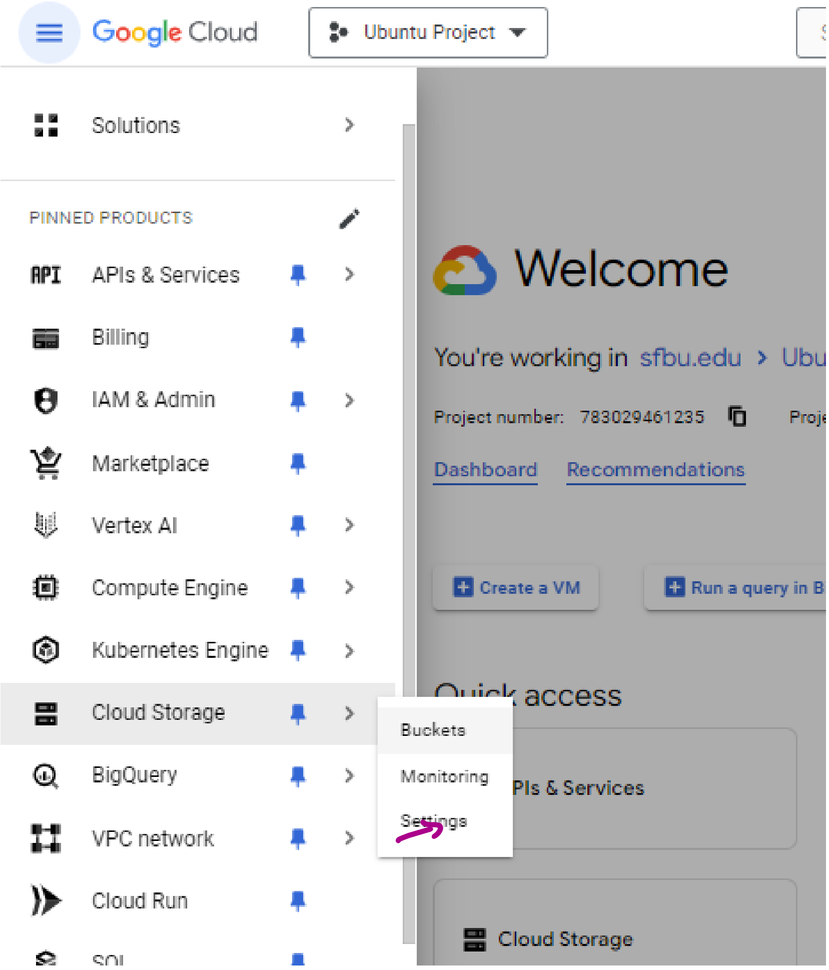
Name: Sened Desalegn

Pyspark on GCP

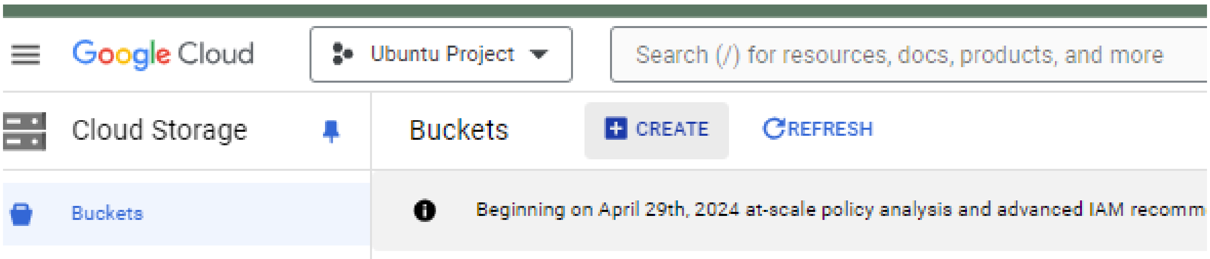
Create a Bucket

1. To create a bucket in Google Cloud Storage, navigate to the Cloud Console, select or create a project, then go to Storage > Browser, click "Create Bucket", name it uniquely, choose a location, and click "Create".



2. Click Create bucket.

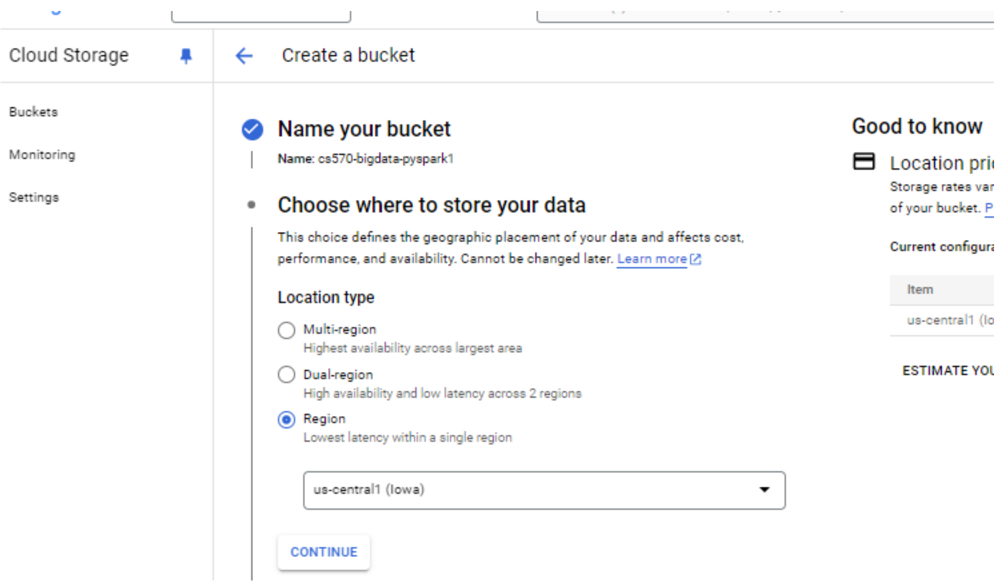
3. On the Create a Bucket page, enter your bucket information. To go to the next step, click Continue.



Beginning on April 29th, 2024 at-scale policy analysis and advanced IAM recomm

4. For Name your bucket, enter a name that meets the bucket name requirements. Choose the cheaper region. And rest things can be default values.

Name: Sened Desalegn



Name your bucket

Name: cs570-bigdata-pyspark1

Choose where to store your data

This choice defines the geographic placement of your data and affects cost, performance, and availability. Cannot be changed later. Learn more

Location type

○ Multi-region
Highest availability across largest area

○ Dual-region
High availability and low latency across 2 regions

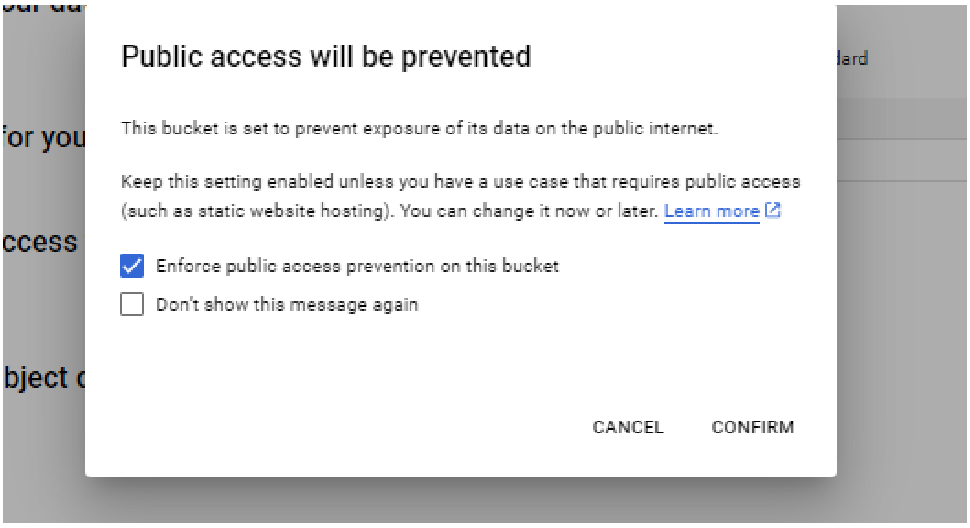
● Region
Lowest latency within a single region

us-central1 (Iowa)

Storage rates var of your bucket. P

After you click “Create”, you might face a pop up message like this:

Just click “Confirm”.



