3803ICT
Big Data Analysis

Lab 11 – Data Analytics in Cloud

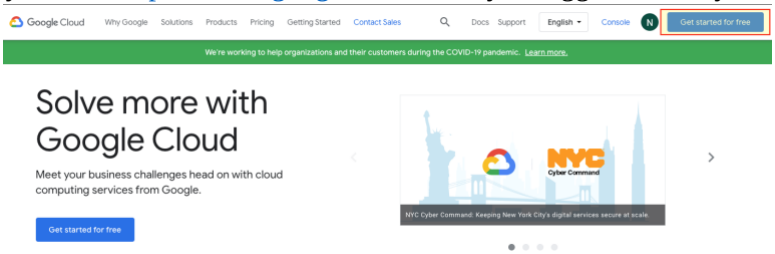
Trimester 1 - 2022

Table of Contents

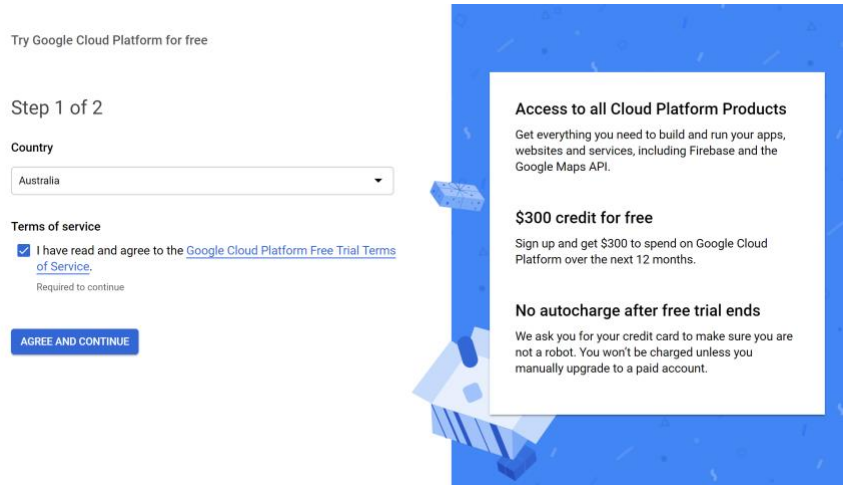
ATTENTION—DO NOT FORGET TO STOP THE VM INSTANCE!!

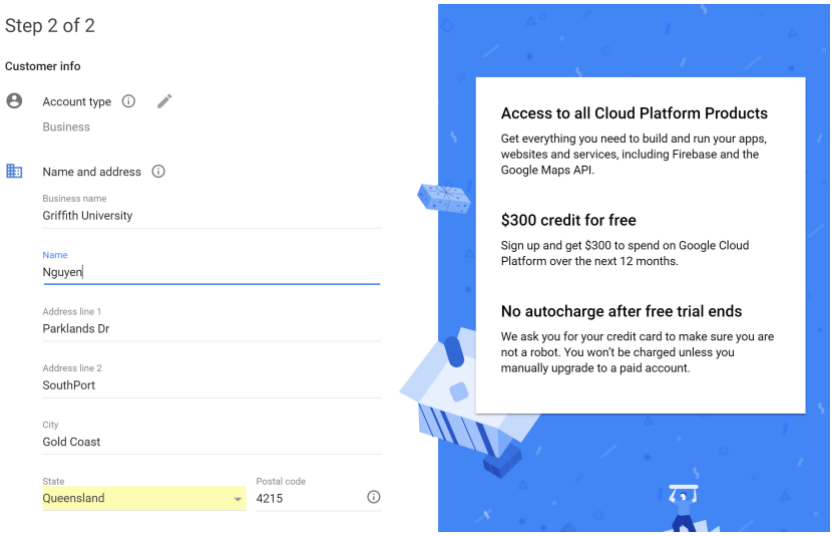
I. Account Registration

Click Try Free in https://cloud.google.com/ after you logged in with your google account.

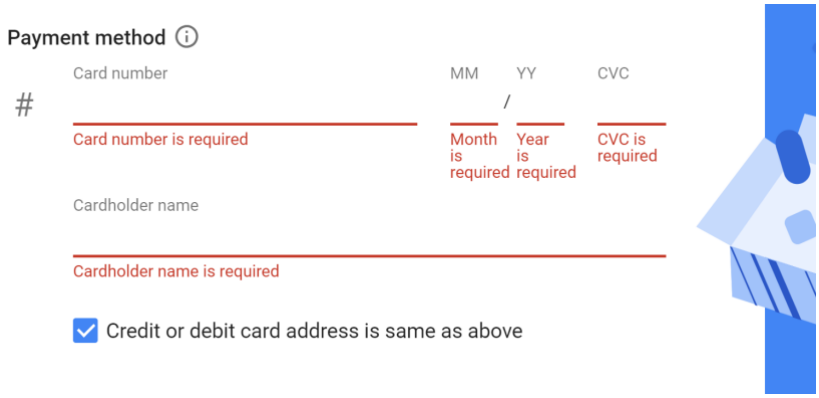


Provide your information as needed.





Input your payment information. You will have $300 for your trial. **Then you can CANCEL this after the course.**

