This manual will help you set up GCP and will cover the following:

1. How to redeem your credits
2. Creating a project and check billing
3. Set up the environment to working with pyspark on GCP
4. Creating a bucket on google storage
5. Working with Dataproc and creating a cluster
6. Figure out the cost breakdown of your GCP usage.

Please note that we might not cover everything, but GCP documentation online is clear and easy to follow.

**\*\* The setup will take about an hour to complete. \*\***

**Redeem credits**

Before we dig in, we first need to create our GCP account. Each student has $50 to work on GCP and you should redeem your credits before getting started.

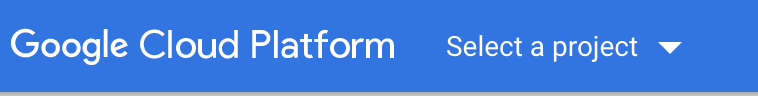
To redeem your credits you should enter one of the following links ([here](https://gcp.secure.force.com/GCPEDU?cid=bx1a0s4Y9hGpQ44CXwktW7aXJdEWozhg8kibHpyf7QJAsECLH6dTXe%2Fk%2FhF1Nnyq/)). Please try the first link and if it doesn’t work enter the second one. **Make sure you use your post account.**

**You don’t need to insert your credit card to use the credits.** However, you should make sure you have enough credits because we can’t get more.If you finish your credits before the end of the course you will be able to enter a free trial for a month and get $300 free credits, but at this point you will need to enter your credit card details.

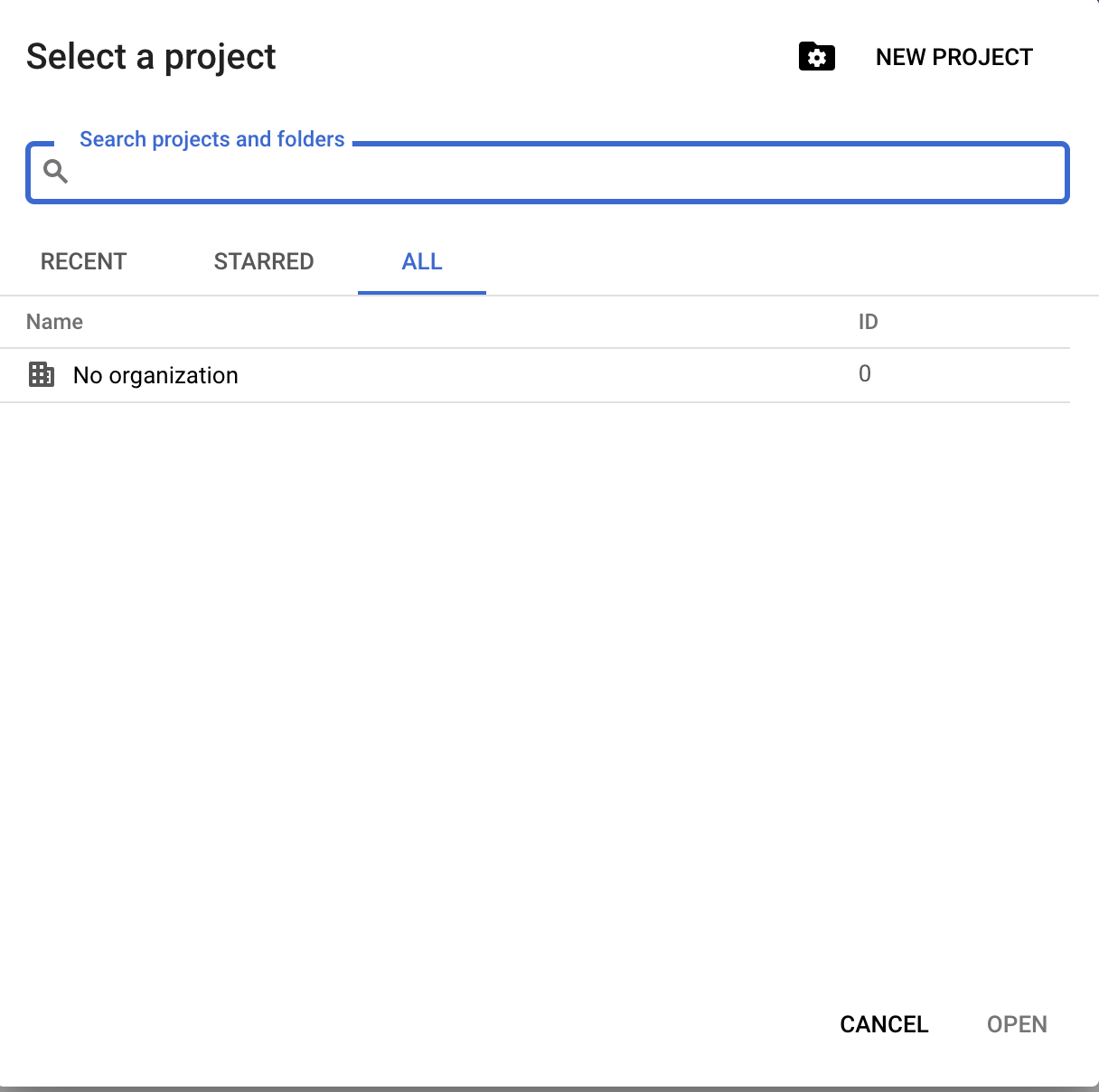
Redeeming your credits automatically creates your GCP account.