Public access will be prevented

This bucket is set to prevent exposure of its data on the public internet.

Keep this setting enabled unless you have a use case that requires public access (such as static website hosting). You can change it now or later. Learn more

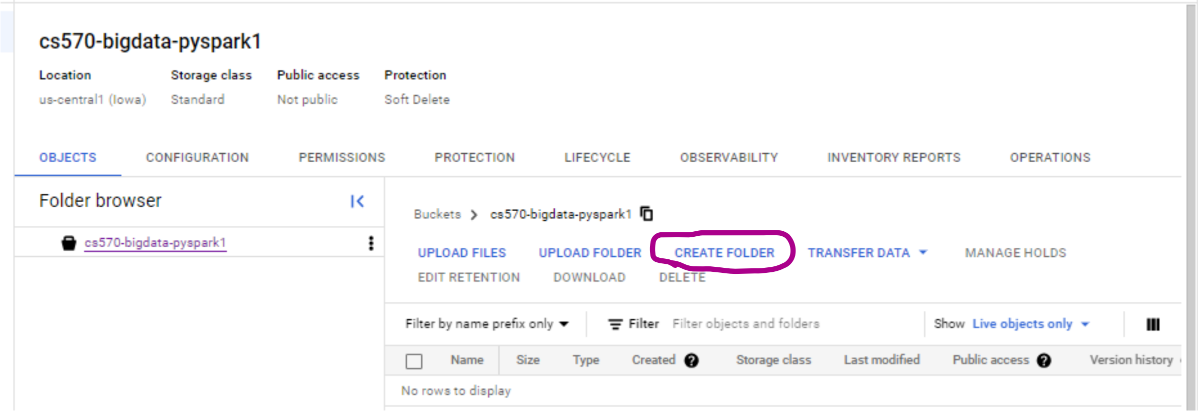
☑ Enforce public access prevention on this bucket

□ Don't show this message again

5. Now, we create a folder in the bucket and add sample.txt, you can follow these steps:

- ✓ In the Google Cloud console, go to the Cloud Storage Buckets page.
- ✓ Click on the name of the bucket you created

Name: Sened Desalegn

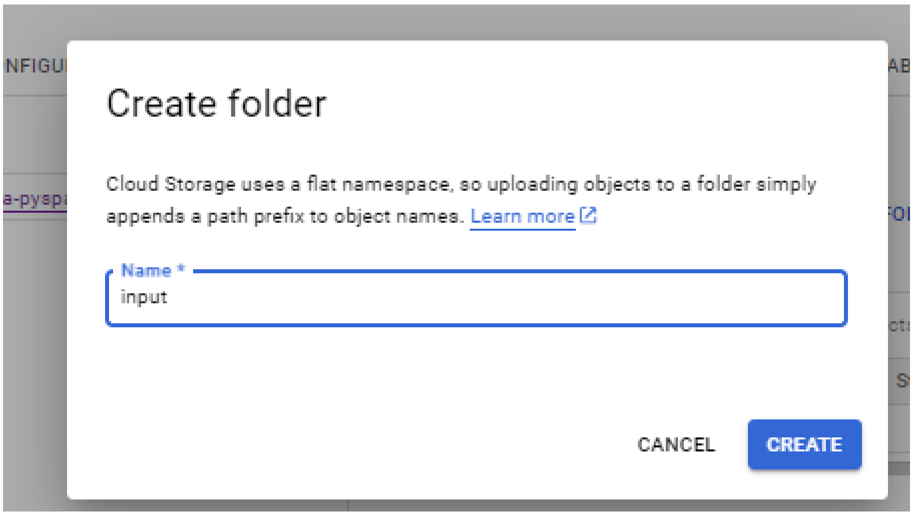

cs570-bigdata-pyspark1

Location	Storage class	Public access	Protection
us-central1 (Iowa)	Standard	Not public	Soft Delete

☐	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history

No rows to display

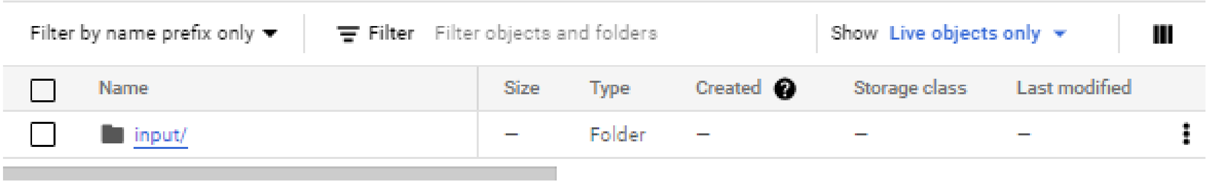
- ✓ Click on the Create folder button.
- ✓ Enter a name for your folder and click Create.


Create folder

Cloud Storage uses a flat namespace, so uploading objects to a folder simply appends a path prefix to object names. Learn more

Name *: input

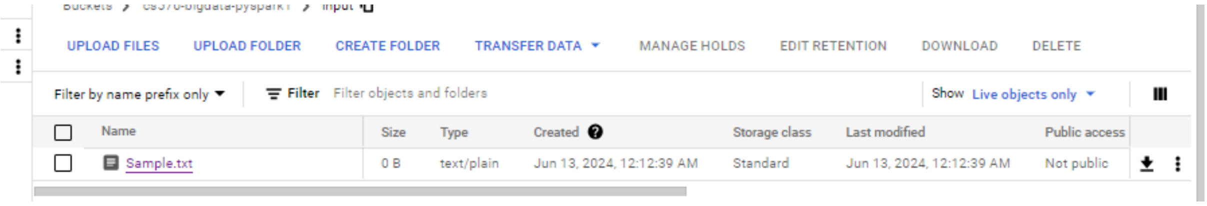
6. Click on the name of the folder you just created.



☐	Name	Size	Type	Created	Storage class	Last modified
☐	input/	–	Folder	–	–	–

7. Click on the Upload files button.

After I have uploaded the sample.txt file I have created.

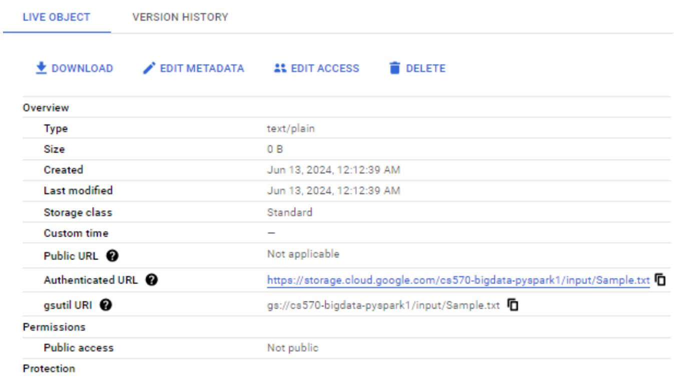


☐	Name	Size	Type	Created	Storage class	Last modified	Public access
☐	Sample.txt	0 B	text/plain	Jun 13, 2024, 12:12:39 AM	Standard	Jun 13, 2024, 12:12:39 AM	Not public

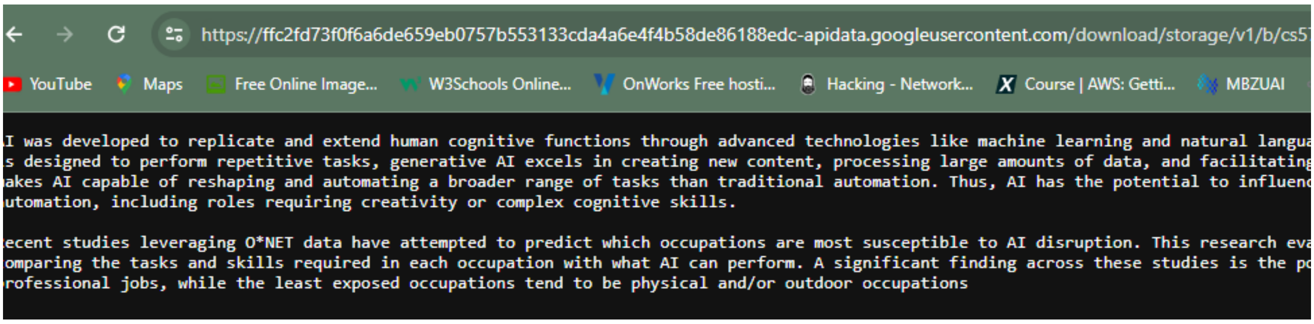
8. Select your sample.txt file and click Open. The sample file can be any text file. I created a sample file containing information about Generative AI.