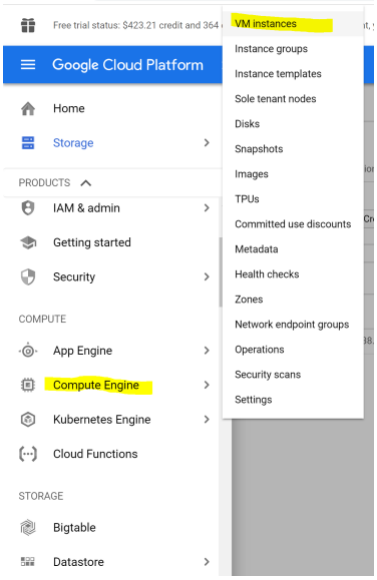


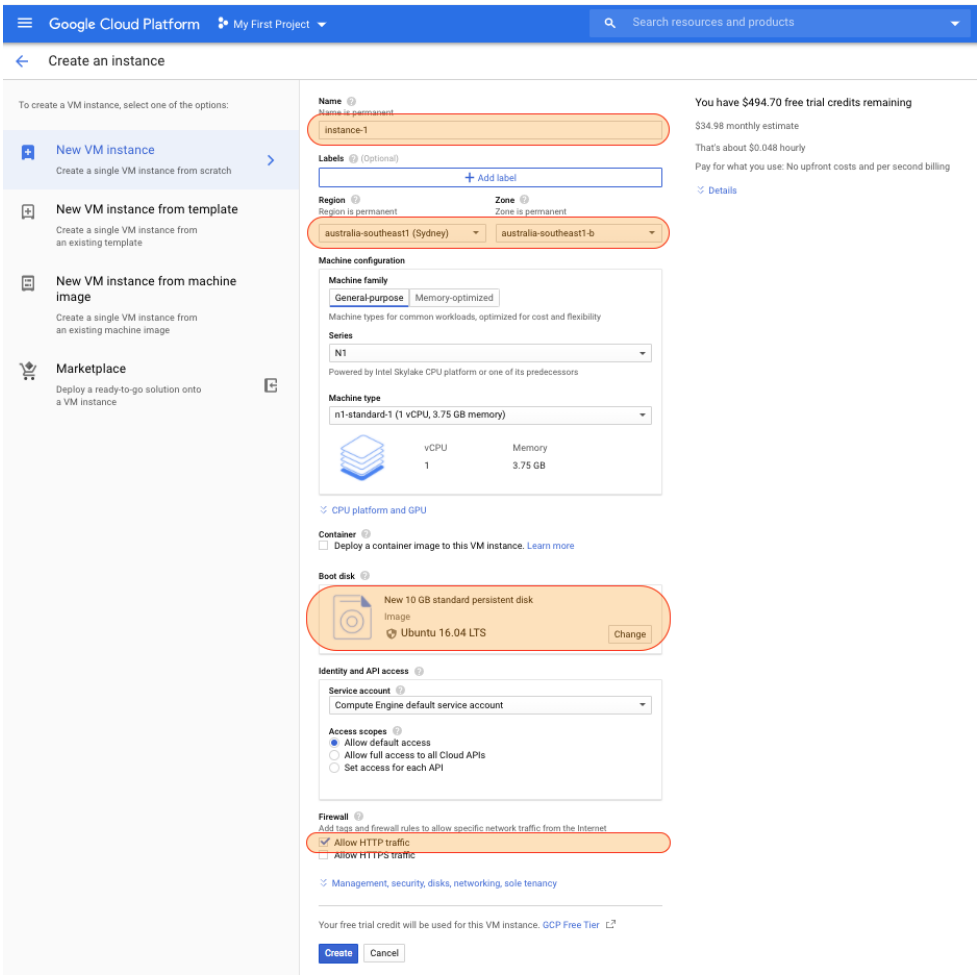
II. Deployment

In this section, you need to deploy a website. The csv file contains emotion analysis of tweet posts over time.

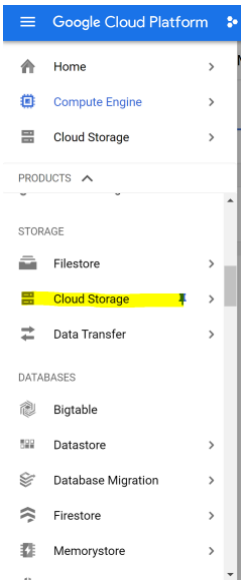
A. Tutorial

1. Create a Compute Engine

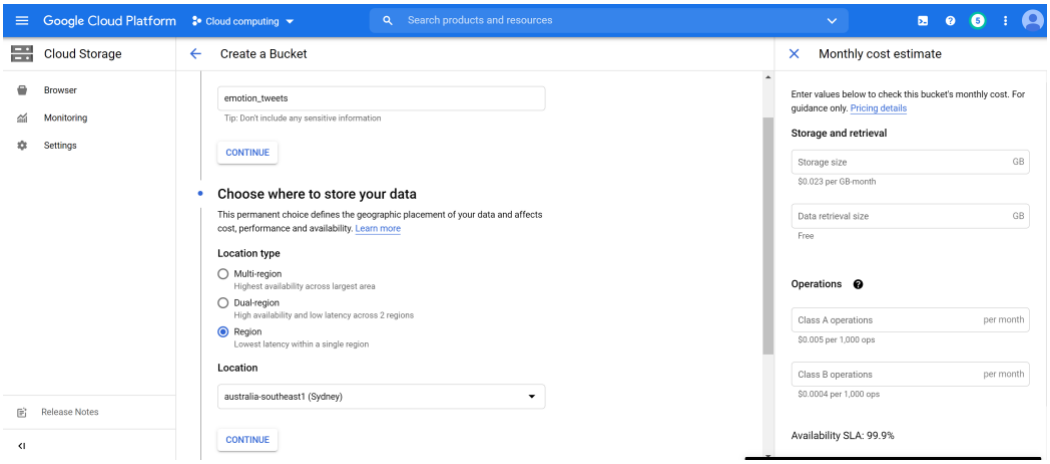




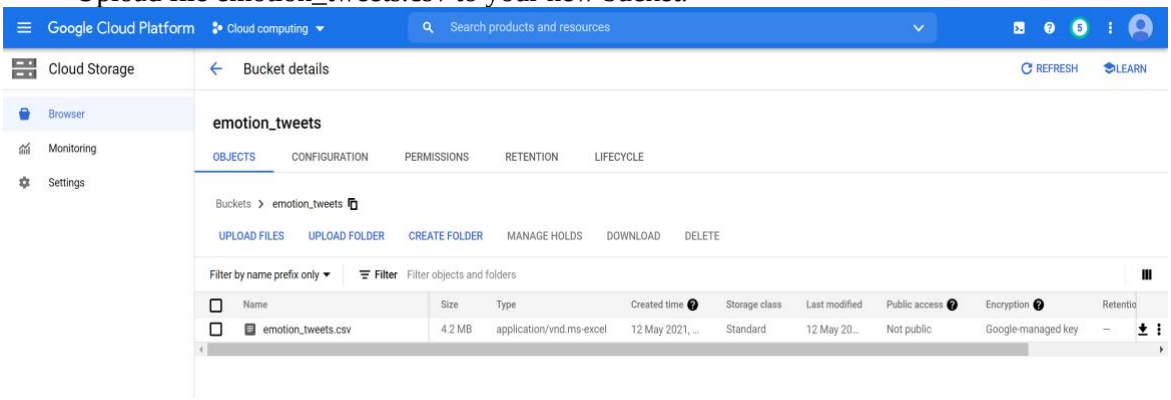
2. Upload file to Google Storage
+ Click Cloud Storage in left panel.



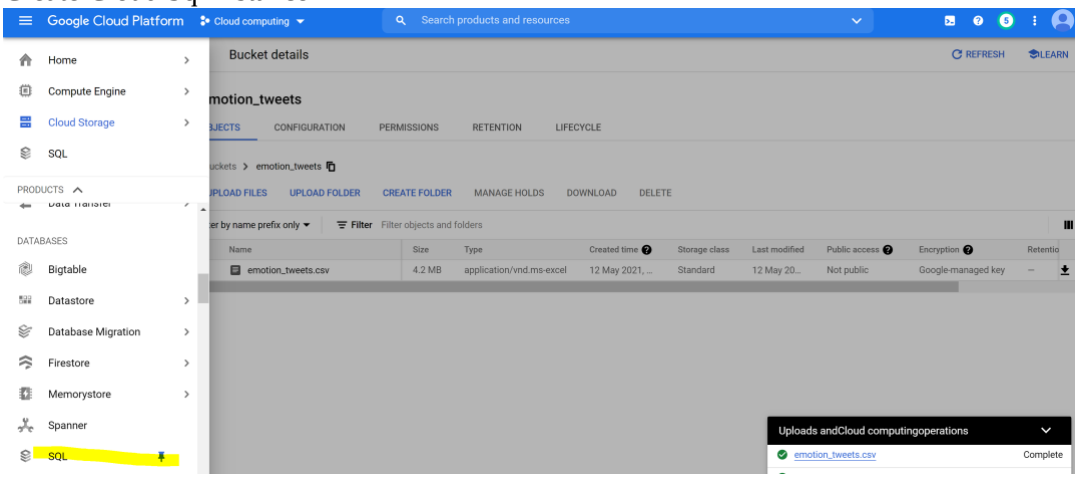
+ Create bucket.
. Put name.
. Choose ‘regional’ as storage class.
. Select ‘australia-southeast1’ as location.
. Default for other information.

