

Instructions for using the Google Cloud Platform for TP2

Dear students, you will find below the instructions on how to use the Google Cloud Platform (GCP) required for the last part of TP2.

1. Obtaining GCP credits

Here is the URL you will need to access in order to request a Google Cloud Platform coupon. You will be asked to provide your school email address and name. An email will be sent to you to confirm these details before a coupon is sent to you.

Student Coupon Retrieval Link

- You will be asked for a name and email address, which needs to match your school domain. A confirmation email will be sent to you with a coupon code.
- You can only request ONE code per unique email address.

Please contact me if you have any questions or issues.

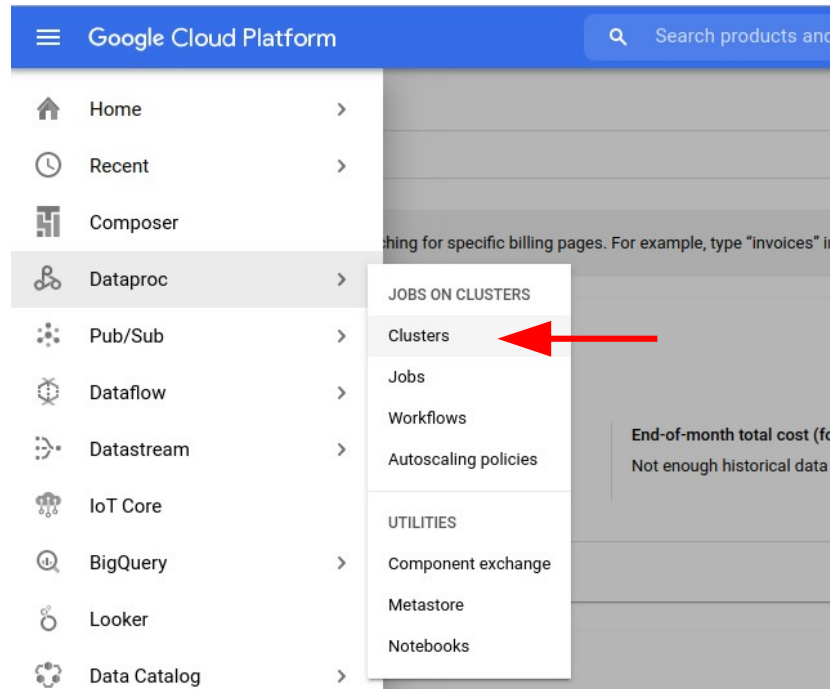
Once you have completed this step, you can see your billing account details with your credits of 50\$ from the following link:

<https://console.cloud.google.com/billing>

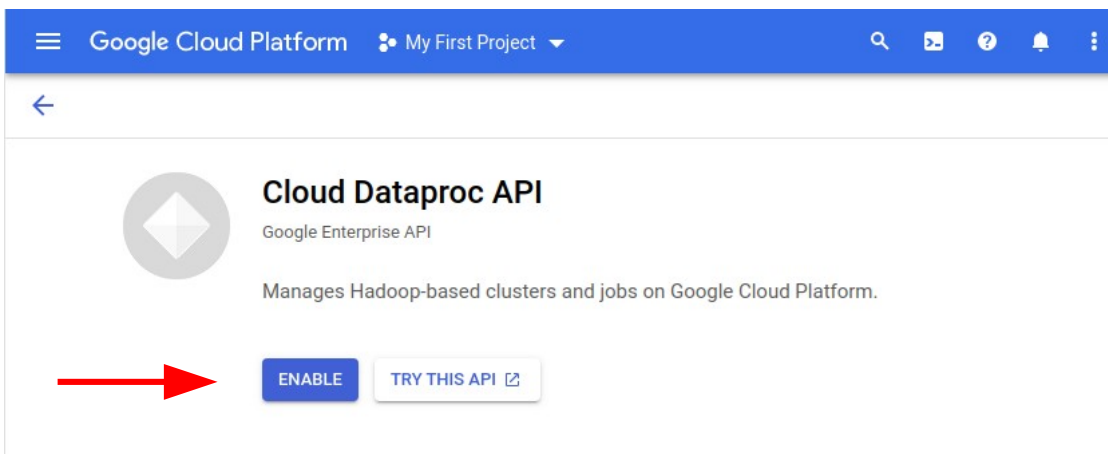
2. Enabling the required APIs

To run our MBA algorithm, we will use the Dataproc service. However, first we need to Enable the APIs.