Name: Sened Desalegn

If you click on the authenticated URL, you will get what is inside the sample.txt file.



Overview	
Type	text/plain
Size	0 B
Created	Jun 13, 2024, 12:12:39 AM
Last modified	Jun 13, 2024, 12:12:39 AM
Storage class	Standard
Custom time	–
Public URL	Not applicable
Authenticated URL	https://storage.cloud.google.com/cs570-bigdata-pyspark1/input/Sample.txt
gsutil URI	gs://cs570-bigdata-pyspark1/input/Sample.txt
Permissions	
Public access	Not public
Protection	



https://ffc2fd73f0f6a6de659eb0757b553133cda4a6e4f4b58de86188edc-apidata.googleusercontent.com/download/storage/v1/b/cs5

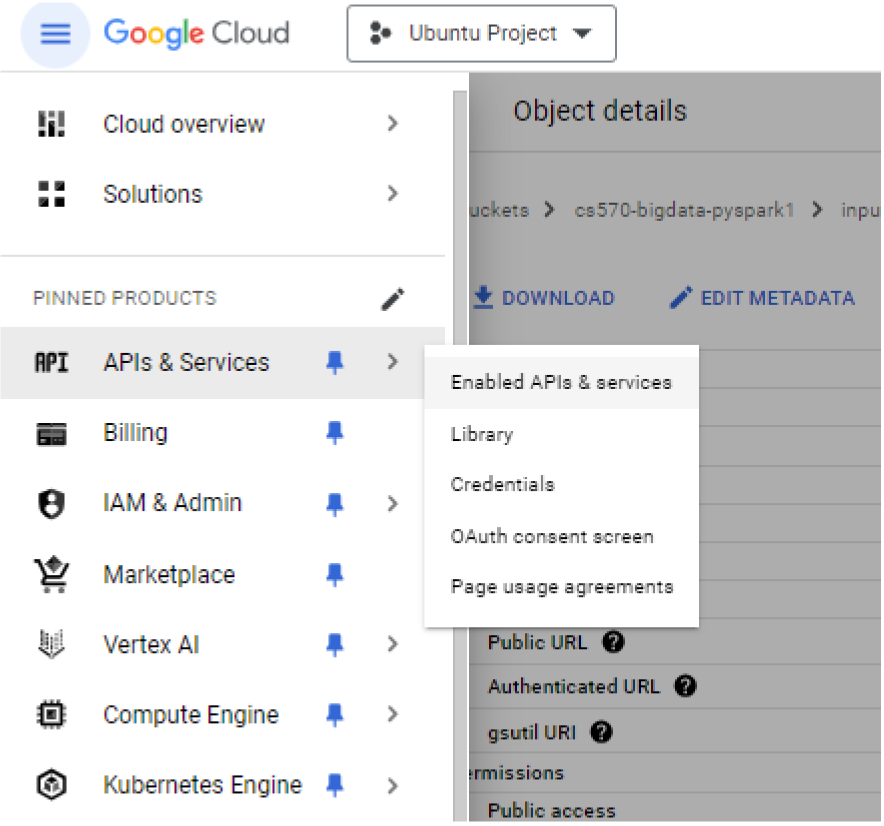
```
I was developed to replicate and extend human cognitive functions through advanced technologies like machine learning and natural langua
s designed to perform repetitive tasks, generative AI excels in creating new content, processing large amounts of data, and facilitating
akes AI capable of reshaping and automating a broader range of tasks than traditional automation. Thus, AI has the potential to influenc
utomation, including roles requiring creativity or complex cognitive skills.

ecent studies leveraging O*NET data have attempted to predict which occupations are most susceptible to AI disruption. This research eva
omparing the tasks and skills required in each occupation with what AI can perform. A significant finding across these studies is the pc
rofessional jobs, while the least exposed occupations tend to be physical and/or outdoor occupations
```

Create a Dataproc cluster:

To create a Dataproc cluster, follow these steps:

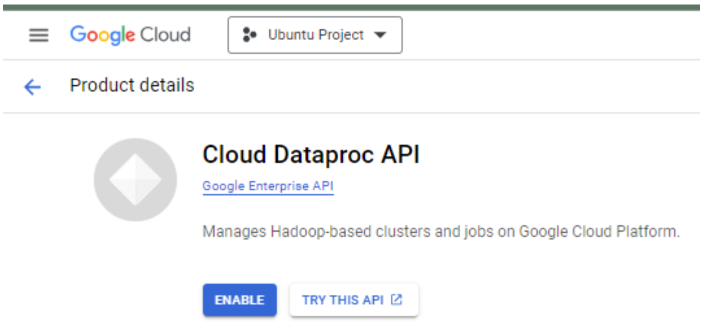
1. Navigate to the APIs & Services page in the Google Cloud Console.
2. Click on "Enable APIs and Services"

Name: Sened Desalegn



3. Search for “Cloud Dataproc API” and select it.

4. Click Enable.

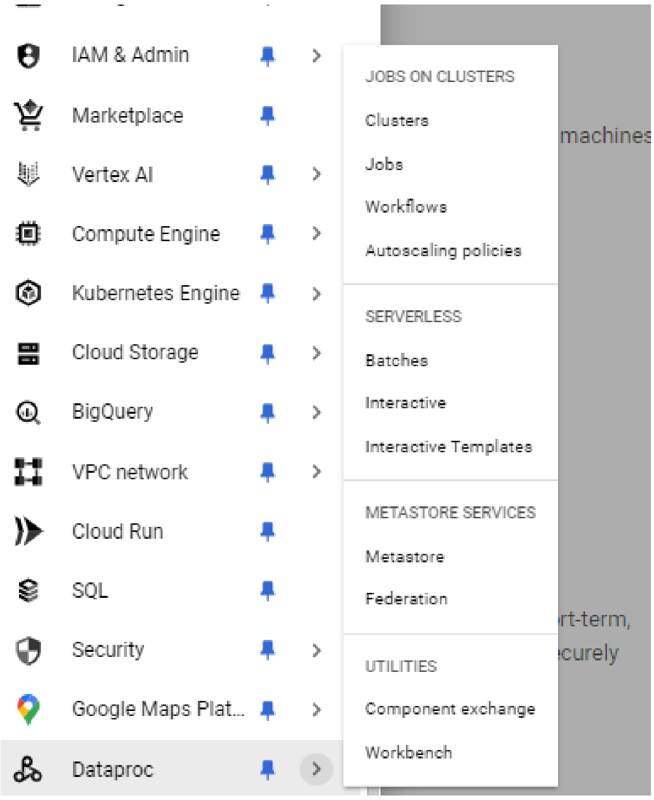


Manages Hadoop-based clusters and jobs on Google Cloud Platform.

