Calculating Pi with PySpark

https://hc.labnet.sfbu.edu/~henry/npu/classes/learning_spark/key_value_pair/slide/exercise_key_value_pair.html

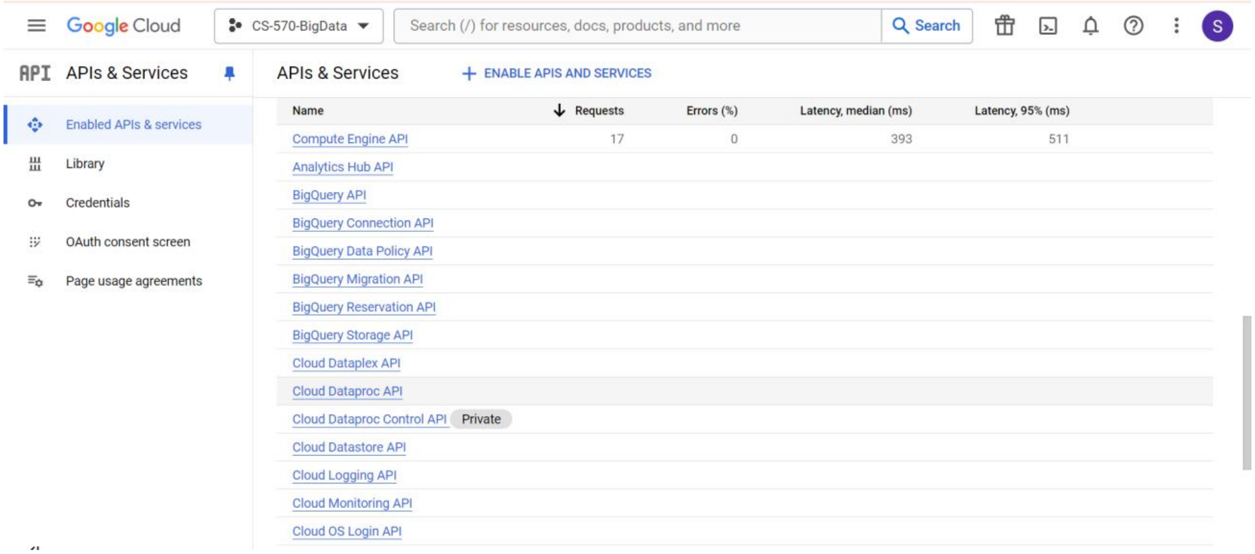
Q24 ==> Project: Calculating Pi

Note:

- If you have an existing cluster go to step 3
- If you have an existing cluster and bucket ready go to step 4

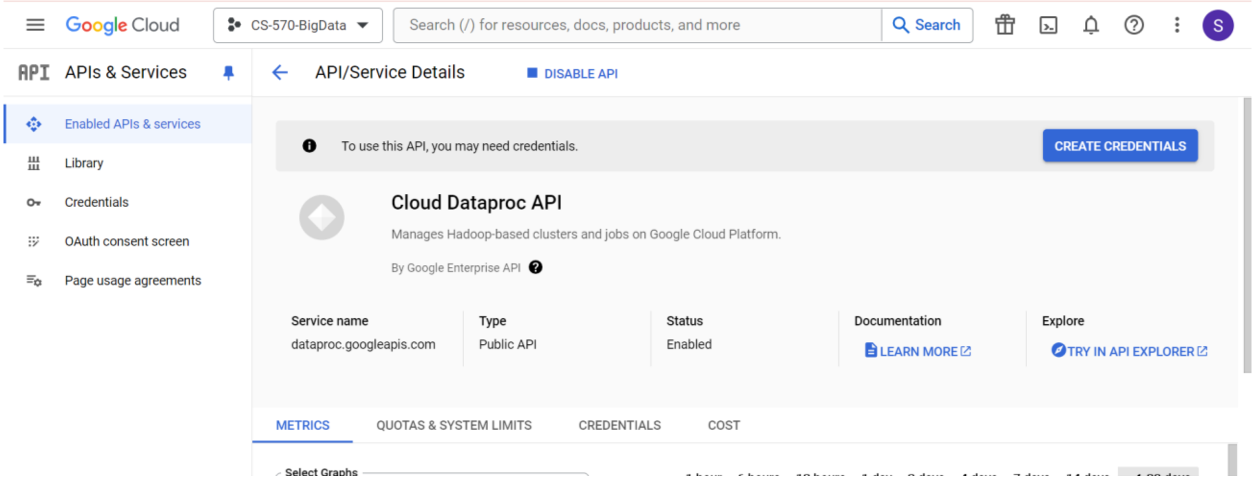
Step 1: Go to your existing vm-project

- Go to navigation menu then API and Services



Name	Requests	Errors (%)	Latency, median (ms)	Latency, 95% (ms)
Compute Engine API	17	0	393	511
Analytics Hub API				
BigQuery API				
BigQuery Connection API				
BigQuery Data Policy API				
BigQuery Migration API				
BigQuery Reservation API				
BigQuery Storage API				
Cloud Dataplex API				
Cloud Dataproc API				
Cloud Dataproc Control API Private				
Cloud Datastore API				
Cloud Logging API				
Cloud Monitoring API				
Cloud OS Login API				

- Enable the Cloud Dataproc API



To use this API, you may need credentials.