On your console, click on the 3 lines on the top left and search for **Dataproc -> Clusters** (Dataproc is located under the category of **big data**).



Next, click on Enable API *. This process can take a few minutes.

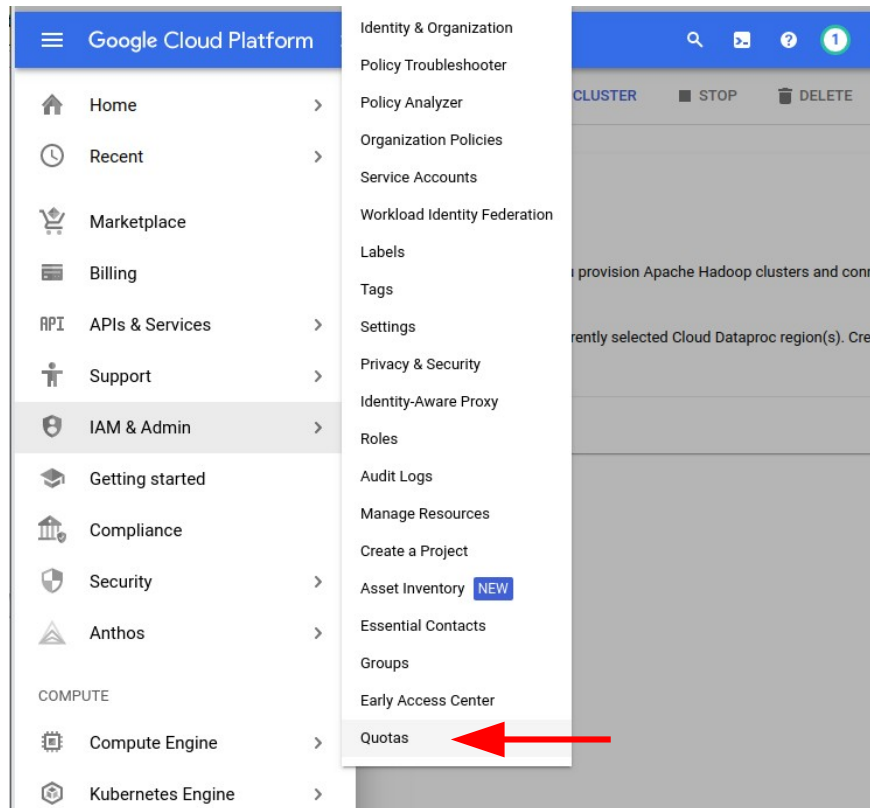


* this may be triggered automatically the first time you access this page.

3. Requiring for more CPU cluster capacity

By default, the maximum number of CPUs allowed by GCP for this student credit account is 24, but we will need much more than that.

On your console, click on the 3 lines on the top left and search for **IAM & admin -> Quotas**



Once there, among the “Compute Engine API” services, look for the limit name of “CPUs (all regions)”. For more convenience, you can sort the table by limit names. Select it and click on Edit Quotas.

The image shows the Google Cloud Platform 'Quotas' page. The table lists various quotas for the Compute Engine API. The 'CPUs (all regions)' quota is selected, and the right panel shows its details.

API	Limit	Status	Action
Compute Engine API	CPUs	✓ All 115 quotas are within limit	ALL QUOTAS
Compute Engine API	CPUs (all regions)	✓ One quota is within limit	ALL QUOTAS
Cloud Datastore API	Create or Delete Index Requests Per Minute	✓ One quota is within limit	ALL QUOTAS
BigQuery Storage API	CreateWriteStream requests per minute	✓ One quota is within limit	ALL QUOTAS
Compute Engine API	Cross Project Networking Service	✓ One quota is within limit	ALL QUOTAS

CPUs (all regions)

Service: **Compute Engine API**
Category: **Location**


☒ Global

Limit: 32

Current usage: 0 (0%)

7 days peak: 0 (0%)


Then enter **128** in the **new limit** field and write something similar to the one showing the image below in the description box.

 1 quota selected

Quota changes
Expand each service card to change individual quotas.

Compute Engine API

Quota: CPUs (all regions)

Current limit: 32
Enter a new quota limit. Your request will be sent to your service provider for approval. 

New limit *
300

Request description *
I am working on an academic project that demands a large cluster to run the application.
Your description will be sent to your service provider and is used to evaluate your request. It's useful to include the intent of the quota usage, future growth plans, region or zone spread, and any additional requirements or dependencies.

DONE

NEXT

You will receive an email confirming your request. Your request usually takes between 30 minutes and 48 hours to be processed.

Then, among the “Compute Engine API” services, look for the limit name of “us-east1” and ask for a new quota limit (for example 128).

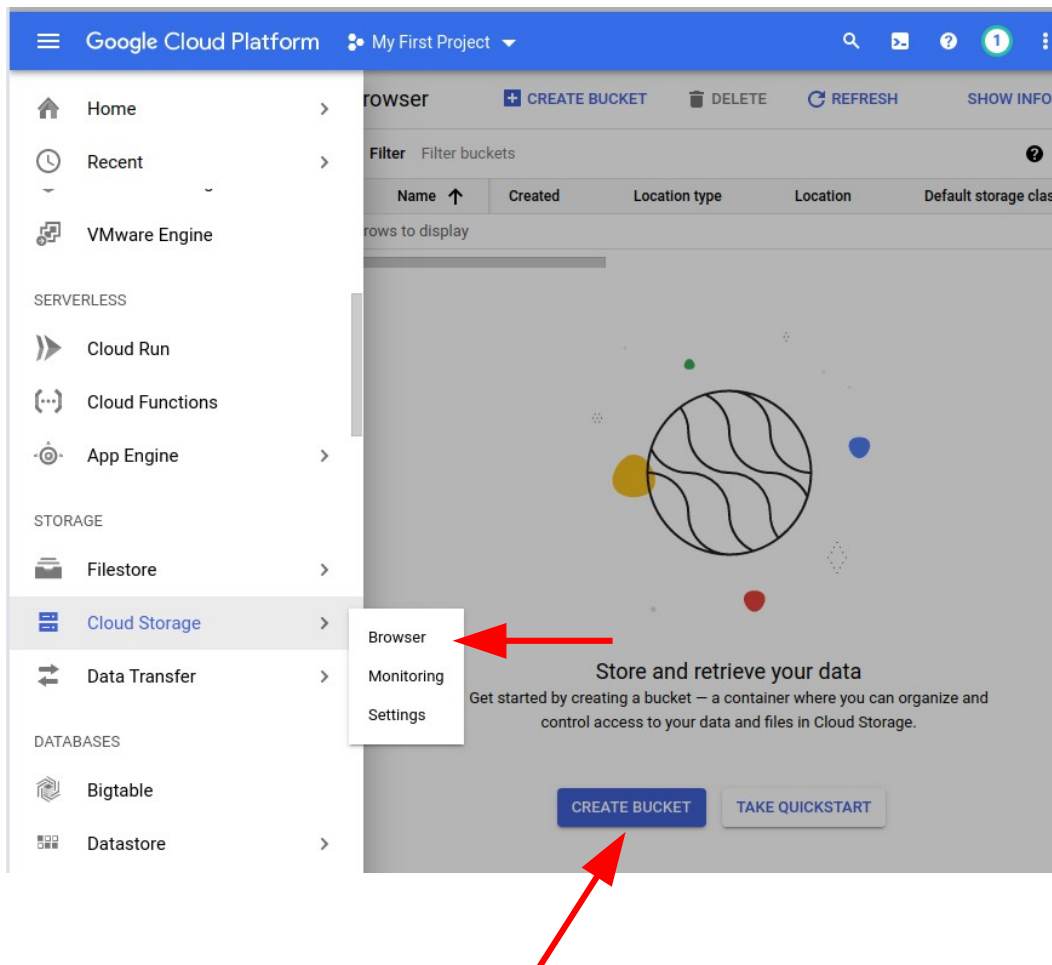
Note: Your request for CPUs (all regions) may be rejected and you may receive a message similar to Rear's message:

“Unfortunately, we are unable to grant you additional quota at this time. If this is a new project please wait 48h until you resubmit the request or until your Billing account has additional history.”

In this case, try to split your desired increase in quota into some steps that could be approved like 32 -> 64 -> 128. In this way, you can acquire 128 CPUs (all regions) or even more.

4. Creating a storage bucket

On your console, click on the 3 lines on the top left and search for **Cloud Storage -> Browser** and click in “**Create bucket**”.