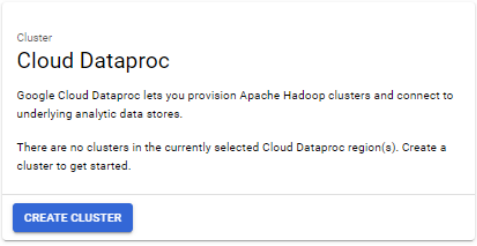
To create a Dataproc cluster in Google Cloud Platform, you can use the Google Cloud Console. Here are the steps:

1. In the Google Cloud console, go to the Dataproc Clusters page.

Name: Sened Desalegn



2. Click Create Cluster.

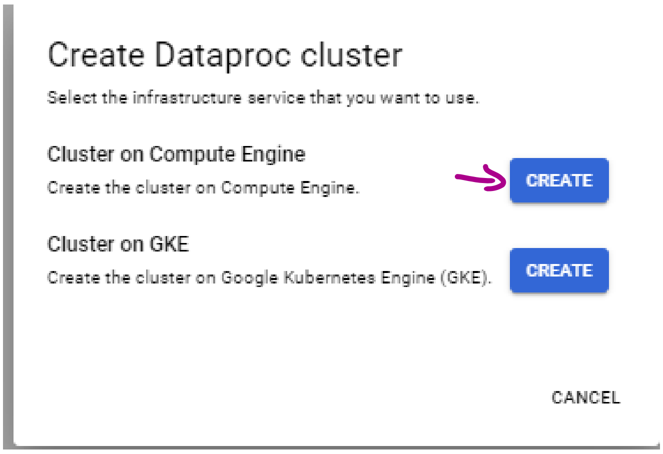


Cloud Dataproc

Google Cloud Dataproc lets you provision Apache Hadoop clusters and connect to underlying analytic data stores.

There are no clusters in the currently selected Cloud Dataproc region(s). Create a cluster to get started.

3. In the Create Dataproc cluster dialog, click Create in the Cluster on Compute engine row.



Create Dataproc cluster

Select the infrastructure service that you want to use.

Cluster on Compute Engine

Create the cluster on Compute Engine.

Cluster on GKE

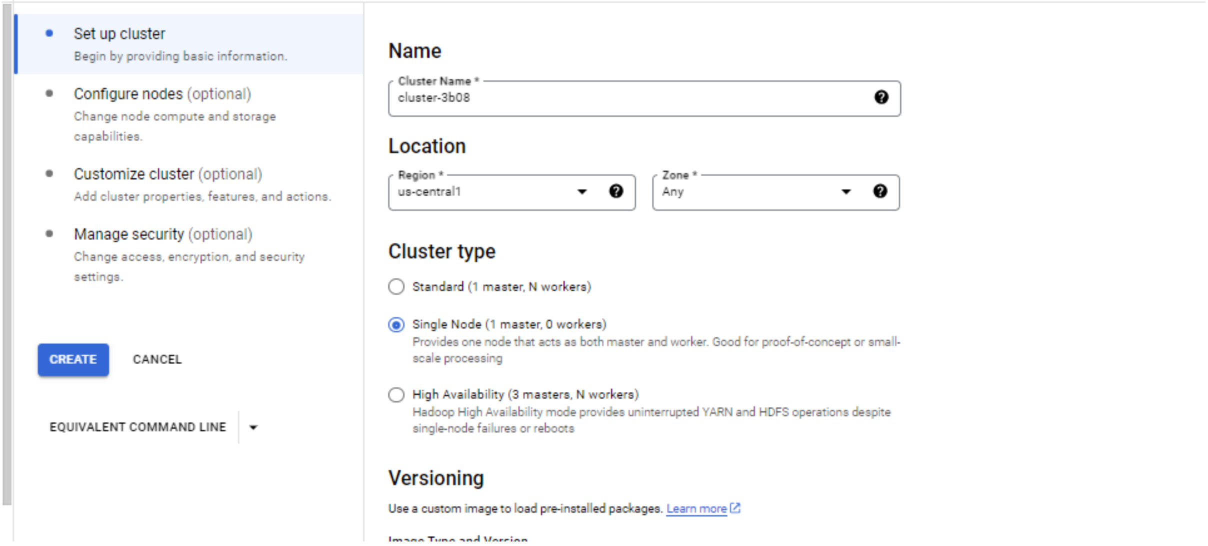
Create the cluster on Google Kubernetes Engine (GKE).

4. In the Cluster Name field, enter a name for your cluster.

5. In the Region and Zone lists, select the same region as the bucket and zone. Choose the cluster type as a single node. And the rest can be the default.

Name: Sened Desalegn

6. For all the other options, use the default settings. 7. To create the cluster, click Create.


• 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

Name

Cluster Name *: cluster-3b08

Location

Region *: us-central1

Zone *: Any

Cluster type

○ Standard (1 master, N workers)

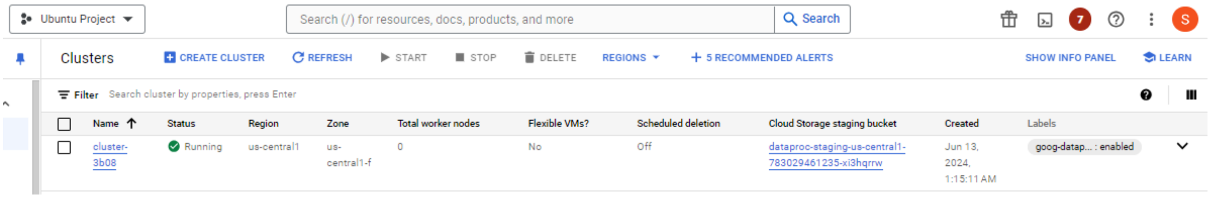
● Single Node (1 master, 0 workers)
Provides one node that acts as both master and worker. Good for proof-of-concept or small-scale processing

○ High Availability (3 masters, N workers)
Hadoop High Availability mode provides uninterrupted YARN and HDFS operations despite single-node failures or reboots

Versioning

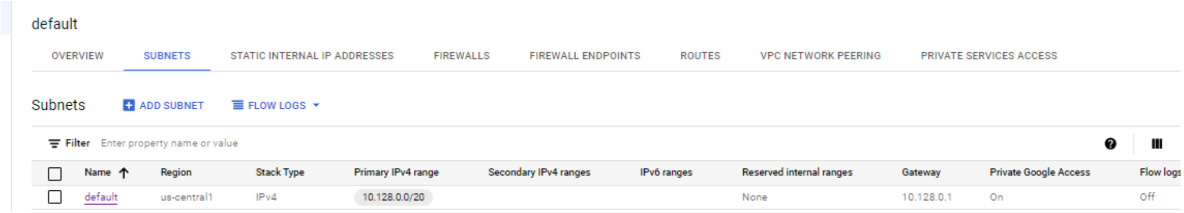
Use a custom image to load pre-installed packages. Learn more

Image Type and Version



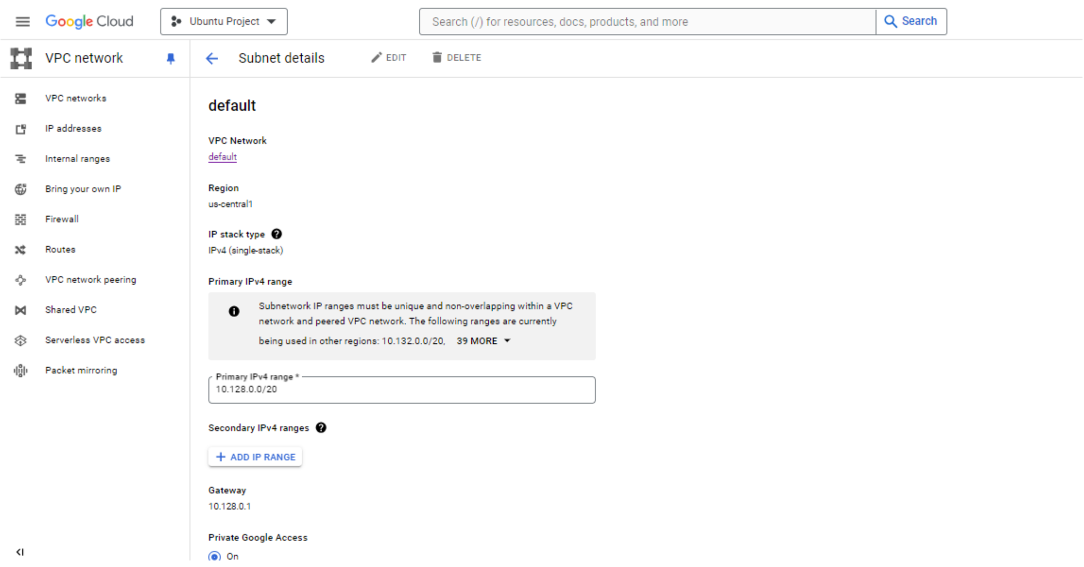
□	Name ↑	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging bucket	Created	Labels
□	cluster-3b08	✓ Running	us-central1	us-central1-f	0	No	Off	dataproc-staging-us-central1-783029461235-xi3hqrrw	Jun 13, 2024, 1:15:11 AM	goog-datap... : enabled