Cloud Dataproc API

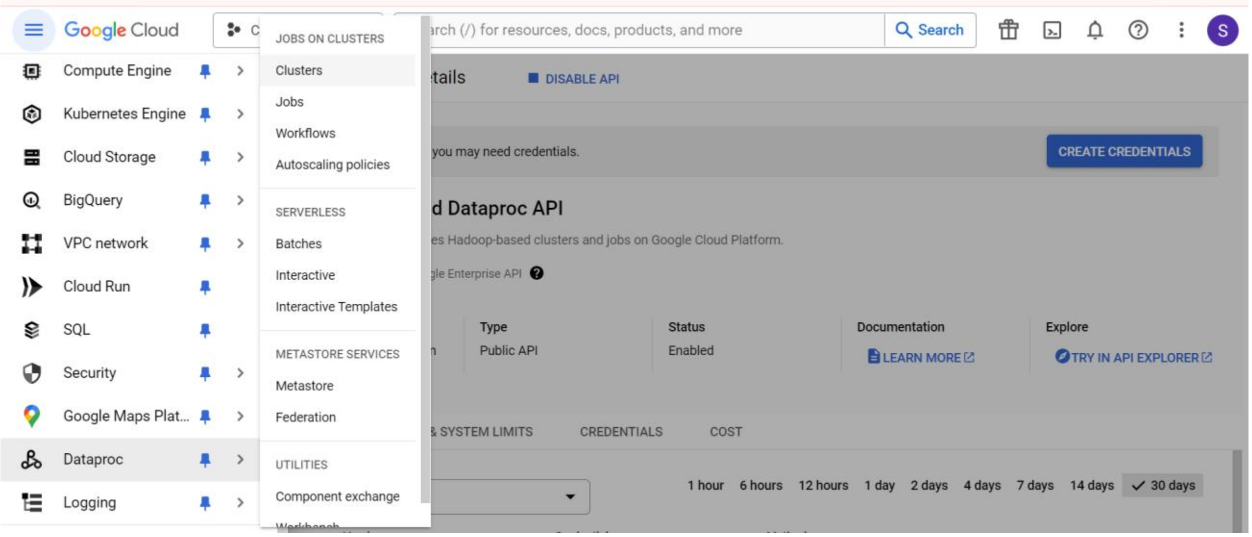
Manages Hadoop-based clusters and jobs on Google Cloud Platform.

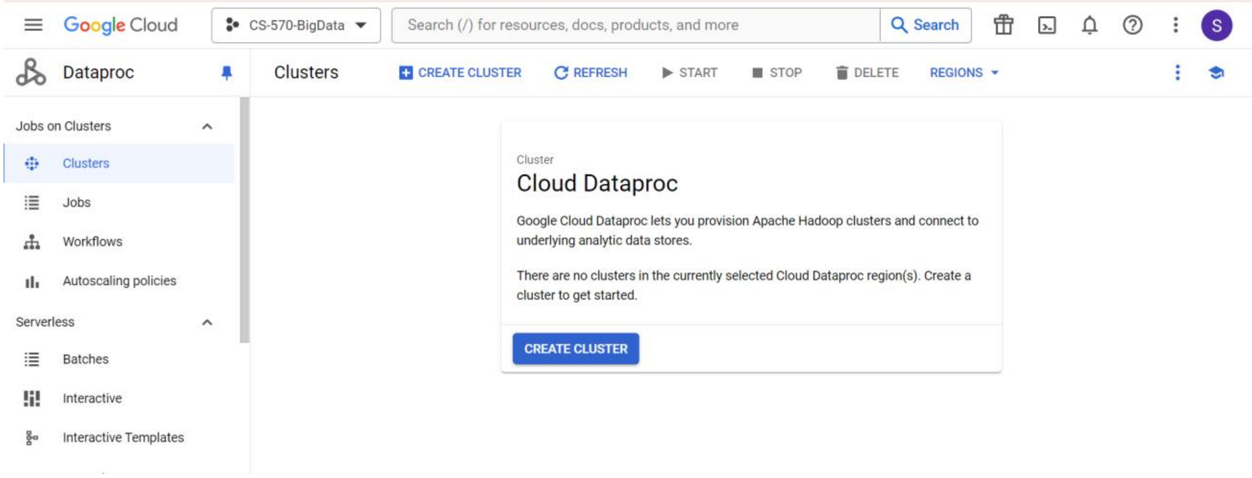
By Google Enterprise API

Service name	Type	Status	Documentation	Explore
dataproc.googleapis.com	Public API	Enabled	LEARN MORE	TRY IN API EXPLORER

Step 2: Create Cluster

- Go to Dataproc on navigation menu
- Create a Dataproc cluster



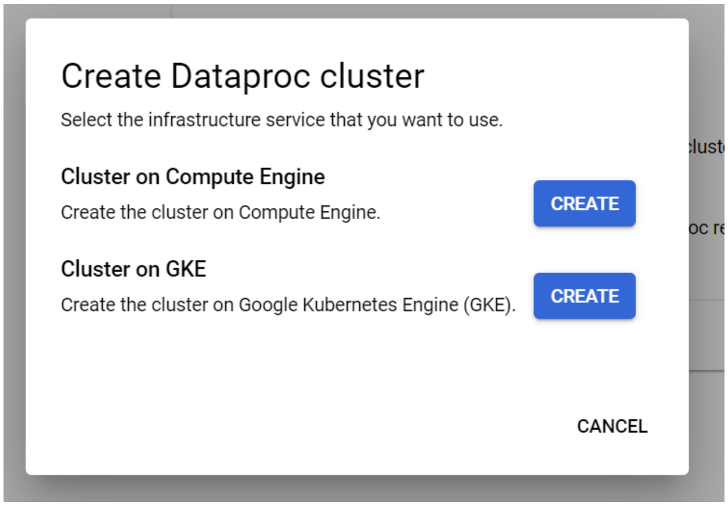


Cluster

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.



Create Dataproc cluster

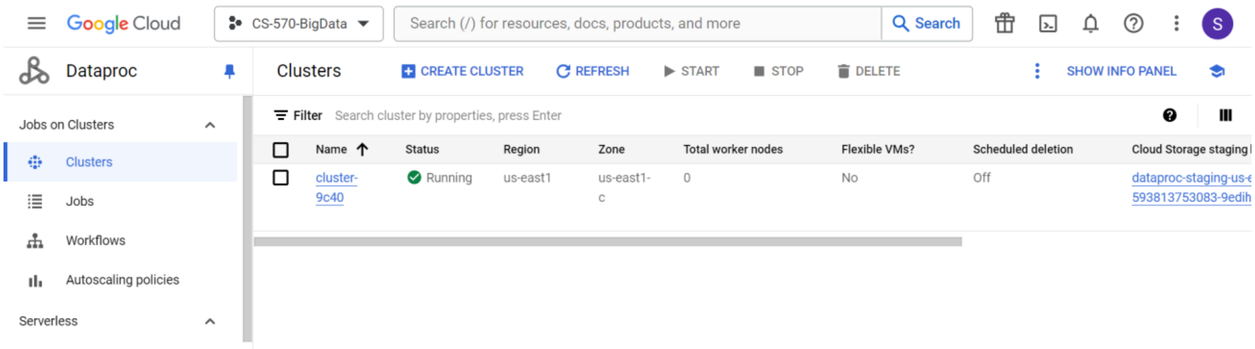
Select the infrastructure service that you want to use.