Then select a name for your bucket:

←

Create a bucket

•

Name your bucket

Pick a globally unique, permanent name. [Naming guidelines](#)

my_bucket_tp

Tip: Don't include any sensitive information

CONTINUE

•

Choose where to store your data

•

Choose a default storage class for your data

•

Choose how to control access to objects

•

Advanced settings (optional)

Under the “Choose where to store your data”, select “**Region**” and search for “**us-east1**”. Also set the access to objects as “**uniform**” and press “**Create**”.

✓

Name your bucket

•

Choose where to store your data

This permanent choice defines the geographic placement of your data and affects cost, performance, and availability. [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

Location

us-east1 (South Carolina)

▼

CONTINUE

•

Choose a default storage class for your data

•

Choose how to control access to objects

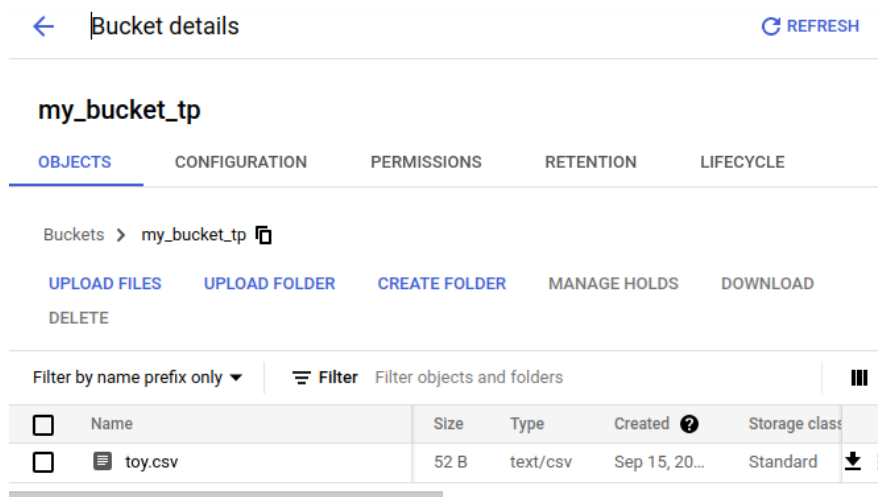
•

Advanced settings (optional)

CREATE

CANCEL

You will be redirected to your bucket page where you can upload your files. As an example, upload the **toy.csv** file to your bucket.



← Bucket details REFRESH

my_bucket_tp

OBJECTS CONFIGURATION PERMISSIONS RETENTION LIFECYCLE

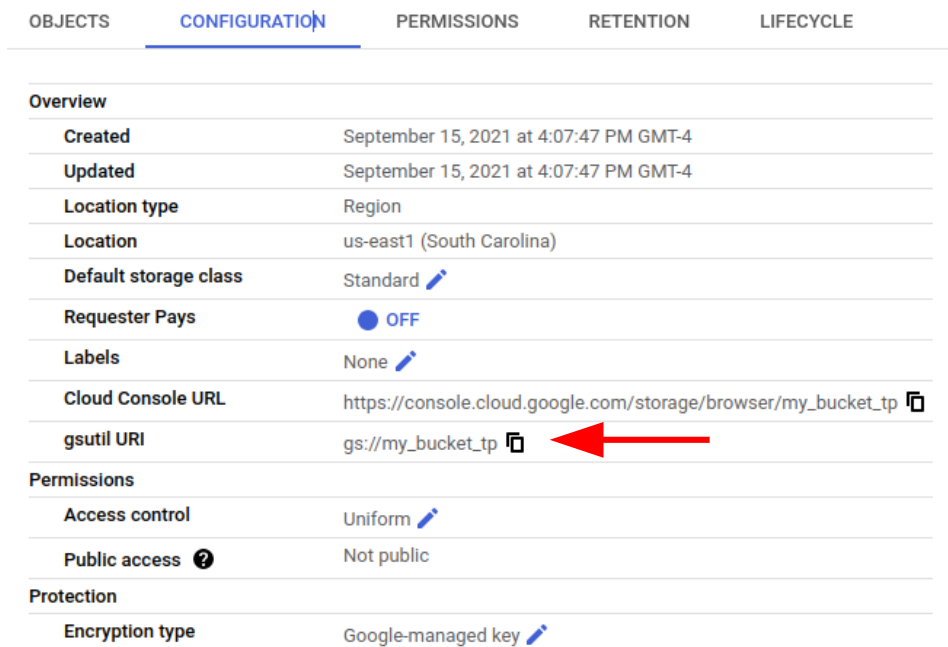
Buckets > my_bucket_tp

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS DOWNLOAD DELETE

Filter by name prefix only Filter objects and folders

<input type="checkbox"/>	Name	Size	Type	Created	Storage class
<input type="checkbox"/>	toy.csv	52 B	text/csv	Sep 15, 20...	Standard

If you go the **CONFIGURATION** tab, the **gsutil URL** gives you the address for your bucket. For example, the path to access **toy.csv** file in this sample bucket would be: "**gs://my_bucket_tp/toy.csv**".