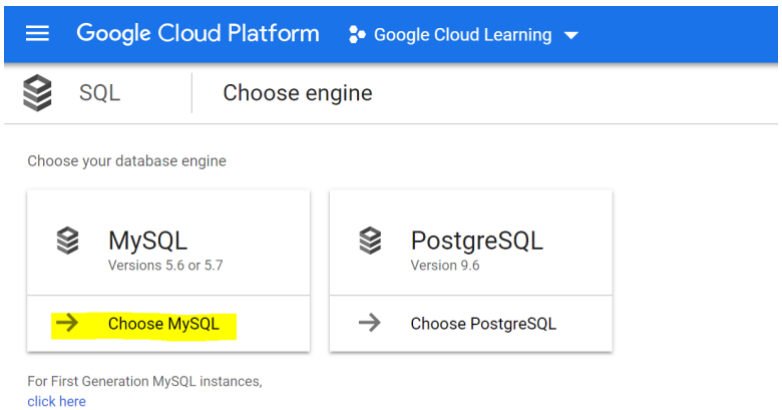


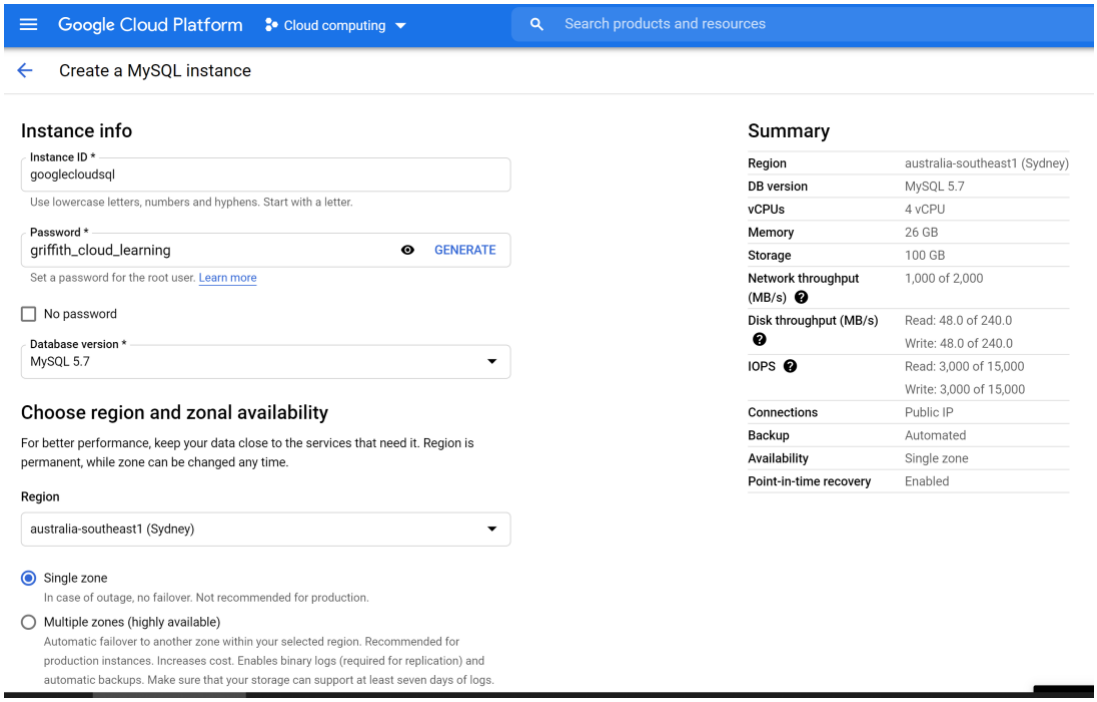
+ Upload file emotion_tweets.csv to your new bucket.



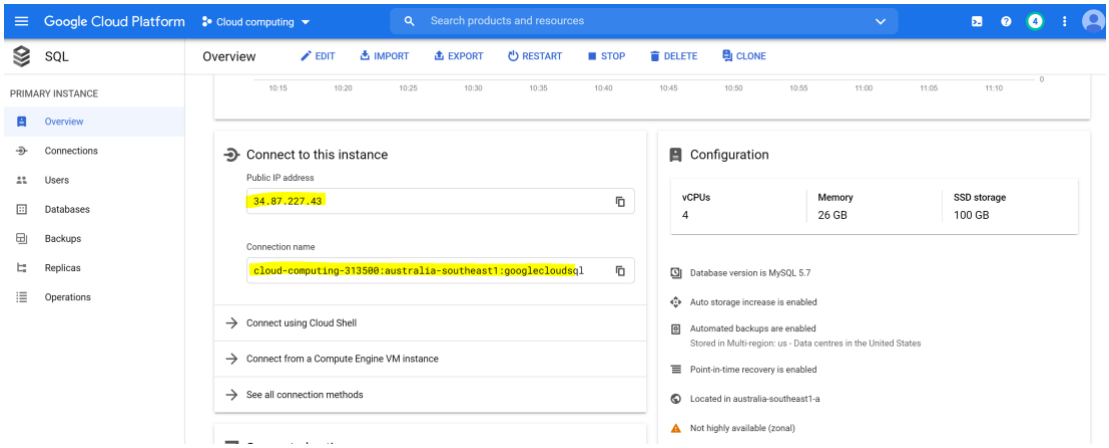
3. Create Cloud Sql instance



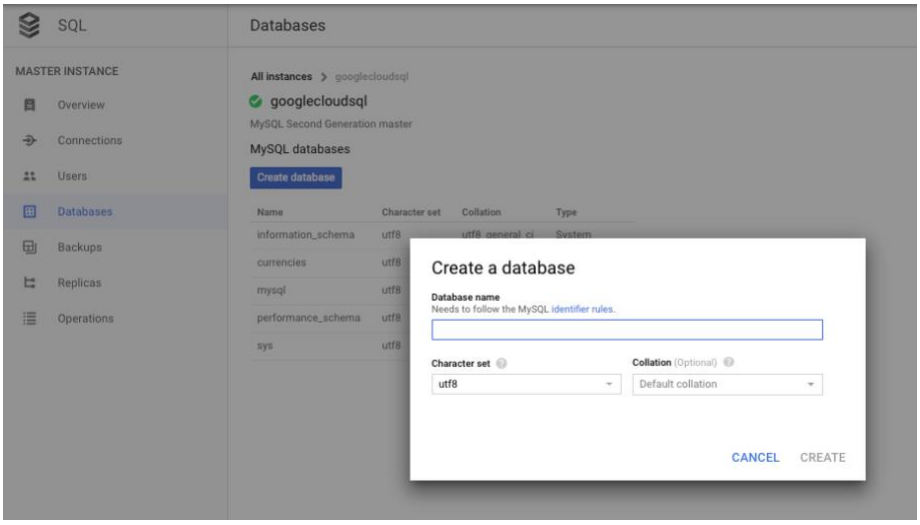




+You can choose your own password, but you should remember it for later steps.
+ Write down your instance IP (ex: **34.87.227.43**) for later usages.