Cluster on Compute Engine

Create the cluster on Compute Engine.

Cluster on GKE

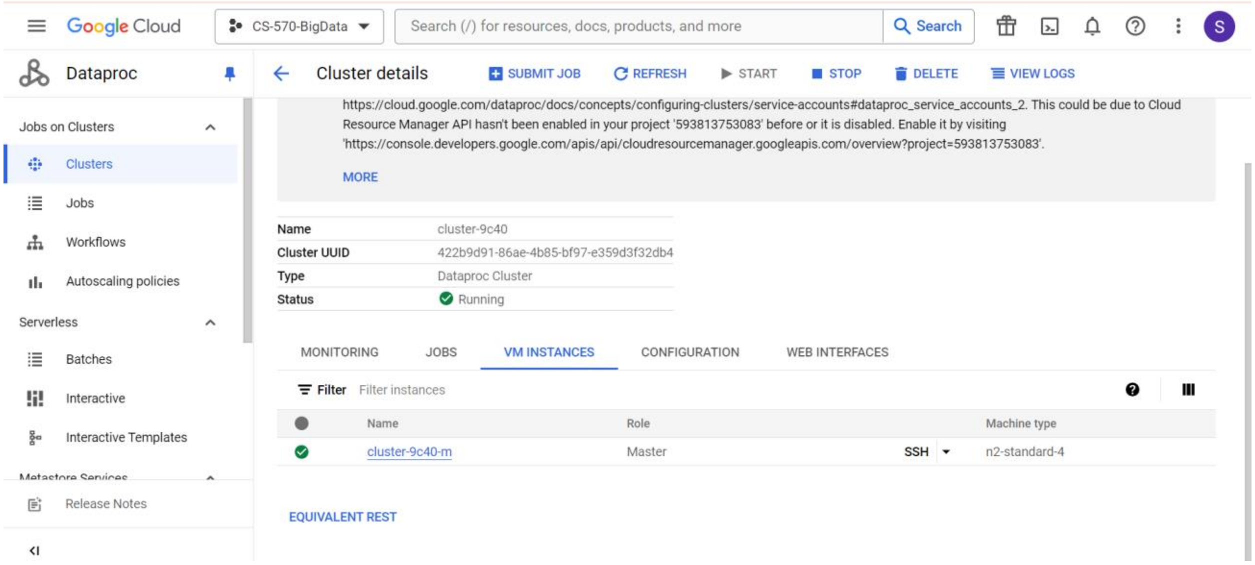
Create the cluster on Google Kubernetes Engine (GKE).



☐	Name ↑	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging
☐	cluster-9c40	✓ Running	us-east1	us-east1-c	0	No	Off	dataproc-staging-us-e 593813753083-9edih

Now our cluster is created we proceed with logging in our ssh-browser the Dataproc master server.

Click on the cluster then go to VM INSTANCES

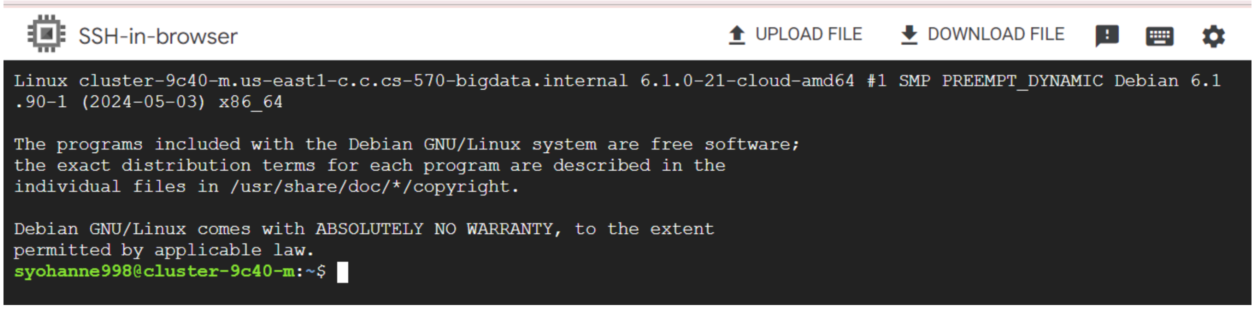


https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-accounts#dataproc_service_accounts_2. This could be due to Cloud Resource Manager API hasn't been enabled in your project '593813753083' before or it is disabled. Enable it by visiting 'https://console.developers.google.com/apis/api/cloudresourcemanager.googleapis.com/overview?project=593813753083'.

Name	cluster-9c40
Cluster UUID	422b9d91-86ae-4b85-bf97-e359d3f32db4
Type	Dataproc Cluster
Status	✓ Running

