google Cloud CLI installer](#).

Alternatively, open a PowerShell terminal and run the following PowerShell commands:

```
(New-Object Net.WebClient).DownloadFile("https://dl.google.com/dl/cloudsdk/channels/rapid/downloads/google-cloud-sdk-360.0.0-windows-x86_64.zip") & $env:Temp\GoogleCloudSDKInstaller.exe
```


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

Downloads

Name	Date modified	Type	Size
▼ Today			
GoogleCloudSDKInstaller	28-10-2023 12:21	Application	149 KB

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

Google Cloud CLI Setup



Welcome to Google Cloud CLI Setup

This wizard will guide you through the installation of the Google Cloud SDK.

Google Cloud SDK contains tools and libraries that will enable you to easily create and manage resources on Google Cloud Platform.

☐ Turn on screen reader mode

☐ Help make Google Cloud CLI better by automatically sending anonymous usage statistics to Google

[Learn More](#) [Privacy policy](#)

Next > Cancel

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\cmd X + v

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:

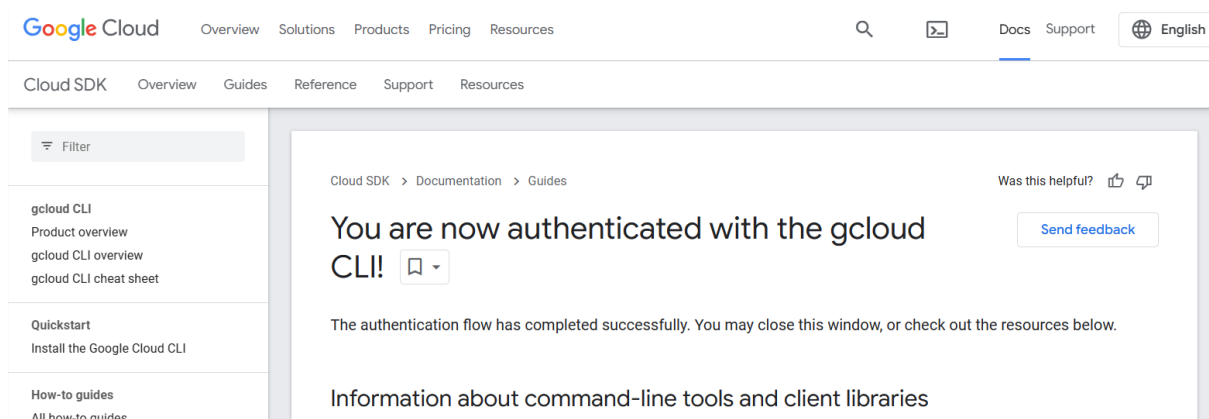
https://accounts.google.com/o/oauth2/auth?response_type=code&client_id=32555940559.apps.goo
%2F&scope=openid+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fuserinfo.email+https%3A%2F%2Fwww.goo
pis.com%2Fauth%2Fappengine.admin+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fsqlservice.login+htt
www.googleapis.com%2Fauth%2Faccounts.reauth&state=rW67tyYi28DQ4puKyTC3yFmLXSezA&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 successfully with the account you provided.

```
Your browser has been opened to visit:

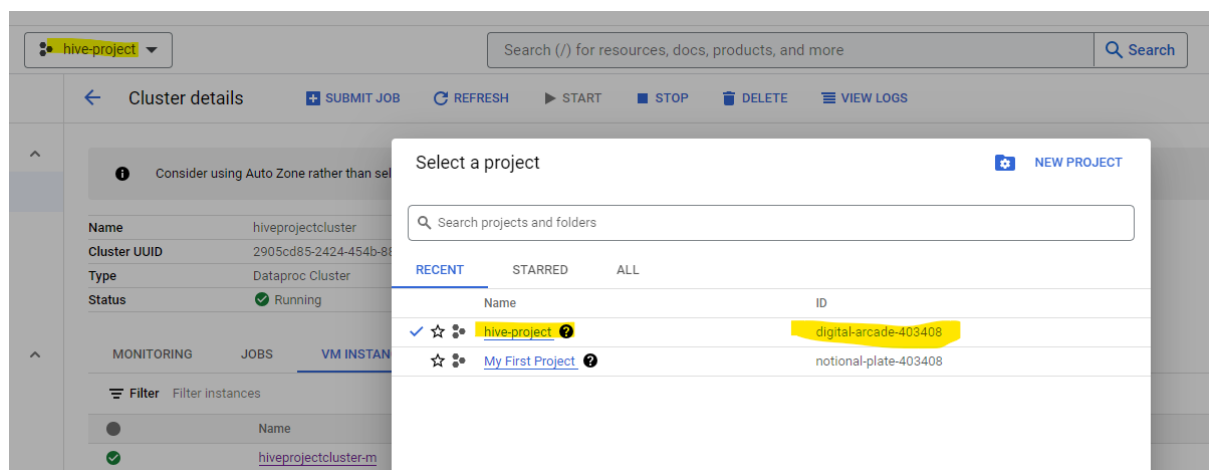
https://accounts.google.com/o/oauth2/auth?response_type=code&client_id=32555940559.apps.googleusercontent.com%2F&scope=openid+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fuserinfo.email+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fengine.admin+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fsqlservice.login+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Faccounts.reauth&state=SJjnb8Ay2NrMIJ226nHWUgDLGAWvCG&access_type=offline&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/LD7WbXbFGRsO/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

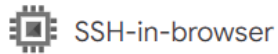


SSH-in-browser