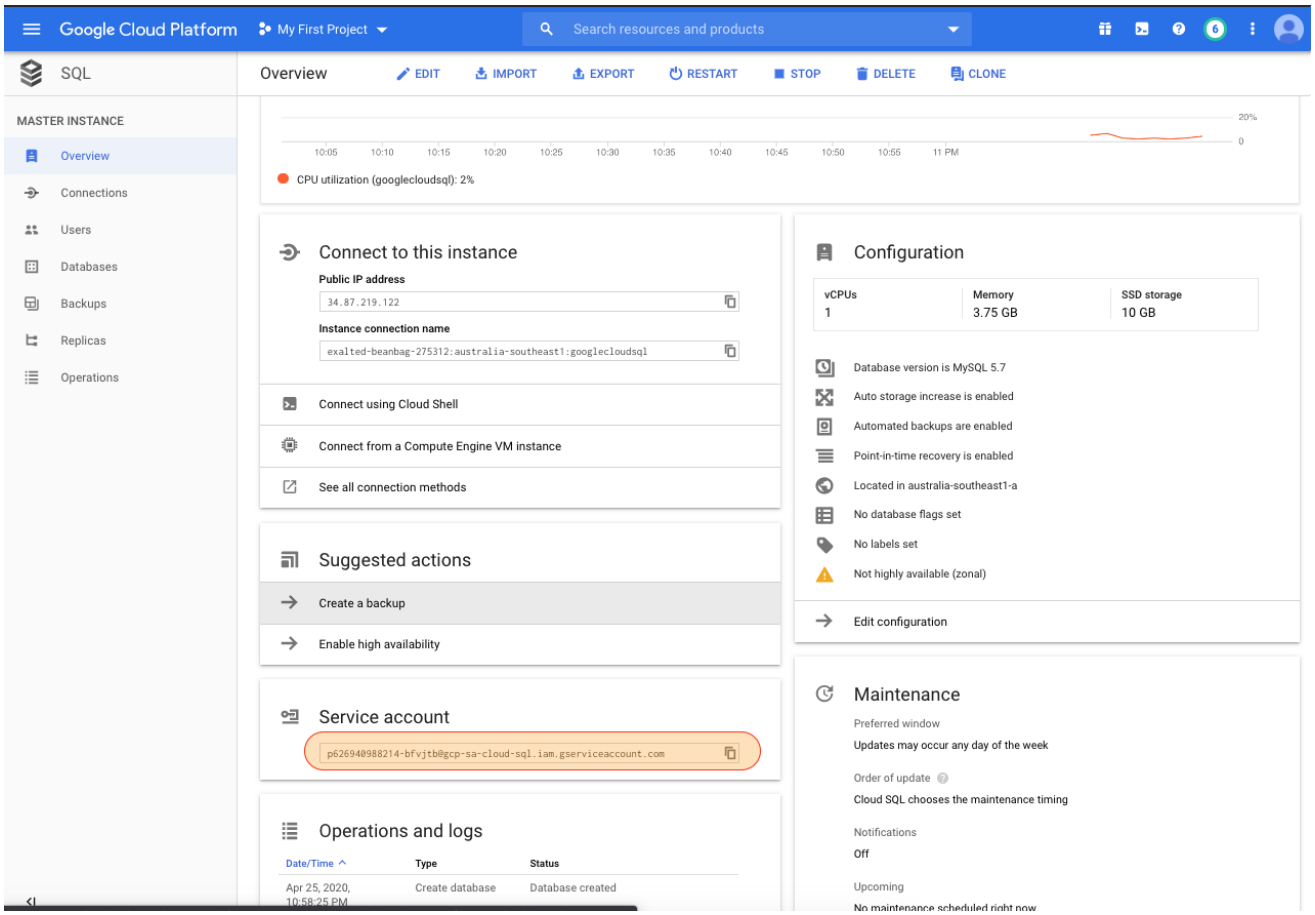
+ Create new database ‘emotions’