●	Name	Role		Machine type
✓	cluster-9c40-m	Master	SSH ▾	n2-standard-4

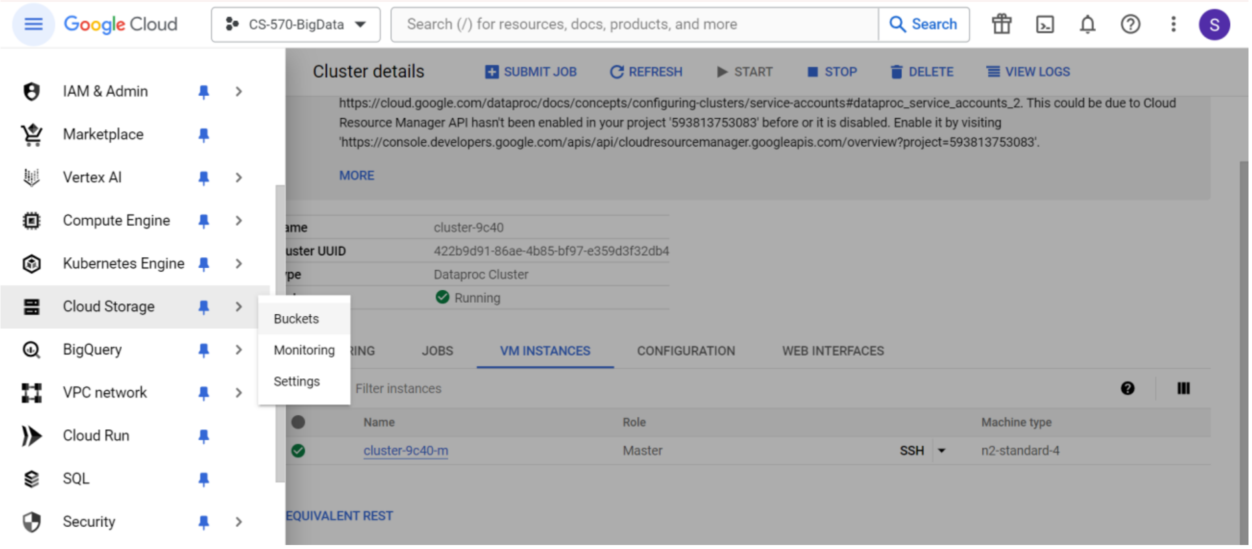
Click on the ssh of the cluster to start working on the ssh-browser



```
Linux cluster-9c40-m.us-east1-c.c.cs-570-bigdata.internal 6.1.0-21-cloud-amd64 #1 SMP PREEMPT_DYNAMIC Debian 6.1
.90-1 (2024-05-03) 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.
syohanne998@cluster-9c40-m:~$
```

Step 3: Create Bucket

Create bucket to store input and output files

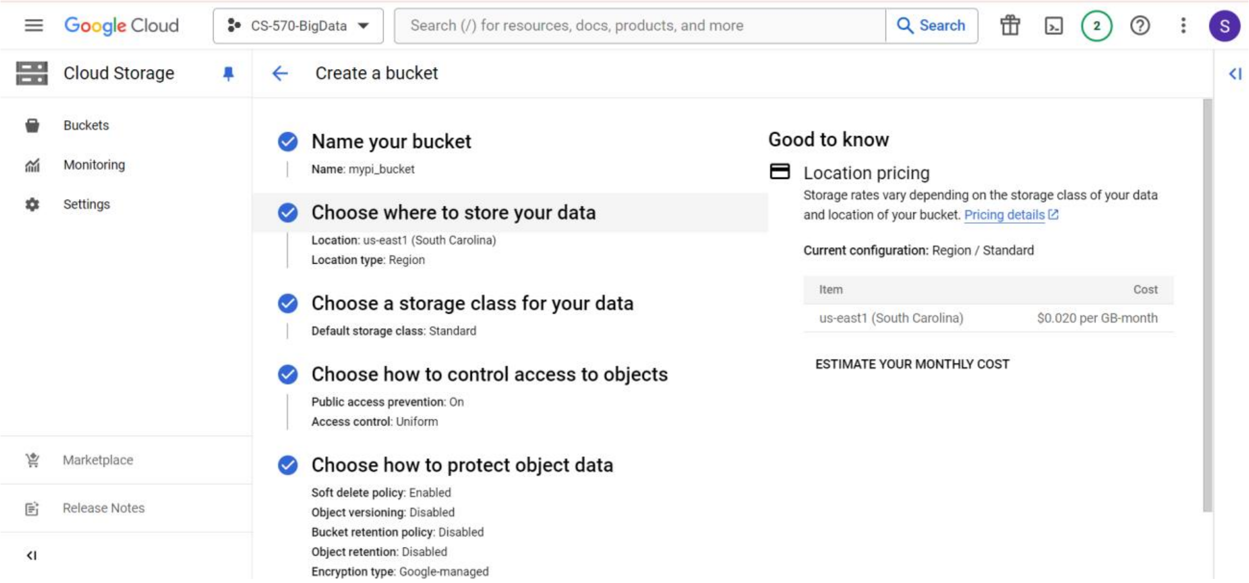


https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-accounts#dataproc_service_accounts_2. This could be due to Cloud Resource Manager API hasn't been enabled in your project '593813753083' before or it is disabled. Enable it by visiting 'https://console.developers.google.com/apis/api/cloudresourcemanager.googleapis.com/overview?project=593813753083'.

ame	cluster-9c40
uster UUID	422b9d91-86ae-4b85-bf97-e359d3f32db4
pe	Dataproc Cluster
	✓ Running

	Name	Role	Machine type
✓	cluster-9c40-m	Master	n2-standard-4

Configure the bucket name, location and storage



Name your bucket

Name: mypi_bucket

Choose where to store your data

Location: us-east1 (South Carolina)
Location type: Region

Choose a storage class for your data

Default storage class: Standard

Choose how to control access to objects

Public access prevention: On
Access control: Uniform

Choose how to protect object data

Soft delete policy: Enabled
Object versioning: Disabled
Bucket retention policy: Disabled
Object retention: Disabled
Encryption type: Google-managed

Good to know

Location pricing

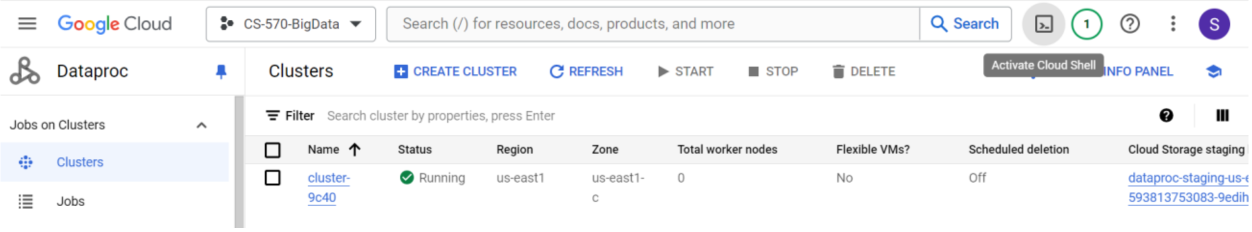
Storage rates vary depending on the storage class of your data and location of your bucket. Pricing details