```
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.

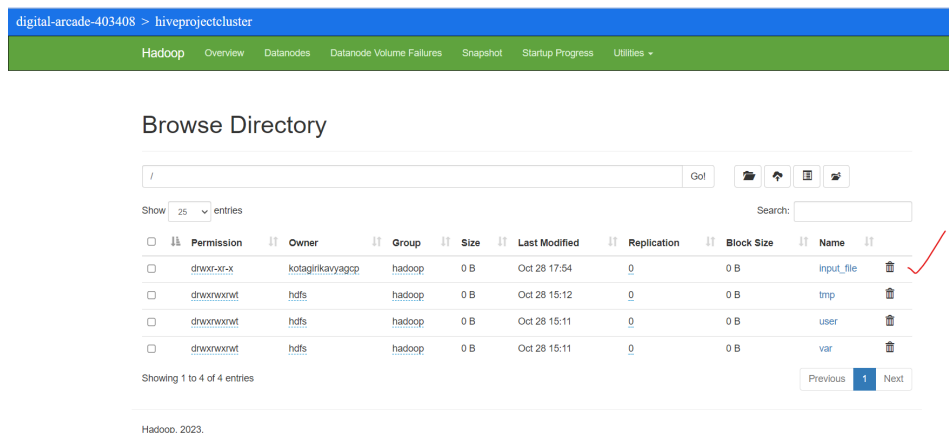
```
hdfs dfs -ls /
```



```
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 directory 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.

digital-arcade-403408 > hiveprojectcluster

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/input_file Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	kotagiriikavyagcp	hadoop	286.97 MB	Oct 28 18:00	2	128 MB	Reviews.csv

Showing 1 to 1 of 1 entries

Previous 1 Next

Hadoop, 2023.

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.

File information - Reviews.csv

Download Head the file (first 32K) Tail the file (last 32K)

Block information -- Block 0

Block ID: 10737418

Block Pool ID: BP-765278231-10.128.0.3-1698485992152

Generation Stamp: 1004

Size: 134217728

Availability:

- hiveprojectcluster-w-0.us-central1-a.c.digital-arcade-403408.internal
- hiveprojectcluster-w-1.us-central1-a.c.digital-arcade-403408.internal

Close

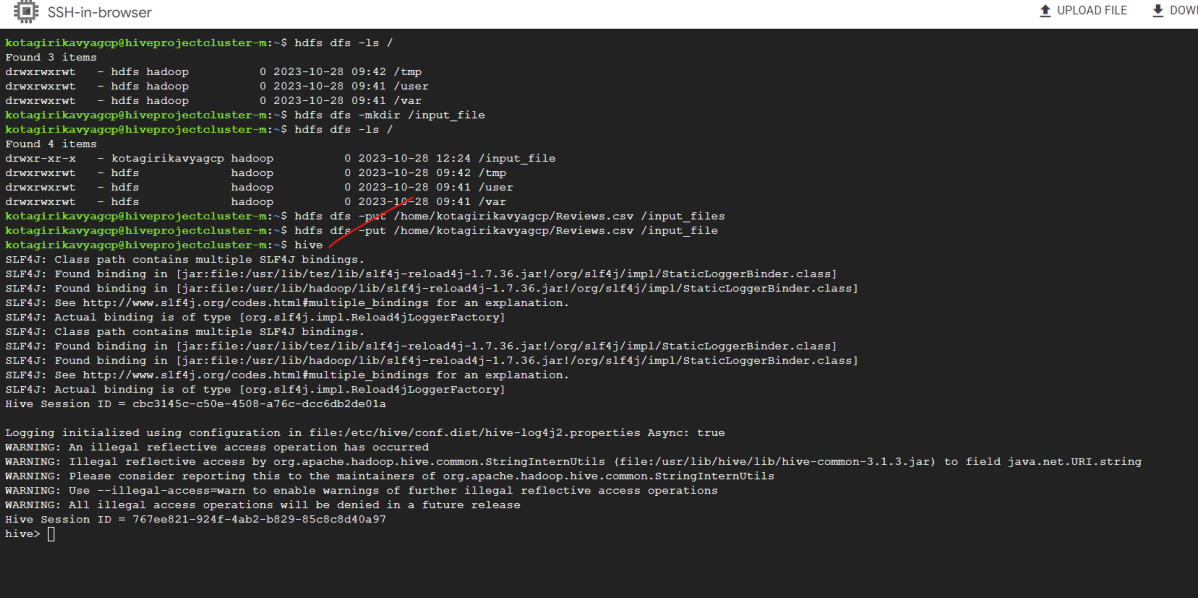
It created 3 blocks as the file size is 286 MB and each block size is 128 MB. So, it created 3 blocks.

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 from our local to hdfs.

44. To execute hive commands in SSH

Just type hive and hit enter.



The image shows a terminal window titled "SSH-in-browser" with a toolbar containing "UPLOAD FILE" and "DOWN" buttons. The terminal content is as follows:

```
kotagirikavyagcp@hiveprojectcluster-m:~$ hdfs dfs -ls /
Found 3 items
drwxr-xr-x - hdfs hadoop          0 2023-10-28 09:42 /tmp
drwxr-xrwt - hdfs hadoop          0 2023-10-28 09:41 /user
drwxr-xrwt - 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
drwxr-xrwt - hdfs hadoop          0 2023-10-28 09:42 /tmp
drwxr-xrwt - hdfs hadoop          0 2023-10-28 09:41 /user
drwxr-xrwt - 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-b829-85c8cd40a97
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.

The screenshot shows the Google Cloud Dataproc Clusters page for a project named 'hive-project'. The left sidebar contains navigation links for Clusters, Jobs, Workflows, Autoscaling policies, Serverless, Batches, Interactive, Metastore Services, and Metastore. The main content area displays a table of clusters. The cluster 'hiveprojectcluster' is in a 'Running' state. Above the table, there are buttons for 'CREATE CLUSTER', 'REFRESH', 'START', 'STOP', and 'DELETE'. The 'STOP' button is highlighted with a red checkmark, indicating the action to be taken.

Name	Status	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket
hiveprojectcluster	Running	us-central1	us-central1-a	2	Off	dataproc-staging-us-central1-565254892126-nwcnwyun

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

The screenshot shows the same Google Cloud Dataproc Clusters page, but the cluster 'hiveprojectcluster' is now in a 'Stopped' state. The 'STOP' button in the top navigation bar is now disabled. The 'Status' column in the table shows 'Stopped' with a red circle around it, indicating the change in status.

Name	Status	Region	Zone	Total worker nodes	Scheduled deletion
hiveprojectcluster	Stopped	us-central1	us-central1-a	2	Off