**Create project**

At the top bar there is a “select a project” button.



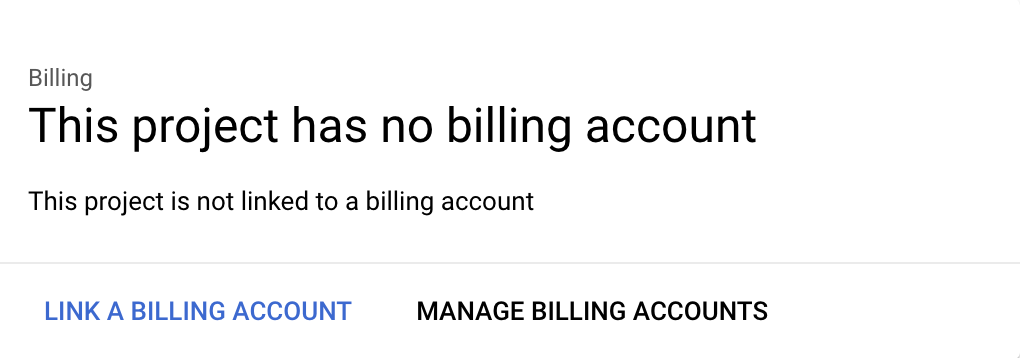
Clicking it will open a window where you can create a new project.



Name the project as you want and click create.

**Check Billing**

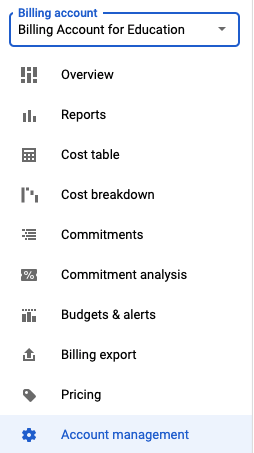
Using the navigation menu at the left of the top bar, access the Billing. You will arrive to this page:



click the “link a billing account” and connect your project to the billing account containing the credits.

If we get to the billing overview page without getting the “link a billing account” msg, it is possible that your project is already connected to the billing account.

In order to check this, go to “account management” and make sure your project appears on the list. The billing account should be the “Billing Account for Education”.



For more information, you can access GCP documentation [on billing](https://cloud.google.com/billing/docs/how-to/modify-project#confirm_billing_is_enabled_on_a_project).

**Setting up the environment**

1. Make sure to enable both the Compute Engine API and the Cloud Storage API.
2. Install cloud SDK following [this](https://cloud.google.com/sdk/docs/install) documentation.
3. Create a service account (more on service accounts can be found [here](https://cloud.google.com/iam/docs/service-accounts)):
   1. Navigation menu → IAM & Admin → Service Accounts → create service account.
   2. In the “Service account name” field, enter a name. The Cloud Console fills in the “Service account ID” field based on this name. In the Service account description field, enter a description. Then click create and continue.
   3. Click the **Select a role** field. Under **Quick access**, click **Basic**, then click **Owner**. Then click continue and then Done to finish creating the service account.
   4. Create a service account key - In the Cloud Console, click the service accounts as previously mentioned, click the project you created, then click the three points manage keys → Click **Keys** → Click **Add key**, then click **Create new key** → Click **Create**. A JSON key file is downloaded to your computer → Click **Close**.
   5. Set the environment variable GOOGLE\_APPLICATION\_CREDENTIALS to the path of the JSON file that contains your service account key. This variable only applies to your current shell session, so if you open a new session, set the variable again.(Replace **KEY\_PATH** with the path of the JSON file that contains your service account key). In the GCP assignment you will need to update this line in the code to have your own local key\_path instead of the path that currently exists.
      1. *export GOOGLE\_APPLICATION\_CREDENTIALS="****KEY\_PATH****"*

**Google Storage**

In the course we work with wiki dumps, that are XML files. In order to save you (and us) time, we created preprocessed files that are used for the project and the 3rd assignment.

All files are in a public Google Storage Bucket, and can be accessed by you at all times. Instructions on how to access the bucket and the files are presented in assignment 3.

Other than the public bucket we created for you that holds all the data, you will need to create your own storage bucket.

1. Go to Cloud storage→ browser → create bucket.
2. Give your bucket a unique name, preferably containing your id (there can’t be two buckets with the same name on google storage).
3. In the location type choose Region then select us-central1.
4. click create

\*After you created the bucket, we should upload the initialization file (.sh file) needed for assignment 3. The file can be found on moodle under assignment 3. Click on the bucket you created then click **upload files** and upload the relevant file through the browser that was opened.

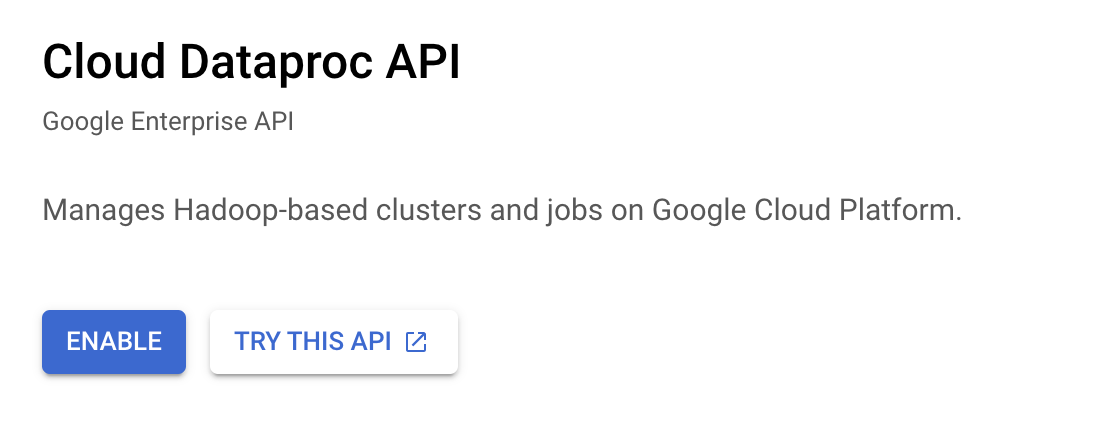
\* You also need to transfer the data to your bucket in order to use him.

1. Click on the bucket you created and then click transfer data → transfer data in.
2. Go to choose a source and paste the name **wikidata20210801\_preprocessed**.
3. Go to choose a destination and select the bucket you just created.
4. Click on create.

After the job is finished your bucket will contain all the multistream files.

**Create cluster**

After we connect the billing account, we can create a cluster that we will use in the project and in the 3rd assignment. To open a cluster we will first need to enable the DataProc API. You can either find the API in the navigation menu, or you can search for it in the search bar at the top of the page. Then, click on the enable button.



Now, we can open a new cluster.

Go to Dataproc → clusters → create cluster -> Cluster on Compute Engine:

1. In the **set up cluster section** you can change the cluster name. Make sure the cluster type is **standard (1 Master, N workers).**

In the region select “us-central1” and in the Zone select “us-central1-a”.

Then, make sure you check the box “enable component gateaway” under components, and check the jupyter notebook box under optional componentes.

1. In the **Configure nodes**, under worker nodes, you can change the number of worker nodes. No need to change anything else, just use the default values.
2. Under **Customize cluster→ Initialization actions,** choose add initialization action. you should upload an initialization file. The initialization file should already be in the bucket we created in google storage so choose the relevant bucket and then the initialization file.
3. Under **Manage Security → Project Access** check the box that says “Allow API access to all Google Cloud services in the same project.”
4. Click create.

**Important** - The main waste of your credits is the **cluster**, so it is highly recommended to delete it at the end of the work, so that you don't run out of money for the project.

**Run spark on GCP**

Now that we have a cluster on Dataproc, we can open a jupyter notebook and start writing code. Go to Dataproc → clusters → click on the name of the cluster we just created.

This will open a window with all the information on the cluster. Click on **web interfaces** and then **jupyter**. Now you are in the jupyter environment. We can create jupyter notebooks in the local disk folder, navigate to the home directory (‘/ Local Disk / home / dataproc’ folder). If you want to run pyspark, click **new** → pySpark.

If you upload the notebook instead of creating it you should change the kernel to pySpark.

Make sure you upload both the `inverted\_index\_gcp.py` file and the `assignment3\_gcp.ipynb` notebook to the directory (‘/ Local Disk / home / dataproc’ folder).

Perfect, now you can start writing your code.

## **Figure out the cost breakdown of your usage**

1. Read this [documentation on the topic](https://cloud.google.com/blog/topics/pricing/shining-a-light-on-your-costs-new-billing-features-from-google-cloud) and make sure you watch the [video there](https://youtu.be/XR_d8u5AGyM).
2. In your GCP console, go to Billing → reports. Select the relevant time range, group by SKU and you'll see a detailed breakdown of costs.

Good luck!