Current configuration: Region / Standard

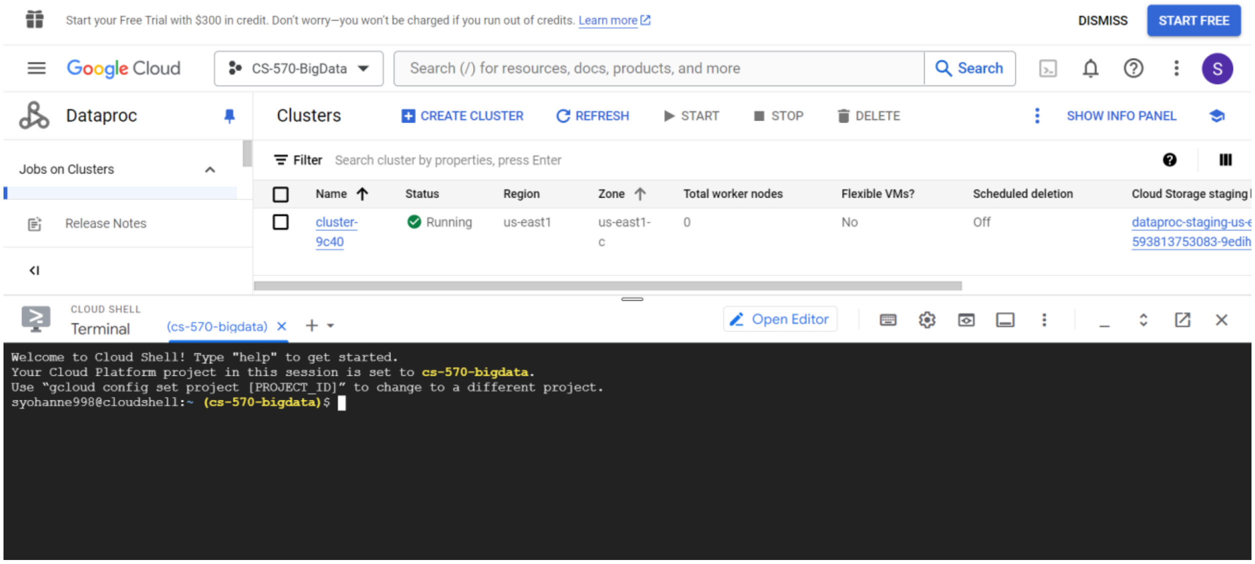
Item	Cost
us-east1 (South Carolina)	$0.020 per GB-month

Step 4: Create Pi_calculate code file

Now go to the cloud shell and create the pi_calculate.py file.



	Name	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging
□	cluster-9c40	✓ Running	us-east1	us-east1-c	0	No	Off	dataproc-staging-us-e 593813753083-9edih



Start your Free Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more

□	Name ↑	Status	Region	Zone ↑	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging
□	cluster-9c40	✓ Running	us-east1	us-east1-c	0	No	Off	dataproc-staging-us-e 593813753083-9edih

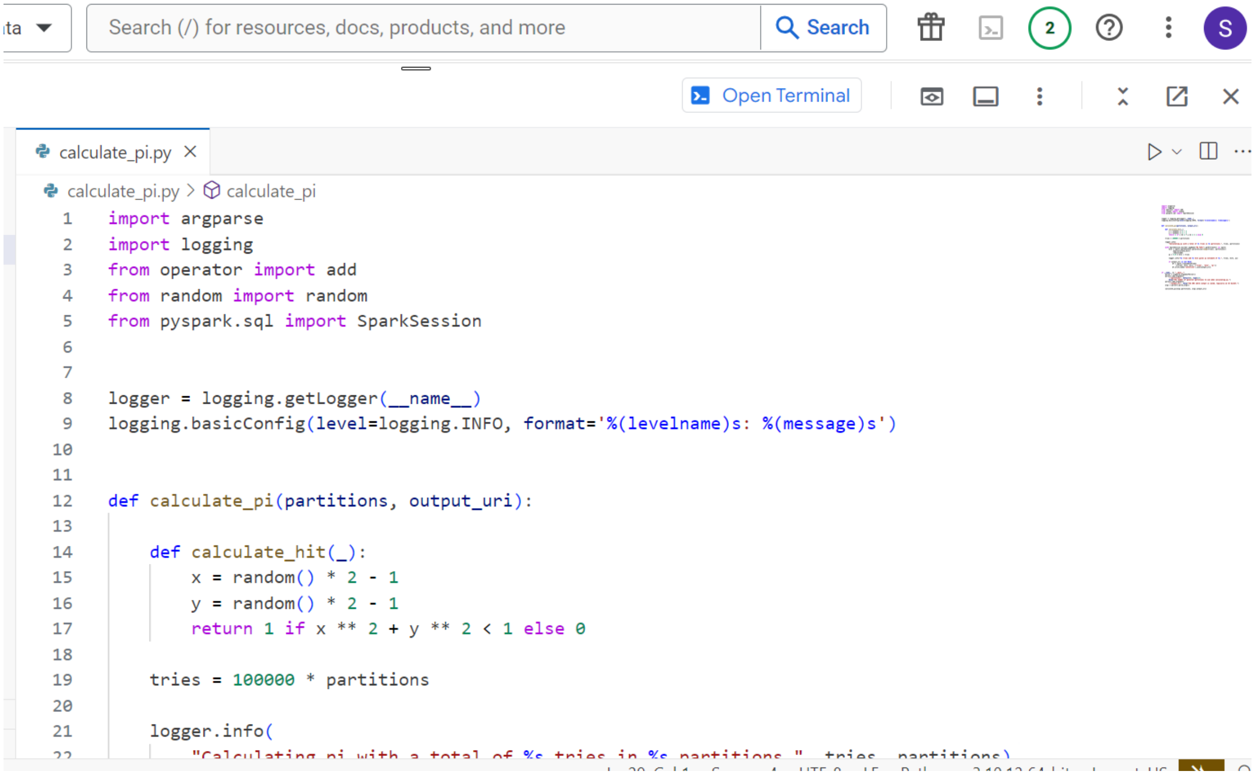
```
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to cs-570-bigdata.
Use "gcloud config set project [PROJECT_ID]" to change to a different project.
syohanne998@cloudshell:~ (cs-570-bigdata)$
```

Go to open Editor and write the following code

....