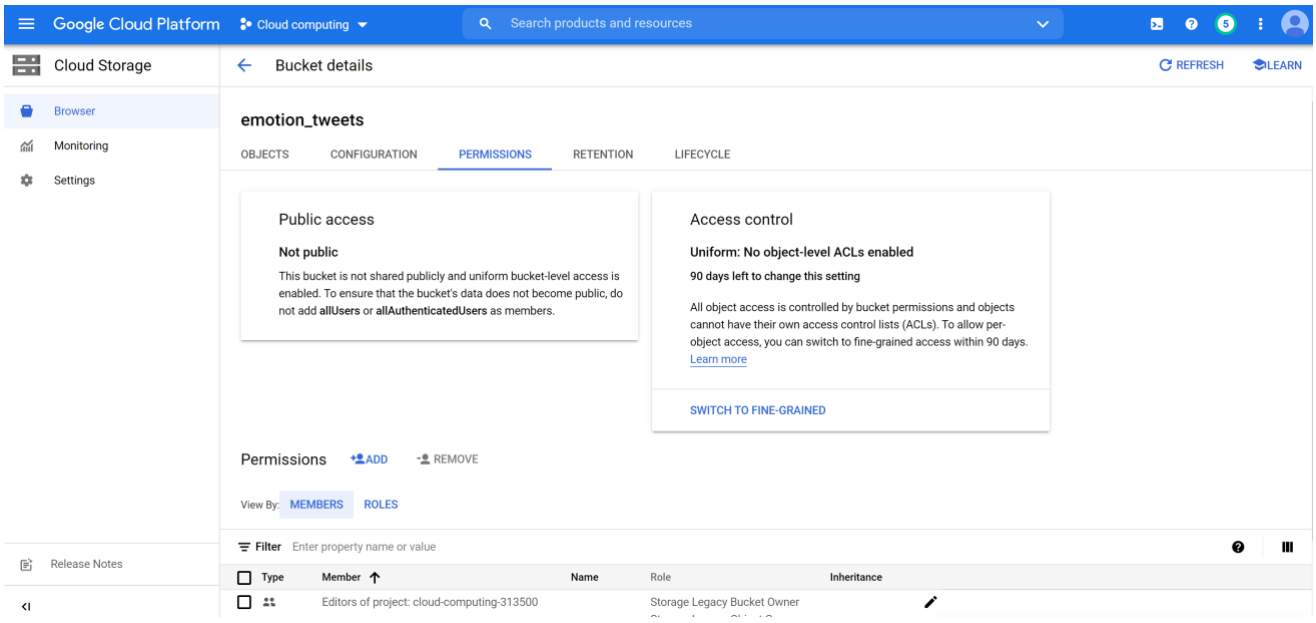
4. Setup permission

a. Allow sql service to access bucket to run import job later

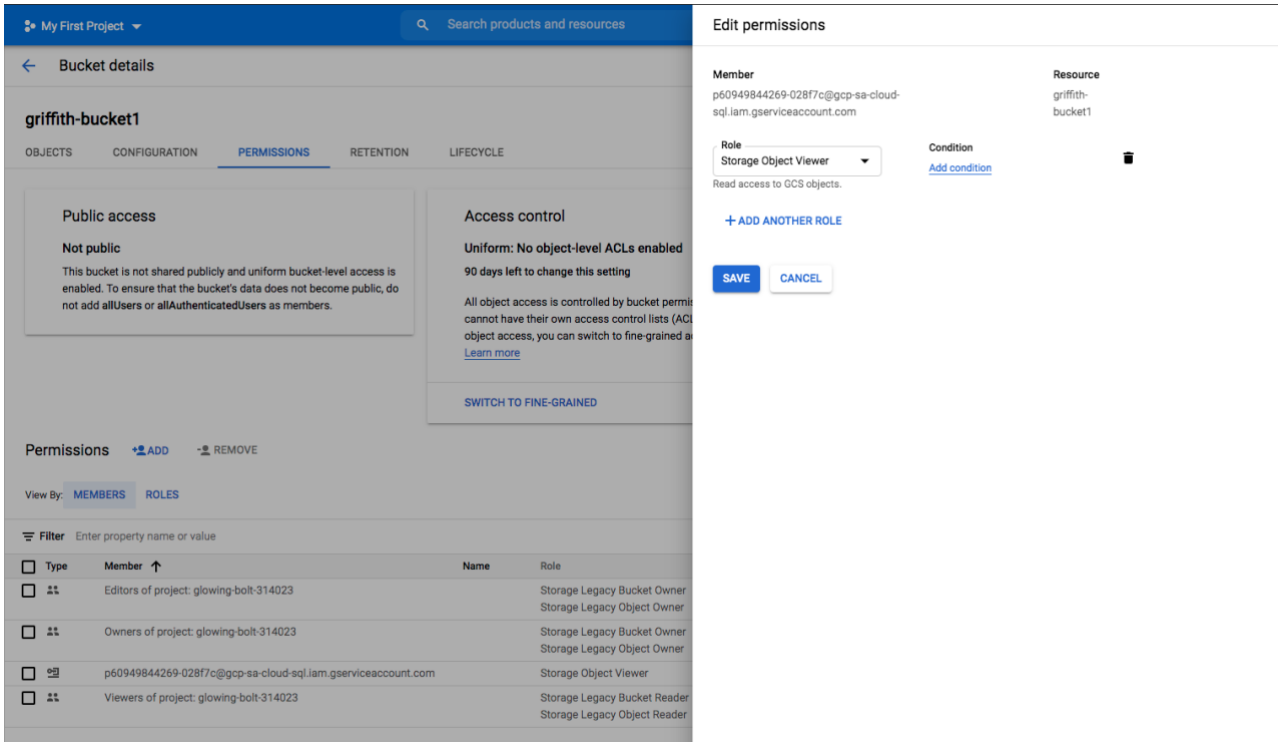
Get Service account from Cloud Sql (ex: p626940988214-bfvjtb@gcp-sa-cloud-sql.iam.gserviceaccount.com)



Open Cloud Storage => Your bucket (ex: emotion_tweets) => Permissions

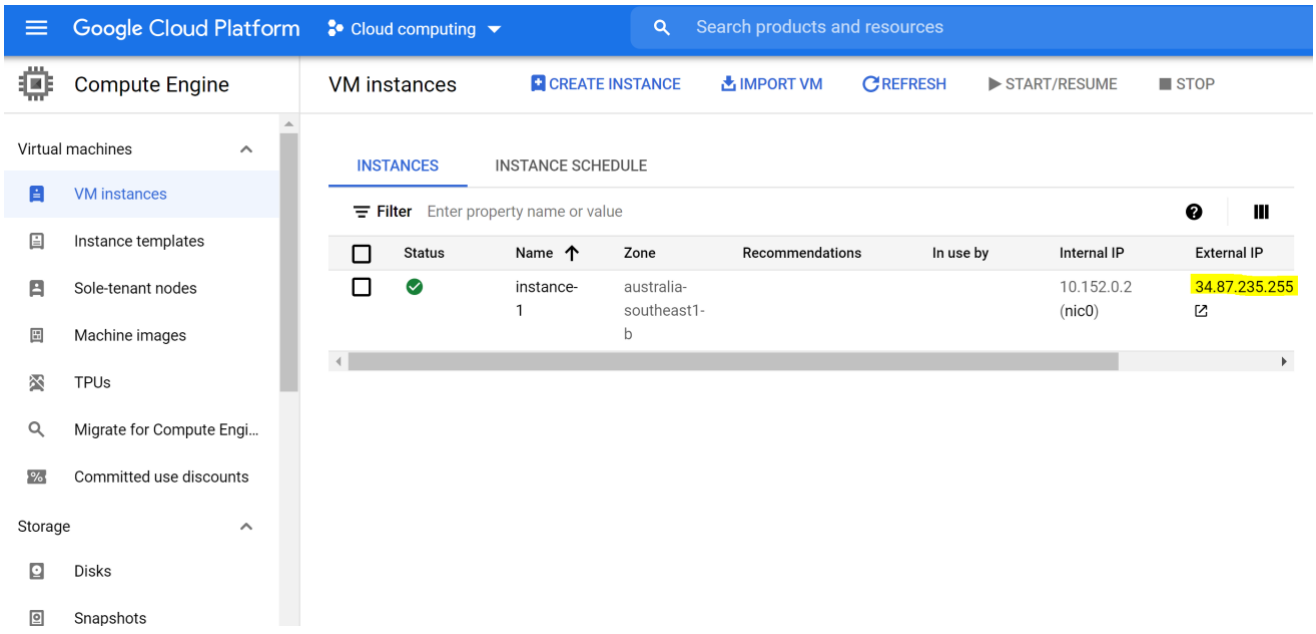


Add sql service account above (p626940988214-bfvjtb@gcp-sa-cloud-sql.iam.gserviceaccount.com) account with role Storage Object Viewer.

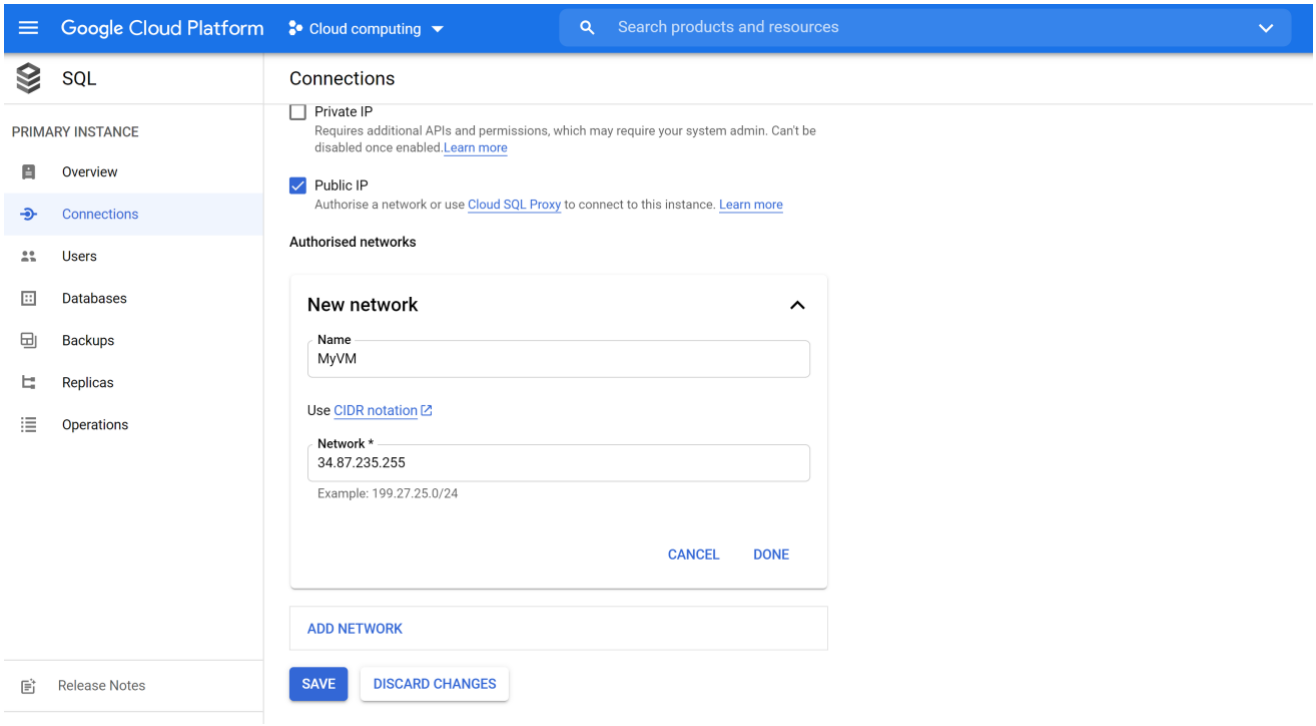


b. Allow VM instance to access sql cloud.

