

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.

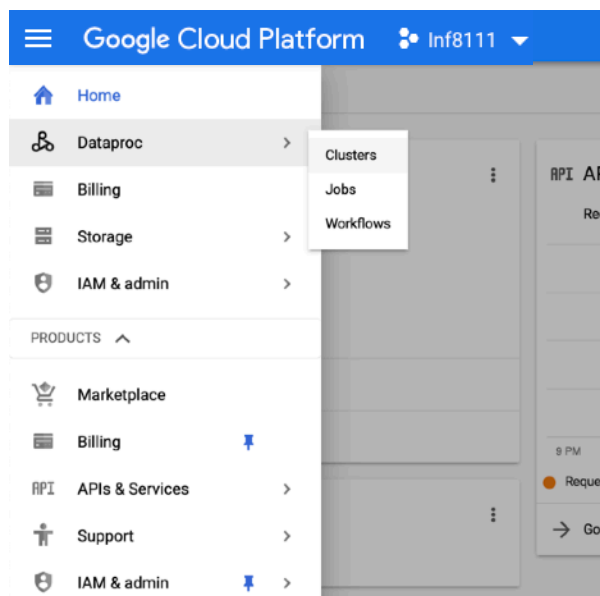
[Student Coupon Retrieval Link](#)

Once you have completed this step, you should have a project named **INF8111 - Fouille des données (Data mining)**. This project is linked to a billing account with your credits of 50\$.

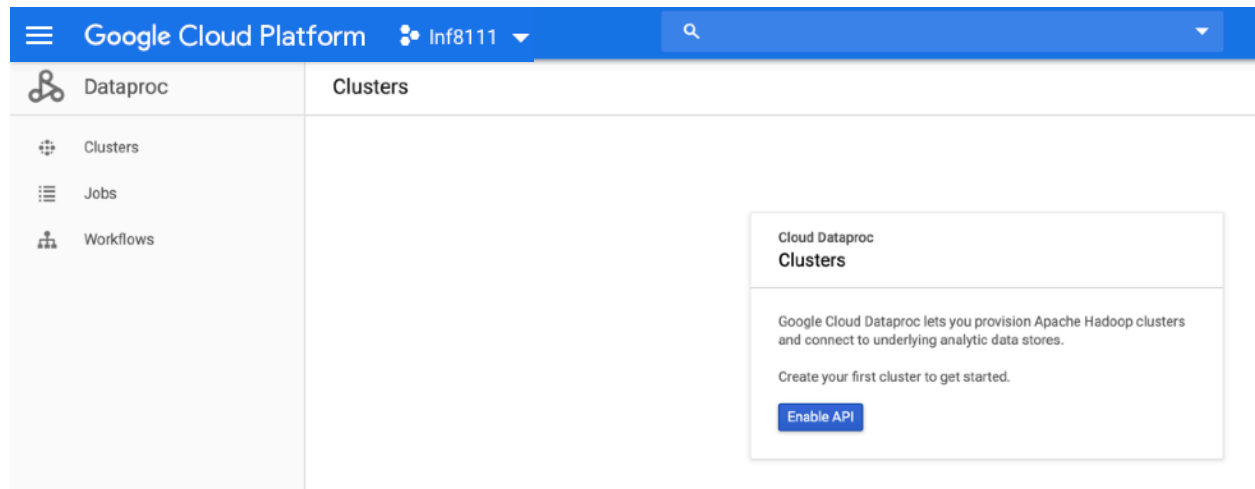
2. Enabling the required APIs.

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



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

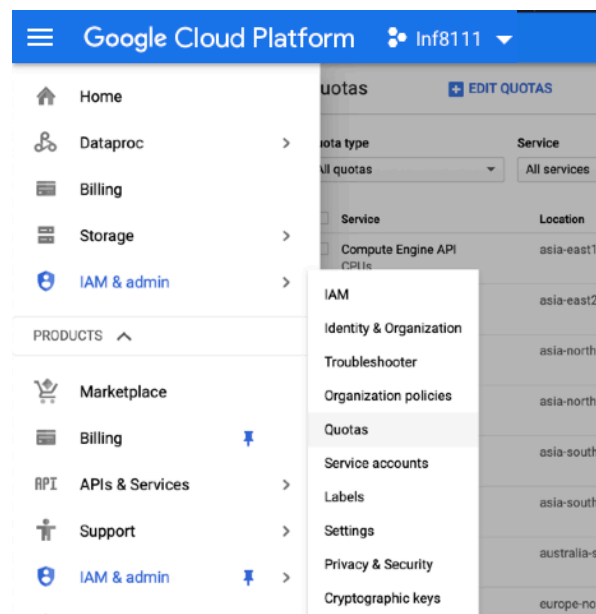


* this may be triggered automatically the first time you access this page.

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, select “CPUs” and “CPUs (all regions)” under the **Metric** select box and look for “Compute Engine API” for the “us-east1” location and CPUs (all regions) for “global”. Select it and click on Edit Quotas.

Quotas + EDIT QUOTAS	
<input type="checkbox"/>	Compute Engine API CPUs europe-west6
<input type="checkbox"/>	Compute Engine API CPUs northamerica-northeast1
<input type="checkbox"/>	Compute Engine API CPUs southamerica-east1
<input type="checkbox"/>	Compute Engine API CPUs us-central1
<input checked="" type="checkbox"/>	Compute Engine API CPUs us-east1
<input type="checkbox"/>	Compute Engine API CPUs us-east4
<input type="checkbox"/>	Compute Engine API CPUs us-west1
<input type="checkbox"/>	Compute Engine API CPUs us-west2
<input type="checkbox"/>	Compute Engine API CPUs us-west3
<input type="checkbox"/>	Compute Engine API CPUs us-west4
<input checked="" type="checkbox"/>	Compute Engine API CPUs (all regions) Global

Once asked for the new quota limit, inform 300 and in the description box write something similar to the one showing the image below.

Compute Engine API

Quota: CPUs - us-east1

New quota limit

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

300

Quota: CPUs (all regions)

New quota limit

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

300

Request description

Required

I am working in an academic project that demands a large cluster to run the application.

Done

Cancel

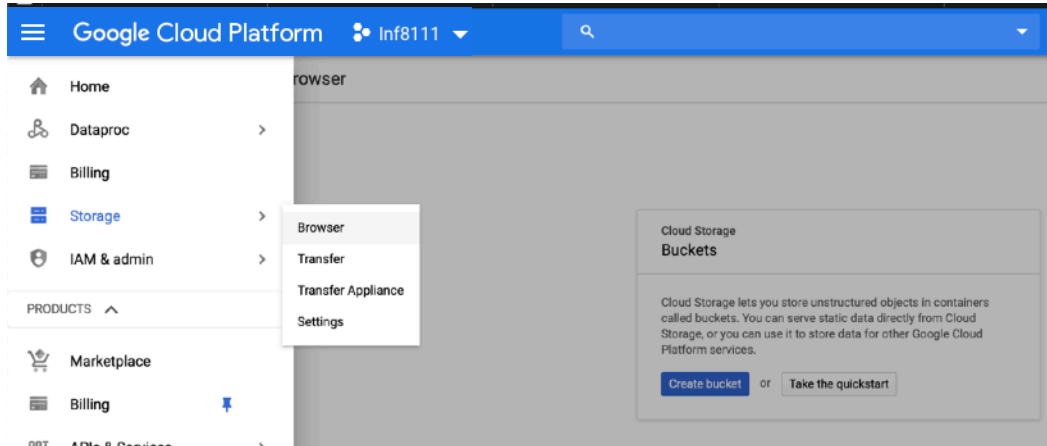
Submit request

Back

You will receive an email confirming your request. GCP usually takes between 30 minutes and a couple hours to process your request.

5. Creating a storage bucket

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



Give your bucket and name and under the “Choose where to store your data”, select Region and search for us-east1 (same as the region that you ask for the quota increment). Also set the access to objects as **uniform** and press “Create”.

A screenshot of the 'Create a bucket' wizard in the Google Cloud Platform console. The wizard has four steps: 1. 'Name your bucket' with a text input field containing 'bucket.tp' and a 'CONTINUE' button. 2. 'Choose where to store your data' with options for 'Location type' (Region, Dual-region, Multi-region) and a 'Location' dropdown menu set to 'us-east1 (South Carolina)'. 3. 'Choose how to control access to objects' with options for 'Access control' (Fine-grained, Uniform) and a 'CONTINUE' button. 4. 'Advanced settings (optional)' with 'CREATE' and 'CANCEL' buttons at the bottom.

You will be redirected to your bucket page from where you starting uploading some files. As an example, upload the toy.csv file to your bucket.

[←](#) Bucket details [EDIT BUCKET](#) [REFRESH BUCKET](#)

bucket_tp

[Objects](#) [Overview](#) [Permissions](#) [Bucket Lock](#)

[Upload files](#) [Upload folder](#) [Create folder](#) [Manage holds](#) [Delete](#)

[Buckets](#) / [bucket_tp](#)

<input type="checkbox"/>	Name	Size	Type	Storage class	Last modified	Public access	Encryption
<input type="checkbox"/>	toy.csv	52 B	text/csv	Standard	5/25/20, 2:28:53 PM UTC-4	Not public	Google-managed key

If you go the the Overview tab, the **Link for gsutil** gives you the address for your bucket. For example, to access my toy.csv file contained in my bucket, it path would be `gs://bucket_tp/toy.csv`.

bucket_tp

[Objects](#) [Overview](#) [Permissions](#) [Bucket Lock](#)

Created	May 25, 2020 at 2:28:30 PM UTC-4
Updated	May 25, 2020 at 2:31:52 PM UTC-4
Location type	Region
Location	us-east1 (South Carolina)
Default storage class	Standard
Access control	Uniform
Requester pays	Off
Encryption type	Google-managed key
Link URL	<input type="text" value="https://console.cloud.google.com/storage/browser/bucket_"/>
Link for gsutil	<input type="text" value="gs://bucket_tp"/>

6. 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 have to change the name for the cluster, but it is necessary to specify the **Region**. Select us-east1 (or the region for which you requested a quota increase).

Now we have to set the number of CPUs that we will use in our cluster. The Cluster mode is the Standard(1 master, N workers)

In our application the most valuable resource is memory. Thus, both for the master node as for the workers nodes will use machines from the type highmem.

- For the master node, select the 32vCPUs of type **n1-highmem-32**.

- For the worker nodes, select 7 nodes of 32vCPUs of type **n2-highmem-32**.

This will give your cluster an total of 256 (32 +224) vCPUs and 1.14 TB of memory on the workers nodes.

Note: 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](#). For example, for this given configuration, we hourly price would be:

Create a cluster

Name: cluster-047d

Region: us-east1 Zone: us-east1-c

Cluster mode: Standard (1 master, N workers)

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

Machine configuration

Machine family: General-purpose
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-8 (8 vCPU, 52 GB memory)

vCPU: 8 Memory: 52 GB

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

Machine configuration

Machine family: General-purpose
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-32 (32 vCPU, 208 GB memory)

vCPU: 32 Memory: 208 GB

Primary disk size (minimum 15 GB): 500 GB Primary disk type: Standard persistent disk

Nodes (minimum 2): 7 Local SSDs (0-8): 0 x 375 GB

YARN cores: 224 YARN memory: 1.14 TB

$1 \times 0.4 + 7 \times 0.4 = \3.2 per hour.

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.

VERY IMPORTANT:

There is still a crucial step in the cluster configuration to be done.

First, select the **Component gateway** option and click to expand the advance options:



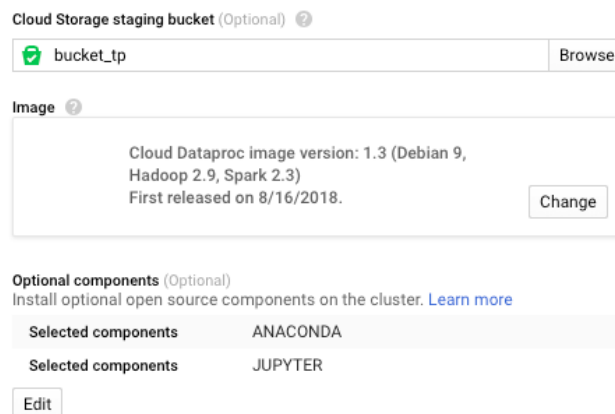
Component gateway

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

⌵ Advanced options

Create Cancel

Look for **Cloud Storage staging bucket** and browser your bucket;
In **Optional components**, select ANACONDA and JUPYTER.



Cloud Storage staging bucket (Optional) ?


 bucket_tp Browse

Image ?

Cloud Dataproc image version: 1.3 (Debian 9, Hadoop 2.9, Spark 2.3)
First released on 8/16/2018. Change

Optional components (Optional)
Install optional open source components on the cluster. [Learn more](#)

Selected components	ANACONDA
Selected components	JUPYTER

Edit

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.

7. Using your cluster

Once your cluster is created, click to open it.

Clusters

+

CREATE CLUSTER

↻

REFRESH

🗑

DELETE

REGIONS

▼

☰

Search clusters, press Enter


?

<div><input type="checkbox"/></div> <div>Name ^</div>	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created	Status
<div><input type="checkbox"/></div> <div><div><div></div><div>✔</div></div>cluster-07d8</div>	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 1TB 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.


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




Databroc

Clusters

 CREATE CLUSTER
  REFRESH

 DELETE

REGIONS ▾


Clusters

Jobs

Workflows

☰

?

✓ Name ^	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created
<div>✓</div> <div>  <div>cluster-07d8</div> </div>	us-east1	us-east1-c	2	Off	bucket_tp3	Nov 4, 2019, 11:29:48 PM