.

```
import argparse
import logging
from operator import add
from random import random
from pyspark.sql import SparkSession

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO, format='%(levelname)s:
%(message)s')

def calculate_pi(partitions, output_uri):
    def calculate_hit(_):
        x = random() * 2 - 1
        y = random() * 2 - 1
        return 1 if x ** 2 + y ** 2 < 1 else 0
    tries = 100000 * partitions
    logger.info(
        "Calculating pi with a total of %s tries in %s partitions.",
tries, partitions)
    with SparkSession.builder.appName("My PyPi").getOrCreate() as spark:
        hits = spark.sparkContext.parallelize(range(tries), partitions)\
            .map(calculate_hit)\
                .reduce(add)
        pi = 4.0 * hits / tries
        logger.info("%s tries and %s hits gives pi estimate of %s.", tries, hits, pi)
        if output_uri is not None:
            df = spark.createDataFrame(
                [(tries, hits, pi)], ['tries', 'hits', 'pi'])
            df.write.mode('overwrite').json(output_uri)
if __name__ == "__main__":
    parser = argparse.ArgumentParser()
    parsers.add_argument(
        '--partitions', default=2, type=int,
        help="The number of parallel partitions to use when calculating pi.")
    parsers.add_argument(
        '--output_uri', help="The URI where output is saved, typically an S3 bucket.")
    args = parsers.parse_args()

    calculate_pi(args.partitions, args.output_uri)
```

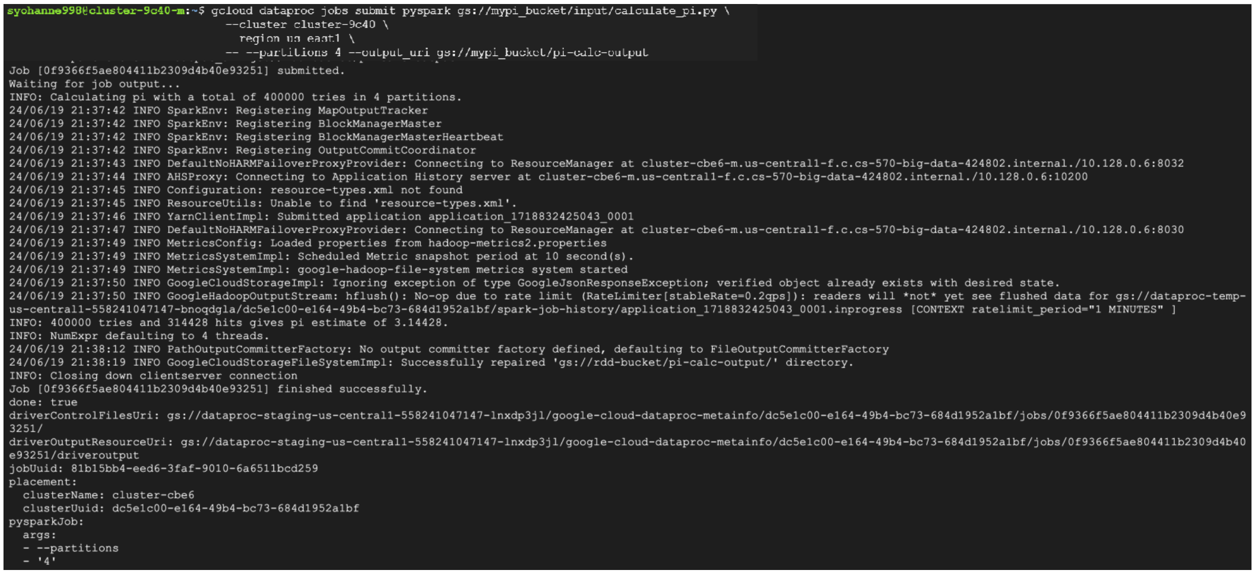

ta
Search (/) for resources, docs, products, and more
Search
2
S
Open Terminal
calculate_pi.py
calculate_pi.py > calculate_pi

```
import argparse
import logging
from operator import add
from random import random
from pyspark.sql import SparkSession


logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO, format='%(levelname)s: %(message)s')


def calculate_pi(partitions, output_uri):

    def calculate_hit(_):
        x = random() * 2 - 1
        y = random() * 2 - 1
        return 1 if x ** 2 + y ** 2 < 1 else 0

    tries = 100000 * partitions

    logger.info(
        "Calculating pi with a total of %s tries in %s partitions.", tries, partitions)
```