Note: If you get error for subnetting issues, first you have to navigate to VPC networks, and under the VPC network, there is default and under default you have to select the subnet based on the zone you have and edit the network.



□	Name ↑	Region	Stack Type	Primary IPv4 range	Secondary IPv4 ranges	IPv6 ranges	Reserved internal ranges	Gateway	Private Google Access	Flow logs
□	default	us-central1	IPv4	10.128.0.0/20			None	10.128.0.1	On	Off

Name: Sened Desalegn



default

VPC Network
default

Region
us-central1

IP stack type
IPv4 (single-stack)

Primary IPv4 range

Subnetwork IP ranges must be unique and non-overlapping within a VPC network and peered VPC network. The following ranges are currently being used in other regions: 10.132.0.0/20, 39 MORE

Primary IPv4 range *
10.128.0.0/20

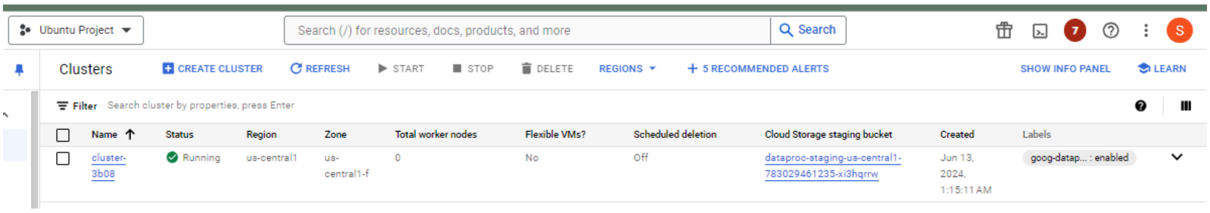
Secondary IPv4 ranges

Gateway
10.128.0.1

Private Google Access
● On

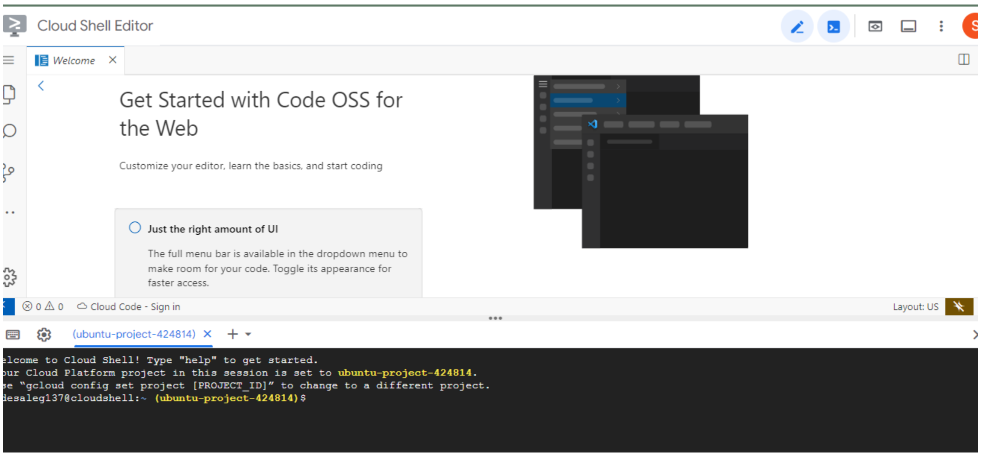
To create and save the spark job python file:

1. In the top right corner of the console, click the Activate Cloud Shell button.



□	Name	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging bucket	Created	Labels
□	cluster-3b08	✓ Running	us-central1	us-central1-f	0	No	Off	dataproc-staging-us-central1-783029461235-xi3hqrrw	Jun 13, 2024, 1:15:11 AM	goog-datap... : enabled

2. Once the Cloud Shell is activated, click on the Open Editor button in the top right corner of the Cloud Shell window.



Get Started with Code OSS for the Web

Customize your editor, learn the basics, and start coding

○ Just the right amount of UI

The full menu bar is available in the dropdown menu to make room for your code. Toggle its appearance for faster access.

```
elcome to Cloud Shell! Type "help" to get started.
our Cloud Platform project in this session is set to ubuntu-project-424814.
se "gcloud config set project [PROJECT_ID]" to change to a different project.
desaleg137@cloudshell:~ (ubuntu-project-424814)$
```

3. Click on the new file icon beside your username to create a new file

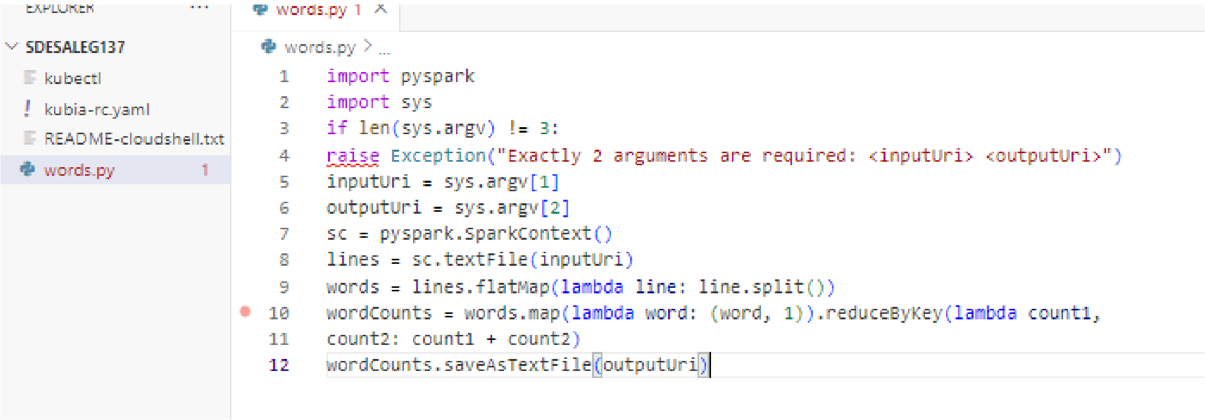
Here is the python code used:

import pyspark

import sys

Name: Sened Desalegn

```
if len(sys.argv) != 3:

raise Exception("Exactly 2 arguments are required: <inputUri> <outputUri>")

inputUri = sys.argv[1]

outputUri = sys.argv[2]

sc = pyspark.SparkContext()

lines = sc.textFile(inputUri)

words = lines.flatMap(lambda line: line.split())

wordCounts = words.map(lambda word: (word, 1)).reduceByKey(lambda count1,

count2: count1 + count2)

wordCounts.saveAsTextFile(outputUri)
```



```
import pyspark
import sys
if len(sys.argv) != 3:
raise Exception("Exactly 2 arguments are required: <inputUri> <outputUri>")
inputUri = sys.argv[1]
outputUri = sys.argv[2]
sc = pyspark.SparkContext()
lines = sc.textFile(inputUri)
words = lines.flatMap(lambda line: line.split())
wordCounts = words.map(lambda word: (word, 1)).reduceByKey(lambda count1,
count2: count1 + count2)
wordCounts.saveAsTextFile(outputUri)
```

Save the file. I saved it as words.py and close the shell.