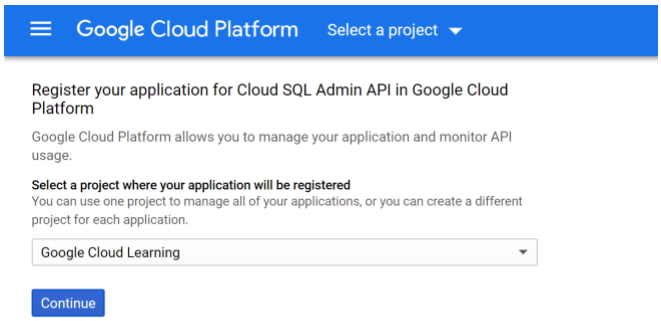
From left panel, click compute engine => copy your instance's external IP. (ex: **34.87.235.255**)



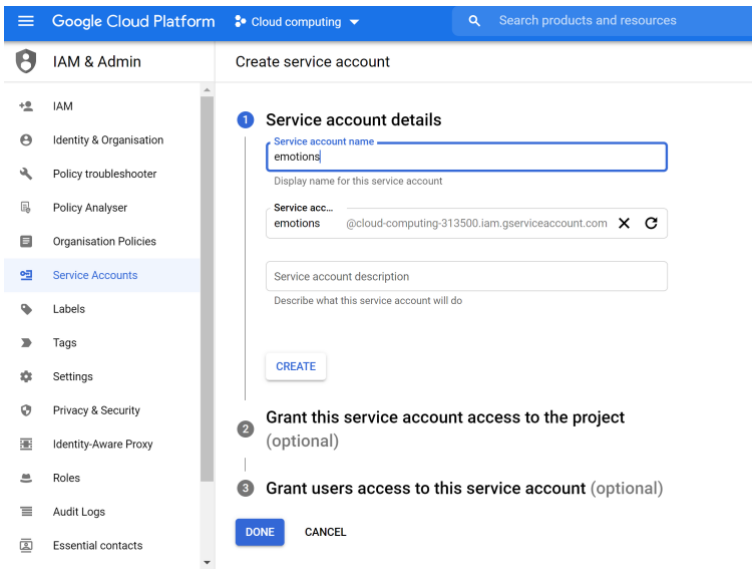
From left panel, click SQL => select your instance => Connections => Click "Add Network"
Copy above IP (ex: **34.87.235.255**) to Network.



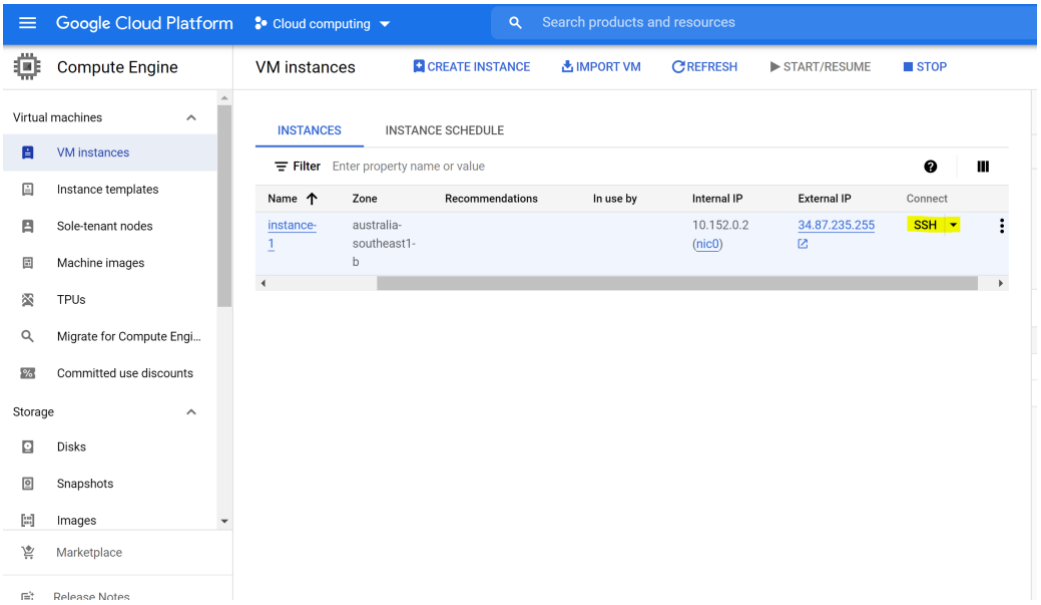
Enable Cloud SQL API: https://console.cloud.google.com/flows/enableapi?apiid=sqladmin. Select your project and then enable:



Follow the intruction to create the service account:



5. Connect My SQL in VM instance and import data
From left panel => choose Compute Engine => connect SSH



Install mysql client using command: sudo apt-get install mysql-client

```
[redacted]@instance-1:~$ sudo apt-get install mysql-client
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  libaio1 mysql-client-5.7 mysql-client-core-5.7 mysql-common
The following NEW packages will be installed:
  libaio1 mysql-client mysql-client-5.7 mysql-client-core-5.7 mysql-common
0 upgraded, 5 newly installed, 0 to remove and 0 not upgraded.
Need to get 8,431 kB of archives.
After this operation, 65.7 MB of additional disk space will be used.
Do you want to continue? [Y/n] Y
Get:1 http://australia-southeast1.gce.archive.ubuntu.com/ubuntu xenial/main amd64 libaio1 amd64 0.3.110-2 [
Get:2 http://australia-southeast1.gce.archive.ubuntu.com/ubuntu xenial-updates/main amd64 mysql-client-core
 5.7.25-0ubuntu0.16.04.2 [6,673 kB]
Get:3 http://australia-southeast1.gce.archive.ubuntu.com/ubuntu xenial-updates/main amd64 mysql-common all
untu0.16.04.2 [15.5 kB]
Get:4 http://australia-southeast1.gce.archive.ubuntu.com/ubuntu xenial-updates/main amd64 mysql-client-5.7
25-0ubuntu0.16.04.2 [1,726 kB]
Get:5 http://australia-southeast1.gce.archive.ubuntu.com/ubuntu xenial-updates/main amd64 mysql-client all
untu0.16.04.2 [10.0 kB]
Fetched 8,431 kB in 1s (5,424 kB/s)
Selecting previously unselected package libaio1:amd64.
(Reading database ... 71063 files and directories currently installed.)
Preparing to unpack .../libaio1_0.3.110-2_amd64.deb ...
Unpacking libaio1:amd64 (0.3.110-2) ...
Selecting previously unselected package mysql-client-core-5.7.
Preparing to unpack .../mysql-client-core-5.7_5.7.25-0ubuntu0.16.04.2_amd64.deb ...
Unpacking mysql-client-core-5.7 (5.7.25-0ubuntu0.16.04.2) ...
Selecting previously unselected package mysql-common.
Preparing to unpack .../mysql-common_5.7.25-0ubuntu0.16.04.2_all.deb ...
Unpacking mysql-common (5.7.25-0ubuntu0.16.04.2) ...
Selecting previously unselected package mysql-client-5.7.
Preparing to unpack .../mysql-client-5.7_5.7.25-0ubuntu0.16.04.2_amd64.deb ...
Unpacking mysql-client-5.7 (5.7.25-0ubuntu0.16.04.2) ...
Selecting previously unselected package mysql-client.
Preparing to unpack .../mysql-client_5.7.25-0ubuntu0.16.04.2_all.deb ...
Unpacking mysql-client (5.7.25-0ubuntu0.16.04.2) ...
Processing triggers for libc-bin (2.23-0ubuntu10) ...
Processing triggers for man-db (2.7.5-1) ...
Setting up libaio1:amd64 (0.3.110-2) ...
Setting up mysql-client-core-5.7 (5.7.25-0ubuntu0.16.04.2) ...
Setting up mysql-common (5.7.25-0ubuntu0.16.04.2) ...
update-alternatives: using /etc/mysql/my.cnf.fallback to provide /etc/mysql/my.cnf (my.cnf) in auto mode
Setting up mysql-client-5.7 (5.7.25-0ubuntu0.16.04.2) ...
```

Connect to your mysql using comman: mysql --host=[your sql IP] --user=root --password=[your root password]

mysql --host=34.87.227.43 --user=root --password=griffith_cloud_learning

```
[redacted]nstance-1:~$ mysql --host=35.197.176.134 --user=root --password=griffith_cloud_learning
mysql: [Warning] Using a password on the command line interface can be insecure.
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 186
Server version: 5.7.14-google-log (Google)

Copyright (c) 2000, 2019, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql>
```

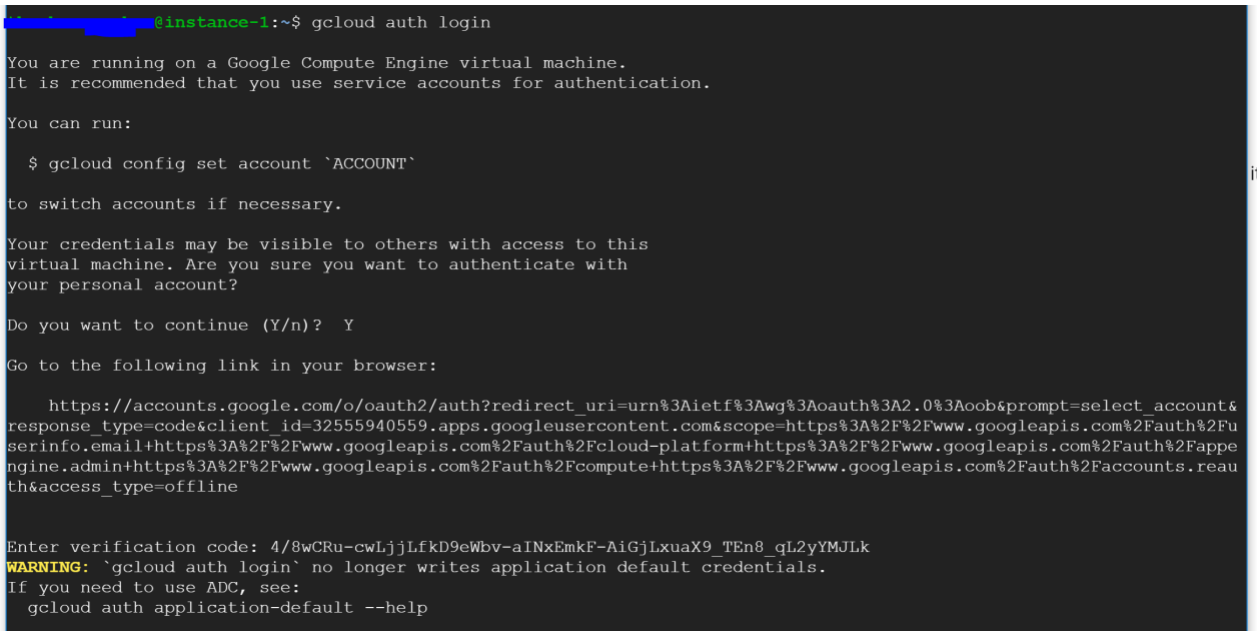
Create table by running these commands in console.

+ use emotions;
+ CREATE TABLE IF NOT EXISTS emotions (
  tweet_id bigint,

sentiment text,
author text,
content text
);

+ quit

Change gcloud authentication to run import task: gcloud auth login => then follow their instruction



Import data from emotion_tweets.csv file(in your Storeage bucket) by running this command:

```
gcloud sql import csv [INSTANCE_NAME] gs://[BUCKET_NAME]/[FILE_NAME] \
                --database=[DATABASE_NAME] --table=[TABLE_NAME]
```

+ gcloud sql import csv googlecloudsql gs://emotion_tweets/emotion_tweets.csv \
--database=emotions --table=emotions

6. Deploy sample app

+ Unzip app.zip file
+ Open main.py and change your mysql information.

```
### TO EDIT ###
sql_host = '35.197.176.134'
sql_connection_name='lexical-archery-231806:australia-southeast1:googlecloudsql'
sql_port = 3306
sql_database = 'currencies'
sql_user = 'root'
sql_password = 'griffith_cloud_learning'
### END TO EDIT ###
```

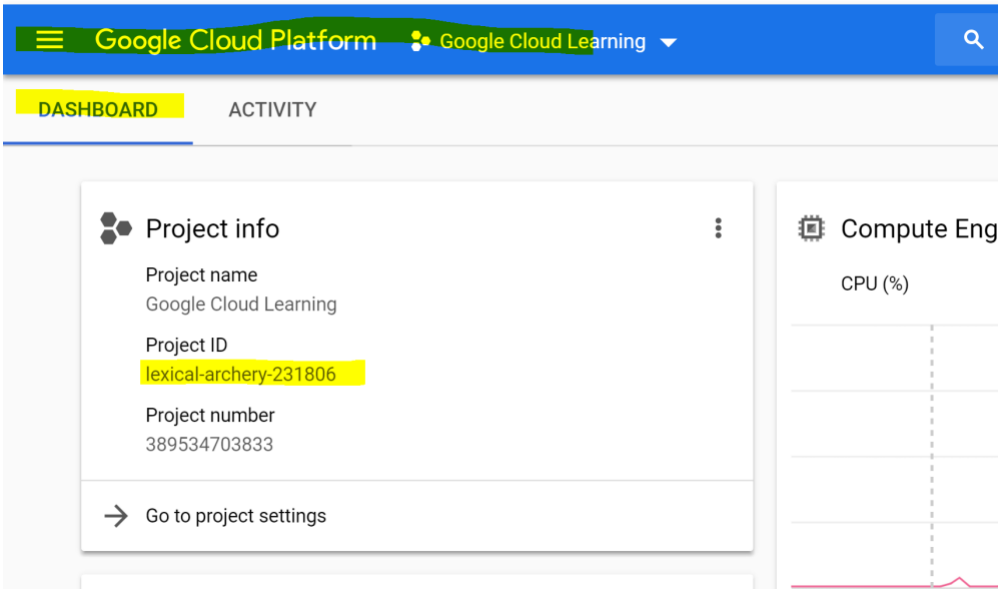
+ Open app.yaml and change to your mysql information.