Step 5: Run the code

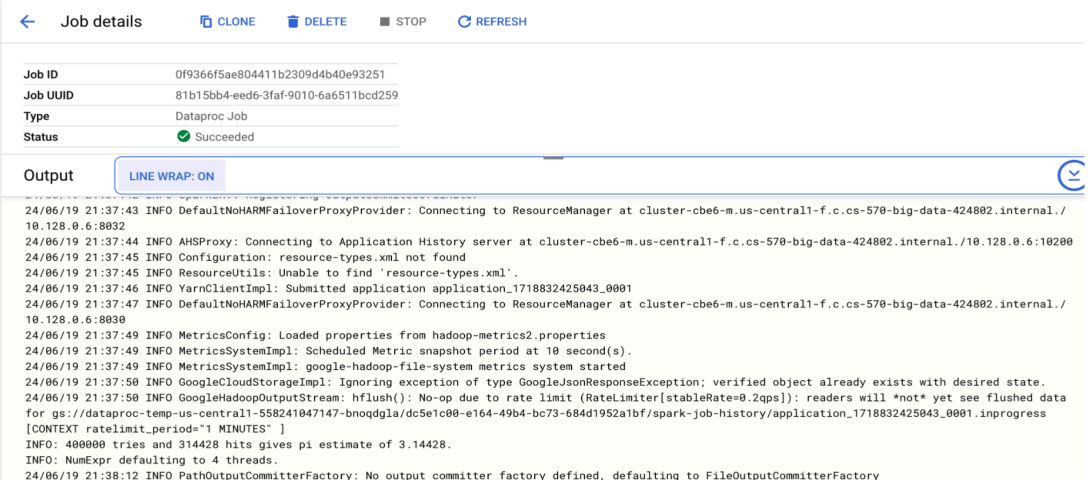
Now click on the cluster and go to vm instances
Then click on the SHH to open SHH-browser

```
gcloud dataproc jobs submit pyspark gs://mypi_bucket/input/calculate_pi.py \
                         --cluster cluster-9c40 \
                         --region us-east1 \
                         -- --partitions 4 --output_uri gs://mypi_bucket/pi-calc-output
```



```
syohanne998@cluster-9c40-m:~$ gcloud dataproc jobs submit pyspark gs://mypi_bucket/input/calculate_pi.py \
                              --cluster cluster-9c40 \
                                region us east1 \
                              -- --partitions 4 --output_uri gs://mypi_bucket/pi-calc-output
Job [0f9366f5ae804411b2309d4b40e93251] submitted.
Waiting for job output...
INFO: Calculating pi with a total of 400000 tries in 4 partitions.
24/06/19 21:37:42 INFO SparkEnv: Registering MapOutputTracker
24/06/19 21:37:42 INFO SparkEnv: Registering BlockManagerMaster
24/06/19 21:37:42 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/06/19 21:37:42 INFO SparkEnv: Registering OutputCommitCoordinator
24/06/19 21:37:43 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-cbe6-m.us-central1-f.c.cs-570-big-data-424802.internal./10.128.0.6:8032
24/06/19 21:37:44 INFO AHSProxy: Connecting to Application History server at cluster-cbe6-m.us-central1-f.c.cs-570-big-data-424802.internal./10.128.0.6:10200
24/06/19 21:37:45 INFO Configuration: resource-types.xml not found
24/06/19 21:37:45 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/06/19 21:37:46 INFO YarnClientImpl: Submitted application application_1718832425043_0001
24/06/19 21:37:47 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-cbe6-m.us-central1-f.c.cs-570-big-data-424802.internal./10.128.0.6:8030
24/06/19 21:37:49 INFO MetricsConfig: Loaded properties from hadoop-metrics2.properties
24/06/19 21:37:49 INFO MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
24/06/19 21:37:49 INFO MetricsSystemImpl: google-hadoop-file-system metrics system started
24/06/19 21:37:50 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
24/06/19 21:37:50 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for gs://dataproc-temp-
us-central1-558241047147-bnoqdgla/dc5e1c00-e164-49b4-bc73-684d1952a1bf/spark-job-history/application_1718832425043_0001.inprogress [CONTEXT ratelimit_period="1 MINUTES" ]
INFO: 400000 tries and 314428 hits gives pi estimate of 3.14428.
INFO: NumExpr defaulting to 4 threads.
24/06/19 21:38:12 INFO PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
24/06/19 21:38:19 INFO GoogleCloudStorageFileSystemImpl: Successfully repaired 'gs://rdd-bucket/pi-calc-output/' directory.
INFO: Closing down clientserver connection
Job [0f9366f5ae804411b2309d4b40e93251] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-central1-558241047147-lnxdp3jl/google-cloud-dataproc-metainfo/dc5e1c00-e164-49b4-bc73-684d1952a1bf/jobs/0f9366f5ae804411b2309d4b40e9
3251/
driverOutputResourceUri: gs://dataproc-staging-us-central1-558241047147-lnxdp3jl/google-cloud-dataproc-metainfo/dc5e1c00-e164-49b4-bc73-684d1952a1bf/jobs/0f9366f5ae804411b2309d4b40
e93251/driveroutput
jobUuid: 81b15bb4-eed6-3faf-9010-6a6511bcd259
placement:
  clusterName: cluster-cbe6
  clusterUuid: dc5e1c00-e164-49b4-bc73-684d1952a1bf
pysparkJob:
  args:
  - --partitions
  - '4'
```

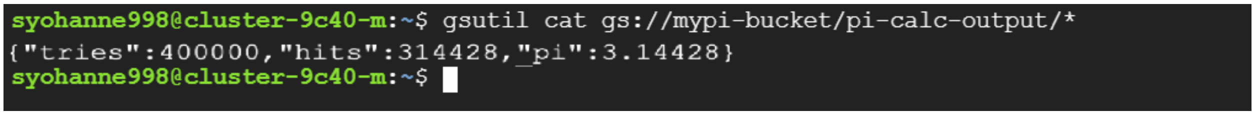
Now go to job details and turn the line warp on



Job ID	0f9366f5ae804411b2309d4b40e93251
Job UUID	81b15bb4-eed6-3faf-9010-6a6511bcd259
Type	Dataproc Job
Status	✓ Succeeded

```
24/06/19 21:37:43 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-cbe6-m.us-central1-f.c.cs-570-big-data-424802.internal./
10.128.0.6:8032
24/06/19 21:37:44 INFO AHSProxy: Connecting to Application History server at cluster-cbe6-m.us-central1-f.c.cs-570-big-data-424802.internal./10.128.0.6:10200
24/06/19 21:37:45 INFO Configuration: resource-types.xml not found
24/06/19 21:37:45 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/06/19 21:37:46 INFO YarnClientImpl: Submitted application application_1718832425043_0001
24/06/19 21:37:47 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at cluster-cbe6-m.us-central1-f.c.cs-570-big-data-424802.internal./
10.128.0.6:8030
24/06/19 21:37:49 INFO MetricsConfig: Loaded properties from hadoop-metrics2.properties
24/06/19 21:37:49 INFO MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
24/06/19 21:37:49 INFO MetricsSystemImpl: google-hadoop-file-system metrics system started
24/06/19 21:37:50 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
24/06/19 21:37:50 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data
for gs://dataproc-temp-us-central1-558241047147-bnoqdgla/dc5e1c00-e164-49b4-bc73-684d1952a1bf/spark-job-history/application_1718832425043_0001.inprogress
[CONTEXT ratelimit_period="1 MINUTES" ]
INFO: 400000 tries and 314428 hits gives pi estimate of 3.14428.
INFO: NumExpr defaulting to 4 threads.
24/06/19 21:38:12 INFO PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
```