Running PySpark Job on Google Cloud Dataproc and Sorting Output Files:

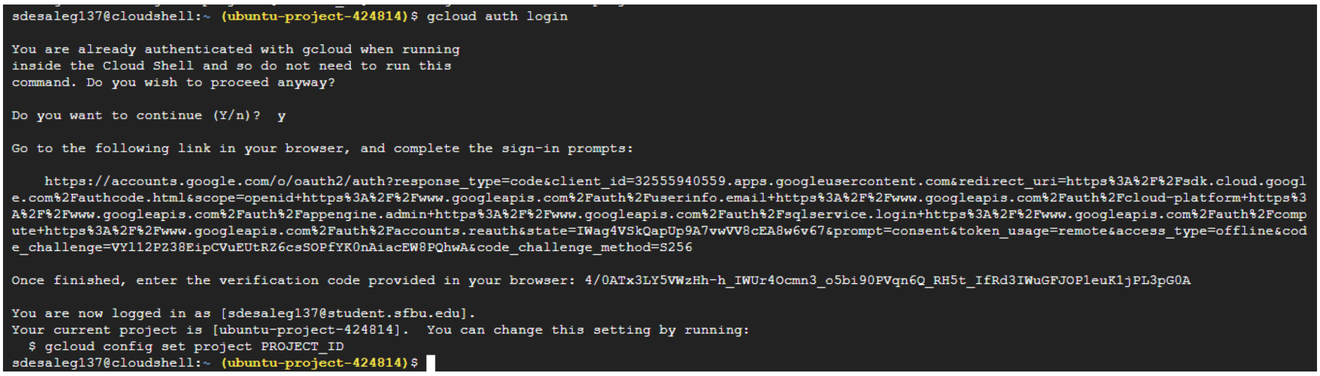
Step 1: Click on "Activate Cloud Shell" like we did previously. Authenticate with Google Cloud Platform (GCP). If you encounter an authentication error, run the command:

```
$ gcloud auth login
```

This command will open a web page where you can authenticate and obtain new credentials. Follow the on-screen instructions to complete the authentication process

Click on the click, authorize, and copy the code. Paste the same code in the authentication code.

Name: Sened Desalegn

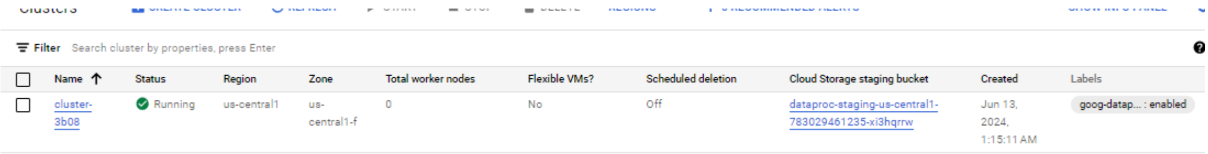


```
sdesaleg137@cloudshell:~ (ubuntu-project-424814)$ gcloud auth login

You are already authenticated with gcloud when running
inside the Cloud Shell and so do not need to run this
command. Do you wish to proceed anyway?

Do you want to continue (Y/n)?  y

Go to the following link in your browser, and complete the sign-in prompts:

    https://accounts.google.com/o/oauth2/auth?response_type=code&client_id=32555940559.apps.googleusercontent.com&redirect_uri=https%3A%2F%2Fsdk.cloud.googl
e.com%2Fauthcode.html&scope=openid+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fuserinfo.email+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fcloud-platform+https%3
A%2F%2Fwww.googleapis.com%2Fauth%2Fappengine.admin+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fsqlservice.login+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fcomp
ute+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Faccounts.reauth&state=IWag4VSkQapUp9A7vwVV8cEA8w6v67&prompt=consent&token_usage=remote&access_type=offline&cod
e_challenge=VYl12PZ38EipCVuEUtRZ6csSOPfYK0nAiacEW8PQhwA&code_challenge_method=S256

Once finished, enter the verification code provided in your browser: 4/0ATx3LY5VWzHh-h_IWUr4Ocmn3_o5bi90PVqn6Q_RH5t_IfRd3IWuGFJOP1euK1jPL3pG0A

You are now logged in as [sdesaleg137@student.sfbu.edu].
Your current project is [ubuntu-project-424814].  You can change this setting by running:
  $ gcloud config set project PROJECT_ID
sdesaleg137@cloudshell:~ (ubuntu-project-424814)$
```

Step 2: Submit the PySpark job to Dataproc

This command submits a PySpark job (words.py) to the Dataproc cluster named cluster-3b08, which is located in the us-central1 region. The job processes an input file (sample.txt) from a specified GCS bucket and writes the output to another GCS directory.

Change the cluster name and Cluster Region, by going back to your cluster.



Filter Search cluster by properties, press Enter

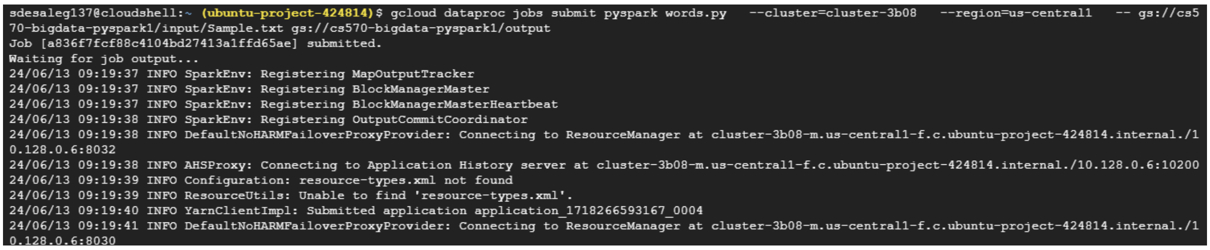
□	Name ↑	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging bucket	Created	Labels
□	cluster-3b08	✓ Running	us-central1	us-central1-f	0	No	Off	dataproc-staging-us-central1-783029461235-xi3hqrrw	Jun 13, 2024, 1:15:11 AM	goog-datap... : enabled

$ gcloud dataproc jobs submit pyspark words.py \

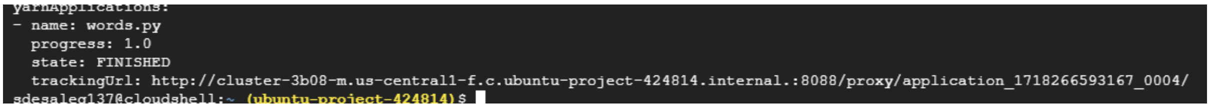
--cluster=cluster-3b08 \

--region=us-central1 \

-- gs://cs570-bigdata-pyspark/input/sample.txt gs://cs570-bigdata-pyspark/output



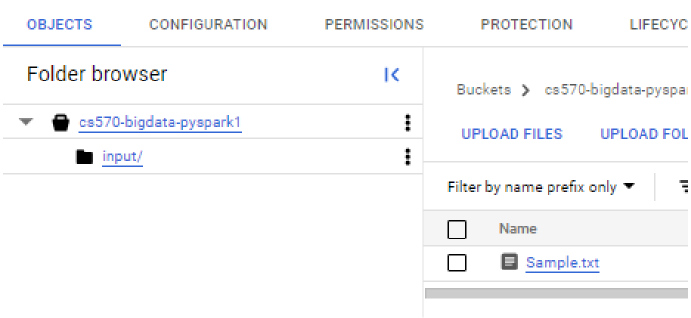
```
sdesaleg137@cloudshell:~ (ubuntu-project-424814)$ gcloud dataproc jobs submit pyspark words.py   --cluster=cluster-3b08   --region=us-central1   -- gs://cs5
70-bigdata-pyspark1/input/Sample.txt gs://cs570-bigdata-pyspark1/output
Job [a836f7fcf88c4104bd27413a1ffd65ae] submitted.
Waiting for job output...
24/06/13 09:19:37 INFO SparkEnv: Registering MapOutputTracker
24/06/13 09:19:37 INFO SparkEnv: Registering BlockManagerMaster
24/06/13 09:19:37 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/06/13 09:19:38 INFO SparkEnv: Registering OutputCommitCoordinator
24/06/13 09:19:38 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-3b08-m.us-central1-f.c.ubuntu-project-424814.internal./1
0.128.0.6:8032
24/06/13 09:19:38 INFO AHSProxy: Connecting to Application History server at cluster-3b08-m.us-central1-f.c.ubuntu-project-424814.internal./10.128.0.6:10200
24/06/13 09:19:39 INFO Configuration: resource-types.xml not found
24/06/13 09:19:39 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/06/13 09:19:40 INFO YarnClientImpl: Submitted application application_1718266593167_0004
24/06/13 09:19:41 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-3b08-m.us-central1-f.c.ubuntu-project-424814.internal./1
0.128.0.6:8030
```



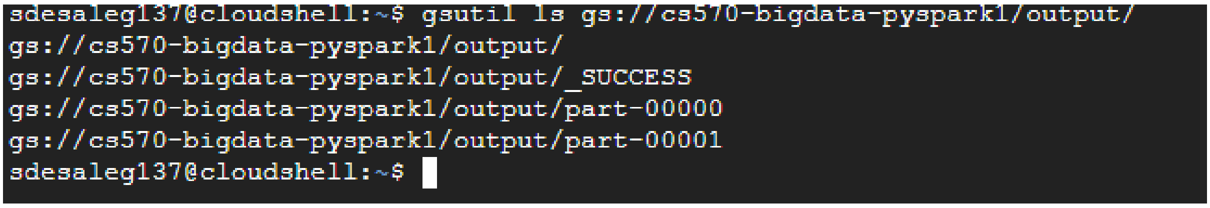
```
yarnApplications:
- name: words.py
  progress: 1.0
  state: FINISHED
  trackingUrl: http://cluster-3b08-m.us-central1-f.c.ubuntu-project-424814.internal.:8088/proxy/application_1718266593167_0004/
sdesaleg137@cloudshell:~ (ubuntu-project-424814)$
```

To get the location, open the bucket we created earlier, and go to the sample.txt. (My bucket name is cs570-bigdata-pyspark)

Name: Sened Desalegn



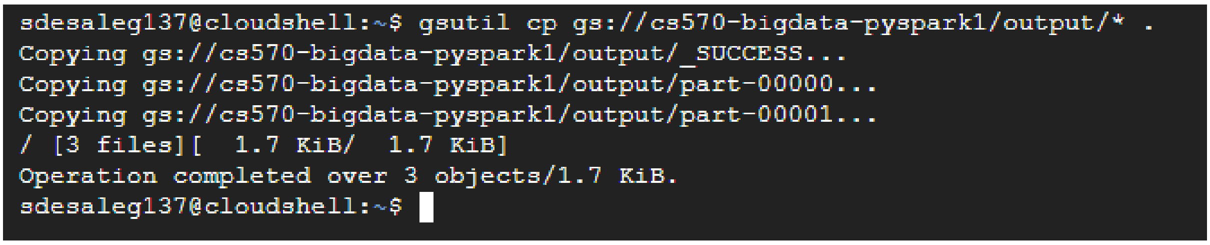
Step 3: List the files in the output directory



```
sdesaleg137@cloudshell:~$ gsutil ls gs://cs570-bigdata-pyspark1/output/
gs://cs570-bigdata-pyspark1/output/
gs://cs570-bigdata-pyspark1/output/_SUCCESS
gs://cs570-bigdata-pyspark1/output/part-00000
gs://cs570-bigdata-pyspark1/output/part-00001
sdesaleg137@cloudshell:~$
```

This command lists the files present in the gs://cs570-bigdata-pyspark1/output/ GCS bucket directory.

Step 4: Copy the output files to the current directory

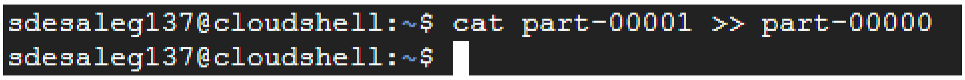


```
sdesaleg137@cloudshell:~$ gsutil cp gs://cs570-bigdata-pyspark1/output/* .
Copying gs://cs570-bigdata-pyspark1/output/_SUCCESS...
Copying gs://cs570-bigdata-pyspark1/output/part-00000...
Copying gs://cs570-bigdata-pyspark1/output/part-00001...
/ [3 files][  1.7 KiB/   1.7 KiB]
Operation completed over 3 objects/1.7 KiB.
sdesaleg137@cloudshell:~$
```

This command copies all the files from the gs://cs570-bigdata-pyspark1/output/ GCS bucket directory to the current directory in the Cloud Shell environment.

Step 5: Combine the contents of part-00000 and part-00001 files

$ cat part-00001 >> part-00000



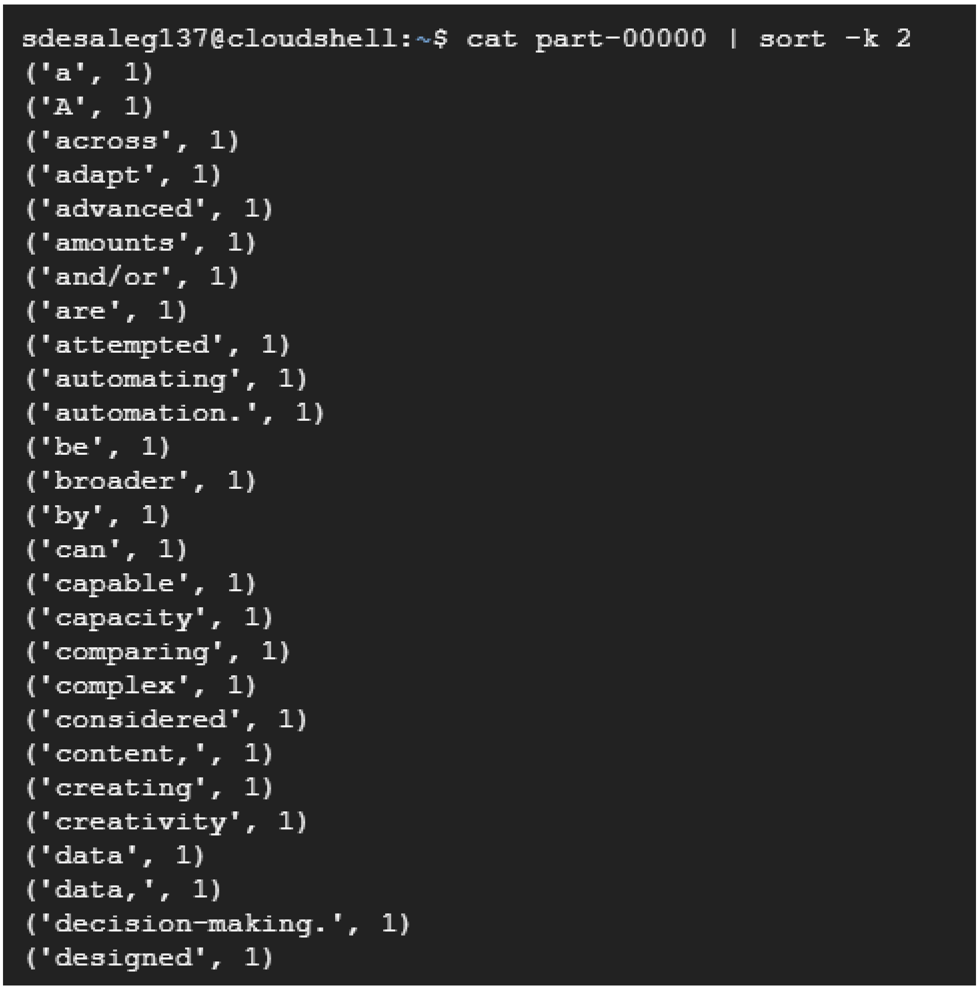
```
sdesaleg137@cloudshell:~$ cat part-00001 >> part-00000
sdesaleg137@cloudshell:~$
```

This command appends the contents of the part-00001 file to the end of the part-00000 file.