OBJECTS CONFIGURATION PERMISSIONS RETENTION LIFECYCLE

Overview

Created	September 15, 2021 at 4:07:47 PM GMT-4
Updated	September 15, 2021 at 4:07:47 PM GMT-4
Location type	Region
Location	us-east1 (South Carolina)
Default storage class	Standard
Requester Pays	OFF
Labels	None
Cloud Console URL	https://console.cloud.google.com/storage/browser/my_bucket_tp
gsutil URI	gs://my_bucket_tp

Permissions

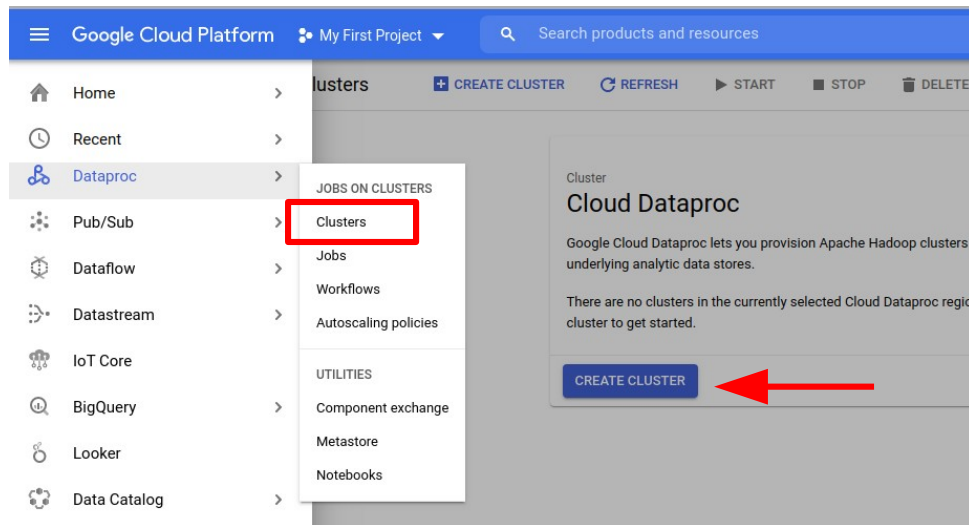
Access control	Uniform
Public access	Not public

Protection

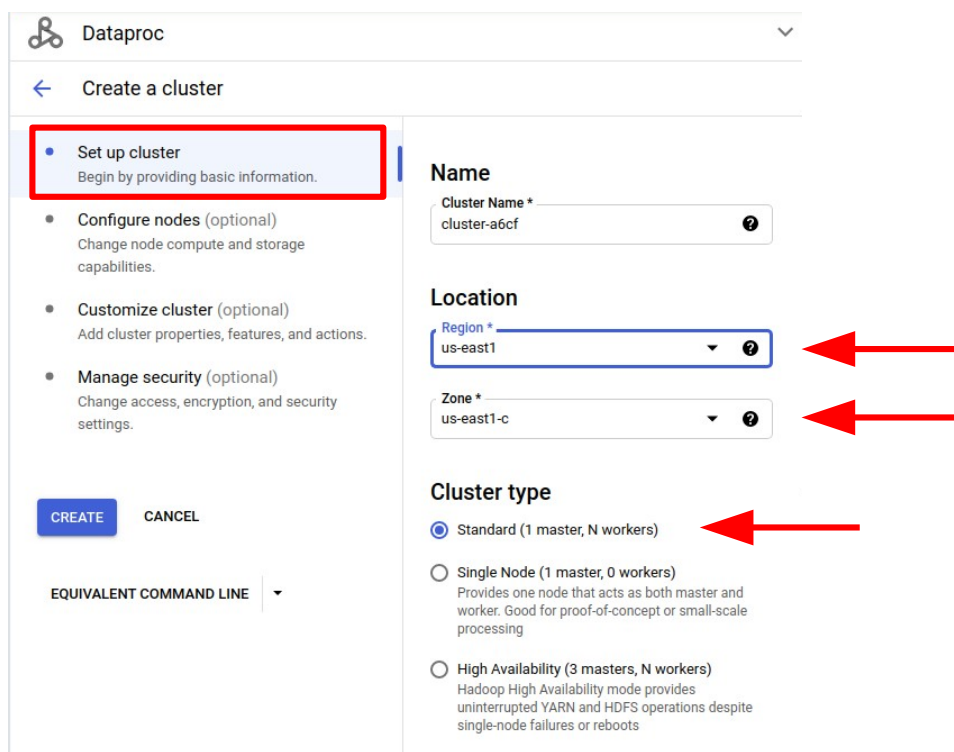
Encryption type	Google-managed key
-----------------	--------------------

5. Creating a computing cluster

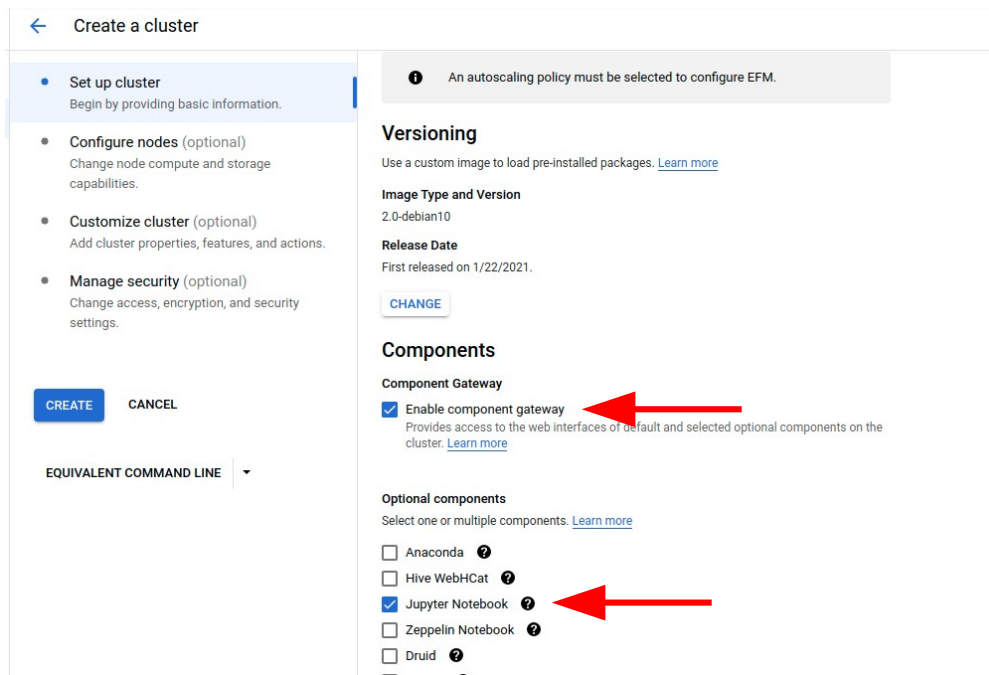
Now everything is set for creating our cluster. Go again to the **Dataproc -> Clusters** and press **Create Cluster**.



You don't need to change the name of the cluster, but it is necessary to specify the **Region**. Select **us-east1** (or the region for which you requested a quota increase). Set the **Cluster type** to **Standard (1 master, N workers)**.



VERY IMPORTANT: There is still a crucial step in the cluster configuration to be done. Select “Enable component gateway” and “Jupyter Notebook” options:



← Create a cluster

- 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 ▾

Versioning
Use a custom image to load pre-installed packages. [Learn more](#)

Image Type and Version
2.0-debian10

Release Date
First released on 1/22/2021.

CHANGE

Components

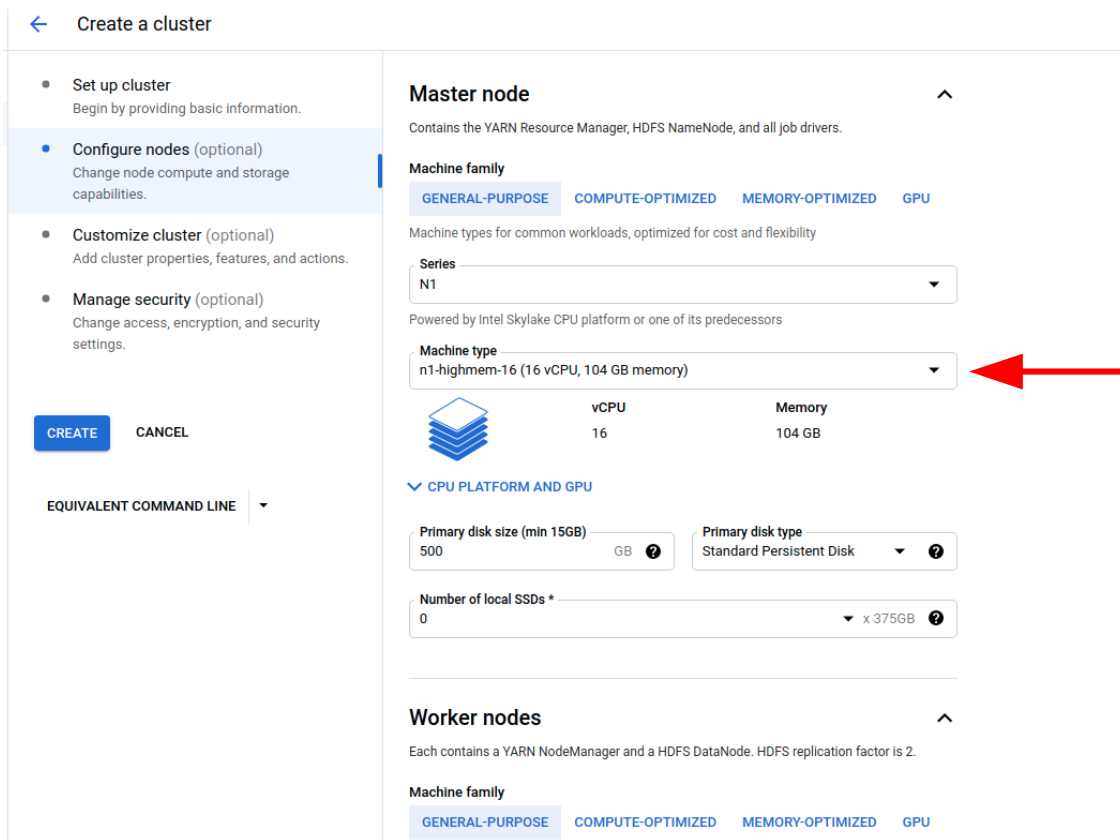
Component Gateway

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

Optional components
Select one or multiple components. [Learn more](#)

- ☐ Anaconda ?
- ☐ Hive WebHCat ?
- ☒ Jupyter Notebook ?
- ☐ Zeppelin Notebook ?
- ☐ Druid ?

Now we have to configure nodes in the the cluster. The Cluster has 1 master and N workers. In our application the most valuable resource is memory. Thus, both for the master node as for the workers nodes we will use machines from the type **highmem**.



← Create a cluster

- 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 ▾

Master node

Contains the YARN Resource Manager, HDFS NameNode, and all job drivers.

Machine family

GENERAL-PURPOSE **COMPUTE-OPTIMIZED** **MEMORY-OPTIMIZED** **GPU**

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-highmem-16 (16 vCPU, 104 GB memory)

	vCPU	Memory
	16	104 GB

CPU PLATFORM AND GPU

Primary disk size (min 15GB)
500 GB

Primary disk type
Standard Persistent Disk

Number of local SSDs *
0 x 375GB

Worker nodes

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

Machine family

GENERAL-PURPOSE **COMPUTE-OPTIMIZED** **MEMORY-OPTIMIZED** **GPU**

Suggestion: make a cluster of a master and 7 workers and select the type of n1-highmem-16 for all nodes. This cluster configuration is only a suggestion and may be advisable to try a smaller cluster in your first run. For example, you could first try to run the section 3.2 with a smaller cluster and then increase it up to this configuration for running the application in 3.3. Also, learn how to calculate the price of a cluster, which can be done [here](#). Alternatively, you can do the pricing calculation using this [application](#). There, navigate on the applications to find Cluster Dataproc and put the the cluster configuration that you want to estimate.

Continue and look for “Cloud Storage staging bucket” in “Customize cluster” and browse your bucket:


The screenshot shows the 'Customize cluster' configuration page in Google Cloud Platform. On the left sidebar, the 'Customize cluster (optional)' section is selected. The main content area is divided into several sections: 'Cluster properties', 'Initialization actions', 'Custom cluster metadata', 'Scheduled deletion', and 'Cloud Storage staging bucket'. The 'Cloud Storage staging bucket' section is at the bottom, showing a text input field with the value 'my_bucket_tp' and a 'BROWSE' button. A red arrow points to the 'BROWSE' button. Below the input field, there is a note: 'Cloud Storage staging bucket to be used for storing cluster job dependencies, job driver output, and cluster config files.'

Warning: as you finish the configuration of your cluster and press create, GCP will start charging your billing account. Always remember to delete the cluster once you have finished your experiment.

Finally, press **Create** to create the cluster. It may take a few minutes until the cluster is created and ready to be used.

6. Using your cluster

Once your cluster is created, click to open it.

Clusters								
<div><div><div></div></div><div><div>CREATE CLUSTER</div></div><div><div>REFRESH</div></div><div><div>DELETE</div></div><div><div>REGIONS</div></div></div>								
<div><div><div></div></div><div><div>Search clusters, press Enter</div></div><div><div></div></div></div>								
<input type="checkbox"/> Name ^	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created	Status	
<input type="checkbox"/>  cluster-07d8	us-east1	us-east1-c	2	Off	bucket_tp3	Nov 4, 2019, 11:29:48 PM	Running	

Go to the **Web Interface** tab and click on JupyterLab

Cluster details

SUBMIT JOB

REFRESH

cluster-07d8

For PD-Standard without local SSDs, we strongly recommend provisioning 1Ti information on disk I/O performance.

Monitoring

Jobs

VM Instances

Configuration

Web Interfaces

SSH tunnel

Create an SSH tunnel to connect to a web interface

Component gateway

YARN ResourceManager

HDFS NameNode

MapReduce Job History

YARN Application Timeline

Spark History Server

Tez

Jupyter

JupyterLab

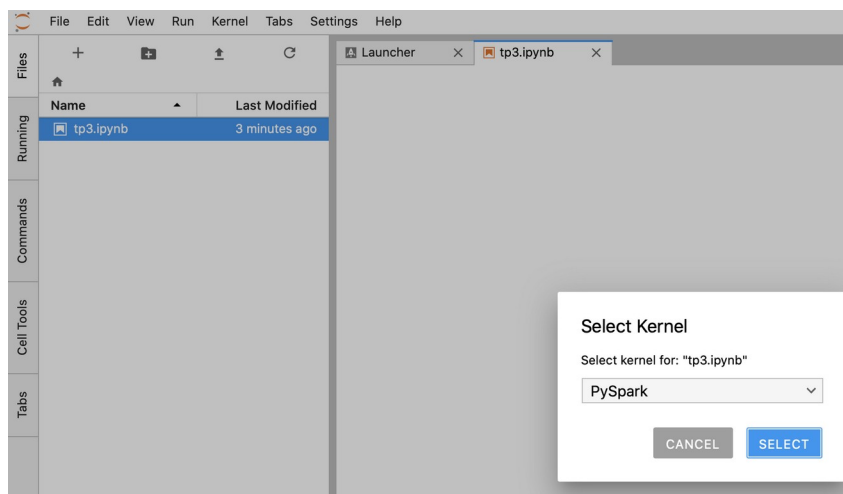
Equivalent REST

Now, go again to **Storage -> Browser** and open your bucket. We will see a notebooks folder.

<input type="checkbox"/>	Name	Size	Type	Storage class	Last modified
<input type="checkbox"/>	google-cloud-dataproc-metainfo/	—	Folder	—	—
<input type="checkbox"/>	notebooks/	—	Folder	—	—
<input type="checkbox"/>	toy.csv	48 B	text/csv	Standard	11/4/19, 10:39:08 PM UTC-5

Go to the **notebooks/jupyter** folder and upload your .ipynb file.

The page that was open when you clicked in JupyterLab now should show your Jupyter file. Open it and select the PySpark kernel.



Just run your notebook as usual.

Dataproc	Clusters + CREATE CLUSTER REFRESH DELETE REGIONS ▼						
	<input type="text" value="Search clusters, press Enter"/>						
Clusters							
Jobs							
Workflows							
	<input checked="" type="checkbox"/> Name ^	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created
	<input checked="" type="checkbox"/> cluster-07d8	us-east1	us-east1-c	2	Off	bucket_tp3	Nov 4, 2019, 11:29:48 PM

Once you have finished using the cluster, go to **Dataproc -> clusters**, select the cluster you desire to exclude and press **Delete**.