Viewing the output of pi value

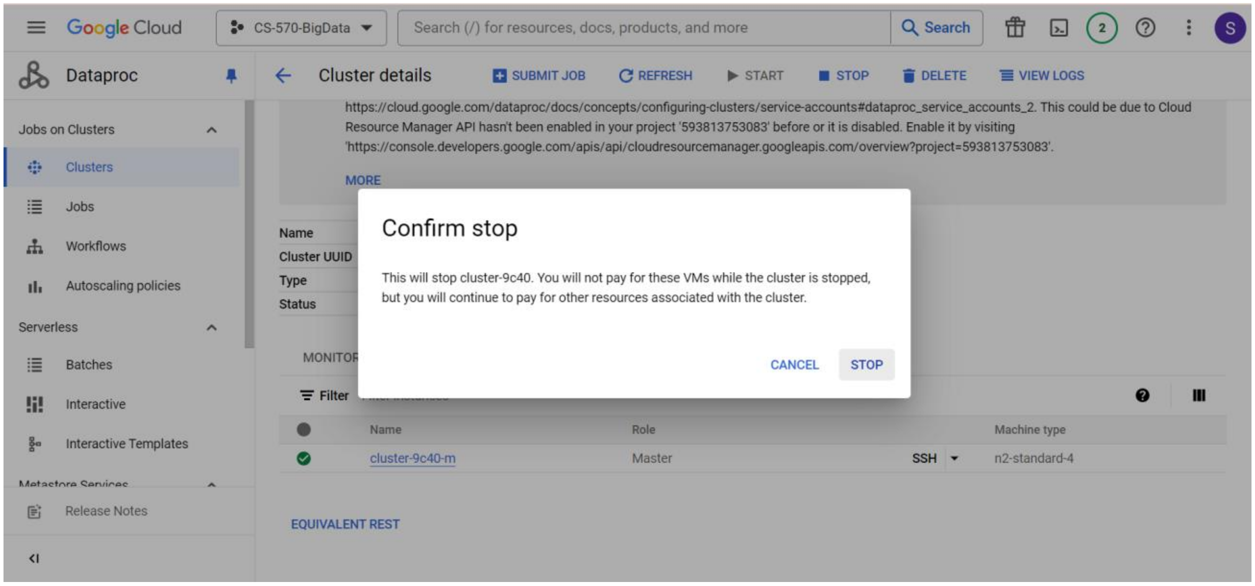


```
syohanne998@cluster-9c40-m:~$ gsutil cat gs://mypi-bucket/pi-calc-output/*
{"tries":400000,"hits":314428,"pi":3.14428}
syohanne998@cluster-9c40-m:~$ 
```

Step 6: Close or delete the cluster and the bucket

Remove bucket and close cluster

I have no use of the bucket and I have stopped my cluster instead of deleting it but you can also delete the cluster if you won't use.

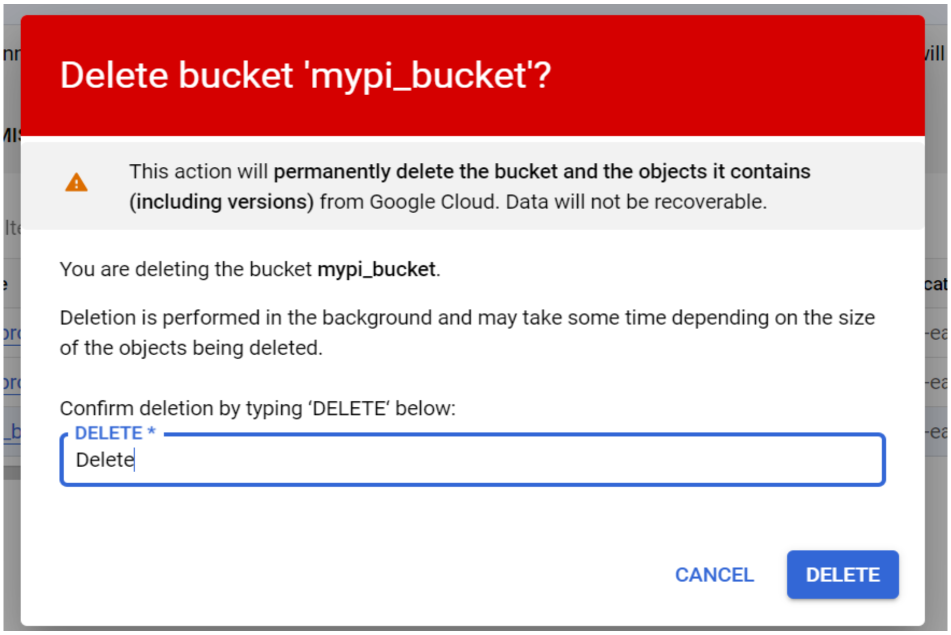


https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-accounts#dataproc_service_accounts_2. This could be due to Cloud Resource Manager API hasn't been enabled in your project '593813753083' before or it is disabled. Enable it by visiting 'https://console.developers.google.com/apis/api/cloudresourcemanager.googleapis.com/overview?project=593813753083'.

Confirm stop

This will stop cluster-9c40. You will not pay for these VMs while the cluster is stopped, but you will continue to pay for other resources associated with the cluster.

CANCEL STOP



Delete bucket 'mypi_bucket'?

This action will **permanently delete the bucket and the objects it contains (including versions)** from Google Cloud. Data will not be recoverable.

You are deleting the bucket **mypi_bucket**.

Deletion is performed in the background and may take some time depending on the size of the objects being deleted.

Confirm deletion by typing 'DELETE' below:

DELETE *

Delete

CANCEL DELETE