Step 6: Sort the combined file based on the second column

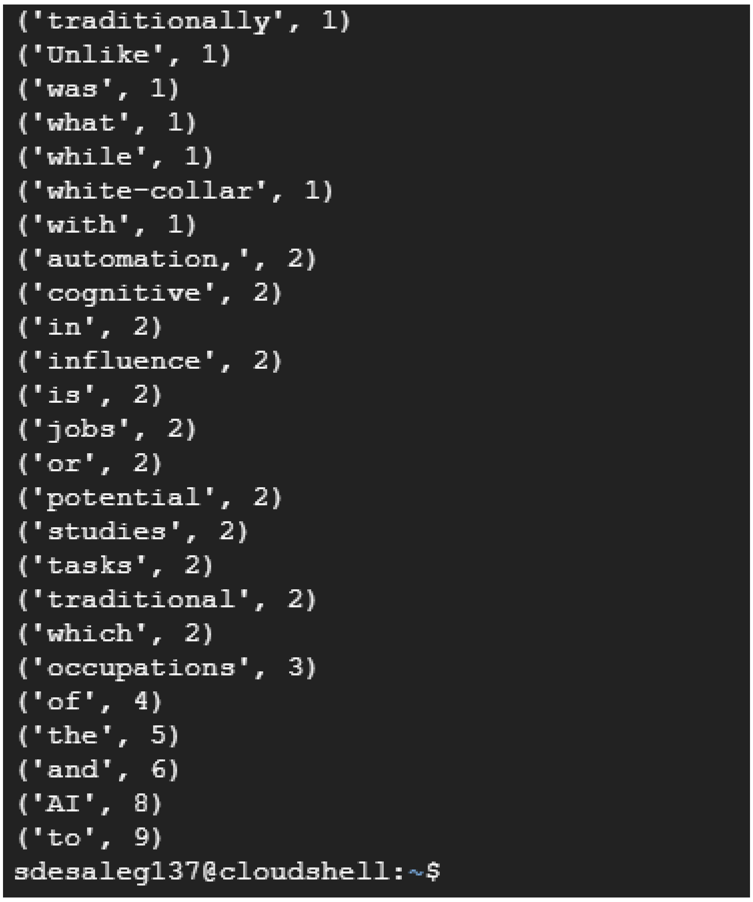
This command reads the contents of the part-00000 file and pipes (|) it as input to the sort command. The sort command then sorts the input based on the second column (-k 2). The sorted output will be displayed in the terminal for easy viewing.

Name: Sened Desalegn



```
sdesaleg137@cloudshell:~$ cat part-00000 | sort -k 2
('a', 1)
('A', 1)
('across', 1)
('adapt', 1)
('advanced', 1)
('amounts', 1)
('and/or', 1)
('are', 1)
('attempted', 1)
('automating', 1)
('automation.', 1)
('be', 1)
('broader', 1)
('by', 1)
('can', 1)
('capable', 1)
('capacity', 1)
('comparing', 1)
('complex', 1)
('considered', 1)
('content,', 1)
('creating', 1)
('creativity', 1)
('data', 1)
('data,', 1)
('decision-making.', 1)
('designed', 1)
```

Name: Sened Desalegn

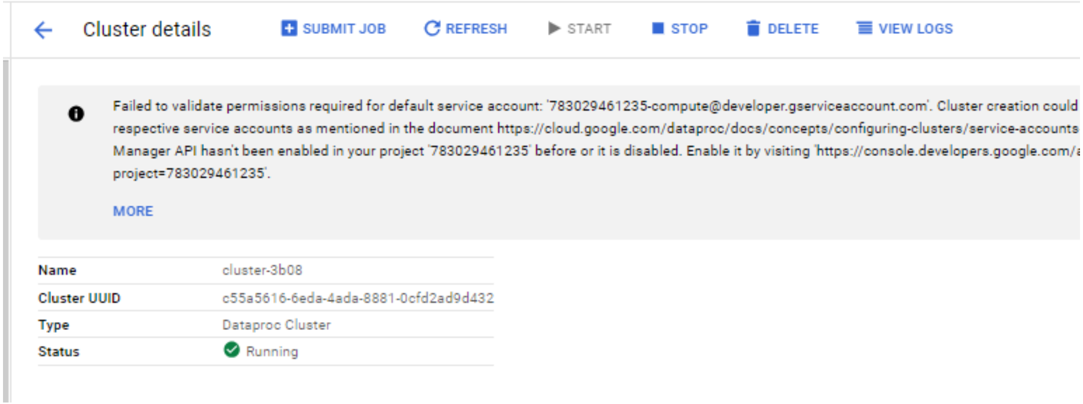


```
('traditionally', 1)
('Unlike', 1)
('was', 1)
('what', 1)
('while', 1)
('white-collar', 1)
('with', 1)
('automation,', 2)
('cognitive', 2)
('in', 2)
('influence', 2)
('is', 2)
('jobs', 2)
('or', 2)
('potential', 2)
('studies', 2)
('tasks', 2)
('traditional', 2)
('which', 2)
('occupations', 3)
('of', 4)
('the', 5)
('and', 6)
('AI', 8)
('to', 9)
sdesaleg137@cloudshell:~$
```

Delete the Dataproc cluster and the bucket:

1. To delete a Dataproc cluster and the bucket we created on GCP, you can follow these steps:

✓ Open the Google Cloud Console.
✓ Navigate to the Dataproc Clusters page in the console.
✓ Select the cluster you wish to delete.
✓ Click the 'Delete' button at the top of the page.



Failed to validate permissions required for default service account: '783029461235-compute@developer.gserviceaccount.com'. Cluster creation could respective service accounts as mentioned in the document https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-accounts Manager API hasn't been enabled in your project '783029461235' before or it is disabled. Enable it by visiting 'https://console.developers.google.com/a project=783029461235'.

Name	cluster-3b08
Cluster UUID	c55a5616-6eda-4ada-8881-0cfd2ad9d432
Type	Dataproc Cluster
Status	✓ Running

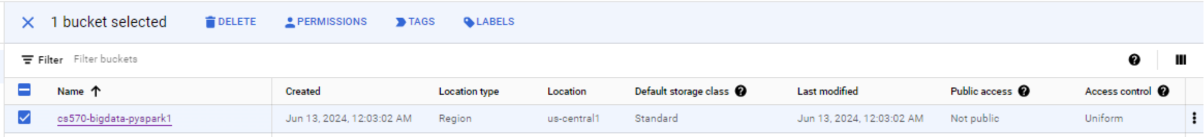
2. In the confirmation dialog box, click on ‘confirm’.

3. Navigate to the Cloud Storage page.

Name: Sened Desalegn

4. Select the bucket you want to delete.

5. Click on the ‘Delete’ button at the top of the page.



Name ↑	Created	Location type	Location	Default storage class	Last modified	Public access	Access control
☑ cs570-bigdata-pyspark1	Jun 13, 2024, 12:03:02 AM	Region	us-central1	Standard	Jun 13, 2024, 12:03:02 AM	Not public	Uniform

6. In the confirmation dialog box, type ‘DELETE’.

Reference

1. https://www.youtube.com/watch?v=_lwrfxE2RtE&ab_channel=SkillCurb
2. https://hc.labnet.sfbu.edu/~henry/npu/classes/cloud_computing/pyspark/hw/q2/2023_summer/CS570_week4_q1_19744_SriVardhan_Kotturu.pdf