+ Push it to your github.
+ Install git in your cloud.
    .sudo apt-get update.
    .sudo apt-get install git.
+ In VM console, clone your repository to your local directory (ex: Cloud/Demo).
+ cd to your local directory
+ Install pip3
    . sudo apt-get update
    . sudo apt-get -y install python3-pip
+ Run “pip3 install -r requirements.txt” to install necessary libraries.
+ Upgrade google-cloud-sdk.

```
# Create environment variable for correct distribution
export CLOUD_SDK_REPO="cloud-sdk-$(lsb_release -c -s)"

# Add the Cloud SDK distribution URI as a package source
echo "deb http://packages.cloud.google.com/apt $CLOUD_SDK_REPO main" | sudo tee -a
/etc/apt/sources.list.d/google-cloud-sdk.list

# Import the Google Cloud Platform public key
curl https://packages.cloud.google.com/apt/doc/apt-key.gpg | sudo apt-key add -

# Update the package list and install the Cloud SDK
sudo apt-get update && sudo apt-get install google-cloud-sdk
```

+ Following these steps to activate your server.
    . gcloud auth application-default login

. gcloud config set project [YOUR PROJECT_ID-]



. gcloud init => choose 1. Then follow the instruction (you need to select 38-australia-southeast1-b as your region).

. gcloud app deploy (this will take time)

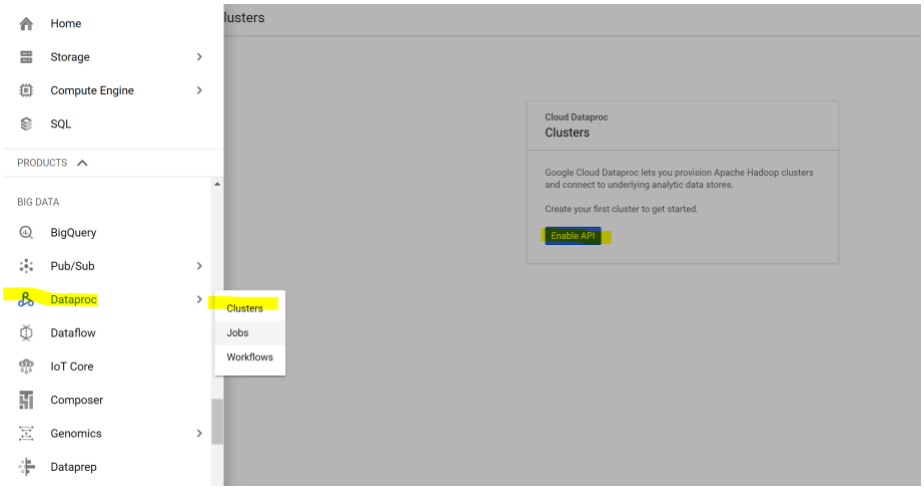
+ Browse your web in browser.

+ If you have any errors, go to Error Reporting in left panel of cloud.

B. Spark – Hello world

You will learn how to submit a job in Google Cloud and execute it.

1. Upload file to bucket
   + Create a bucket in your Google Storage like previous question (ex: **sparklearning**).
   + Upload file hello_world.py to your new bucket
2. Create a new Dataproc Cluster in Google Cloud



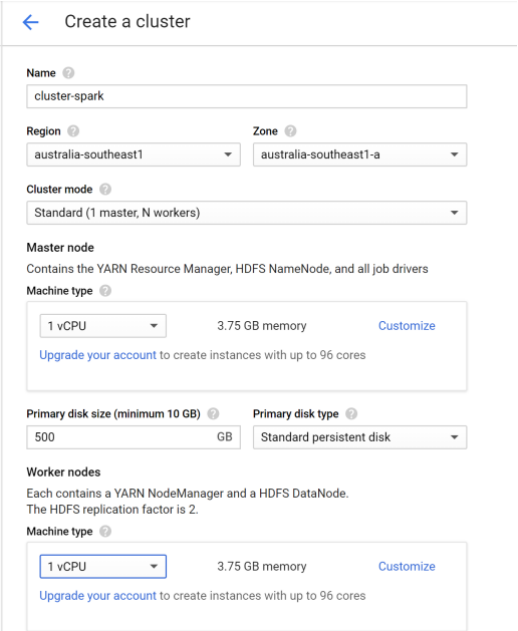
+ Then click create cluster.

. Name: cluster-spark

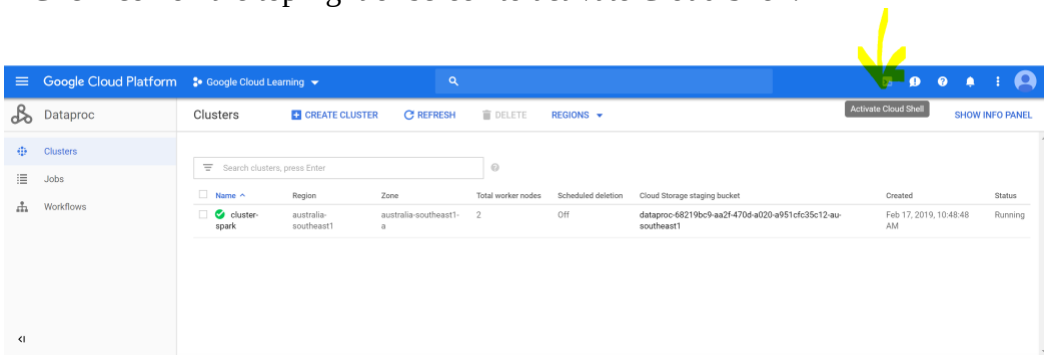
. Region: australia-southeast1, zone: australia-southeast1-a.

. Choose 1vCPU for both master and worker nodes

. Others as default



3. Submit a job via Cloud Shell.
   + Click icon on the top right of screen to activate Cloud Shell.



   + Submit a job via Cloud Shell with this syntax.

```
gcloud dataproc jobs submit job-command \
  --cluster cluster-name --region region \
  job-specific flags and args
```

   . gcloud dataproc jobs submit pyspark --cluster cluster-spark --region australia-southeast1 gs://sparklearning/hello_world.py

   + Wait and observe the result.

C. Practices

- Upload pagerank.py and web-Standford.txt to bucket
- Run PySpark for them

gcloud dataproc jobs submit pyspark --cluster cluster-spark --region australia-southeast1 gs://sparklearning/pagerank.py -- gs://sparklearning/web-Stanford.txt 20

Submission: *Please submit the screenshot of your browser after successfully run pagerank.py. Please include your account's picture (top right corner) as well.*

Sample submission: