



3803ICT
Big Data Analysis

Lab 11 – Data Analytics in Cloud

Trimester 1 - 2022

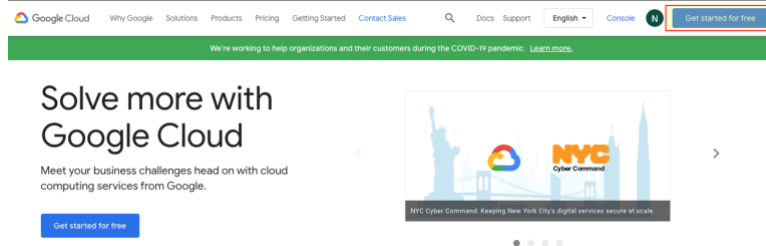
Table of Contents

I. Account Registration	3
II. Deployment	4
A. Tutorial	4
B. Spark – Hello world	14
C. Practices	15

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.

Try Google Cloud Platform for free

Step 1 of 2

Country

Australia

Terms of service

☒ I have read and agree to the [Google Cloud Platform Free Trial Terms of Service](#).

Required to continue

AGREE AND CONTINUE

Access to all Cloud Platform Products

Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

\$300 credit for free

Sign up and get \$300 to spend on Google Cloud Platform over the next 12 months.

No autocharge after free trial ends

We ask you for your credit card to make sure you are not a robot. You won't be charged unless you manually upgrade to a paid account.

Step 2 of 2

Customer info

Account type ⓘ
Business

Name and address ⓘ

Business name

Griffith University

Name

Nguyen

Address line 1

Parklands Dr

Address line 2

SouthPort

City

Gold Coast

State

Queensland

Postal code

4215 ⓘ

Access to all Cloud Platform Products

Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

\$300 credit for free

Sign up and get \$300 to spend on Google Cloud Platform over the next 12 months.

No autocharge after free trial ends

We ask you for your credit card to make sure you are not a robot. You won't be charged unless you manually upgrade to a paid account.

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

Payment method ⓘ

Card number	MM	YY	CVC
#	/		
Card number is required	Month is required	Year is required	CVC is required
Cardholder name			
Cardholder name is required			
<input checked="" type="checkbox"/> Credit or debit card address is same as above			

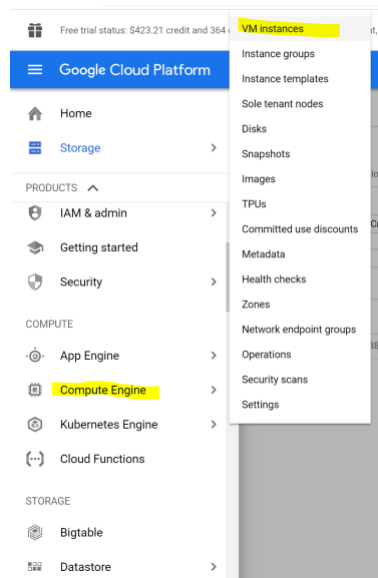


II. Deployment

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



Google Cloud Platform My First Project Search resources and products

← Create an instance

To create a VM instance, select one of the options:

- New VM instance**
Create a single VM instance from scratch
- New VM instance from template**
Create a single VM instance from an existing template
- New VM instance from machine image**
Create a single VM instance from an existing machine image
- Marketplace**
Deploy a ready-to-go solution onto a VM instance

Name (Name is permanent)
instance-1

Labels (Optional)
+ Add label

Region (Region is permanent)
australia-southeast1 (Sydney)

Zone (Zone is permanent)
australia-southeast1-b

Machine configuration

Machine family
General-purpose Memory-optimized
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-standard-1 (1 vCPU, 3.75 GB memory)

	vCPU	Memory
	1	3.75 GB

✓ CPU platform and GPU

Container
☐ Deploy a container image to this VM instance. [Learn more](#)

Boot disk
New 10 GB standard persistent disk
Image
Ubuntu 16.04 LTS [Change](#)

Identity and API access

Service account
Compute Engine default service account

Access scopes
☒ Allow default access
☐ Allow full access to all Cloud APIs
☐ Set access for each API

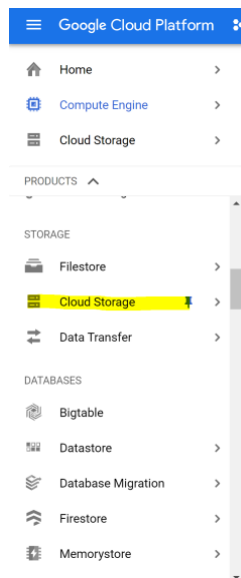
Firewall
Add tags and firewall rules to allow specific network traffic from the Internet
☒ Allow HTTP traffic
☐ Allow HTTPS traffic
✓ Management, security, disks, networking, sole tenancy

Your free trial credit will be used for this VM instance. [GCP Free Tier](#)

[Create](#) [Cancel](#)

You have \$494.70 free trial credits remaining
\$34.98 monthly estimate
That's about \$0.048 hourly
Pay for what you use: No upfront costs and per second billing
[Details](#)

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.

Create a Bucket

emotion_tweets

Tip: Don't include any sensitive information

[CONTINUE](#)

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

australia-southeast1 (Sydney)

[CONTINUE](#)

Monthly cost estimate

Enter values below to check this bucket's monthly cost. For guidance only. [Pricing details](#)

Storage and retrieval

Storage size: GB
\$0.023 per GB-month

Data retrieval size: GB
Free

Operations

Class A operations: per month
\$0.005 per 1,000 ops

Class B operations: per month
\$0.0004 per 1,000 ops

Availability SLA: 99.9%

+ Upload file emotion_tweets.csv to your new bucket.

Bucket details

emotion_tweets

[OBJECTS](#) [CONFIGURATION](#) [PERMISSIONS](#) [RETENTION](#) [LIFECYCLE](#)

[Buckets](#) > emotion_tweets

[UPLOAD FILES](#) [UPLOAD FOLDER](#) [CREATE FOLDER](#) [MANAGE HOLDS](#) [DOWNLOAD](#) [DELETE](#)

Filter by name prefix only [Filter](#) Filter objects and folders

<input type="checkbox"/>	Name	Size	Type	Created time	Storage class	Last modified	Public access	Encryption	Retention
<input checked="" type="checkbox"/>	emotion_tweets.csv	4.2 MB	application/vnd.ms-excel	12 May 2021, ...	Standard	12 May 20...	Not public	Google-managed key	-

3. Create Cloud Sql instance

SQL

emotion_tweets

[OBJECTS](#) [CONFIGURATION](#) [PERMISSIONS](#) [RETENTION](#) [LIFECYCLE](#)

[Buckets](#) > emotion_tweets

[UPLOAD FILES](#) [UPLOAD FOLDER](#) [CREATE FOLDER](#) [MANAGE HOLDS](#) [DOWNLOAD](#) [DELETE](#)

Filter by name prefix only [Filter](#) Filter objects and folders

<input type="checkbox"/>	Name	Size	Type	Created time	Storage class	Last modified	Public access	Encryption	Retention
<input checked="" type="checkbox"/>	emotion_tweets.csv	4.2 MB	application/vnd.ms-excel	12 May 2021, ...	Standard	12 May 20...	Not public	Google-managed key	-

Uploads and Cloud computing operations

emotion_tweets.csv Complete

SQL | Choose engine

Choose your database engine

MySQL
Versions 5.6 or 5.7

[Choose MySQL](#)

PostgreSQL
Version 9.6

[Choose PostgreSQL](#)

For First Generation MySQL instances, [click here](#)

Google Cloud Platform Cloud computing Search products and resources

Create a MySQL instance

Instance info

Instance ID *
googlecloudsql

Use lowercase letters, numbers and hyphens. Start with a letter.

Password *
griffith_cloud_learning

Set a password for the root user. [Learn more](#)

☐ No password

Database version *
MySQL 5.7

Summary

Region	australia-southeast1 (Sydney)
DB version	MySQL 5.7
vCPUs	4 vCPU
Memory	26 GB
Storage	100 GB
Network throughput (MB/s)	1,000 of 2,000
Disk throughput (MB/s)	Read: 48.0 of 240.0 Write: 48.0 of 240.0
IOPS	Read: 3,000 of 15,000 Write: 3,000 of 15,000
Connections	Public IP
Backup	Automated
Availability	Single zone
Point-in-time recovery	Enabled

Choose region and zonal availability

For better performance, keep your data close to the services that need it. Region is permanent, while zone can be changed any time.

Region
australia-southeast1 (Sydney)

☒ Single zone
In case of outage, no failover. Not recommended for production.

☐ Multiple zones (highly available)
Automatic failover to another zone within your selected region. Recommended for production instances. Increases cost. Enables binary logs (required for replication) and automatic backups. Make sure that your storage can support at least seven days of logs.

- + 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.

Google Cloud Platform Cloud computing Search products and resources

SQL Overview EDIT IMPORT EXPORT RESTART STOP DELETE CLONE

PRIMARY INSTANCE

Overview

Connections

Users

Databases

Backups

Replicas

Operations

Connect to this instance

Public IP address
34.87.227.43

Connection name
cloud-computing-313589:australia-southeast1:googlecloudsql

Connect using Cloud Shell

Connect from a Compute Engine VM instance

See all connection methods

Configuration

vCPUs: 4, Memory: 26 GB, SSD storage: 100 GB

Database version is MySQL 5.7

Auto storage increase is enabled

Automated backups are enabled
Stored in Multi-region: us - Data centres in the United States

Point-in-time recovery is enabled

Located in australia-southeast1-a

Not highly available (zonal)

- + Create new database 'emotions'

SQL Databases

MASTER INSTANCE

Overview

Connections

Users

Databases

Backups

Replicas

Operations

All instances > googlecloudsql

googlecloudsql

MySQL Second Generation master

MySQL databases

Create database

Name	Character set	Collation	Type
information_schema	utf8	utf8_general_ci	System
currencies	utf8		
mysql	utf8		
performance_schema	utf8		
sys	utf8		

Create a database

Database name
Needs to follow the MySQL identifier rules.

Character set
utf8

Collation (Optional)
Default collation

CANCEL CREATE

4. Setup permission

- 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)

The screenshot shows the Google Cloud Platform SQL instance overview page. The left sidebar contains a navigation menu with options: Overview, Connections, Users, Databases, Backups, Replicas, and Operations. The main content area is titled 'Overview' and includes a CPU utilization graph showing 2% usage. Below the graph, there are sections for 'Connect to this instance' (with public IP address 34.87.219.122 and instance connection name exalted-beanbag-275312:australia-southeast1:googlecloudsql), 'Connect using Cloud Shell', 'Connect from a Compute Engine VM instance', and 'See all connection methods'. A 'Suggested actions' section lists 'Create a backup' and 'Enable high availability'. The 'Service account' section displays the email p626940988214-bfvjtb@gcp-sa-cloud-sql.iam.gserviceaccount.com. The 'Operations and logs' section shows a table with columns Date/Time, Type, and Status, containing one entry: Apr 25, 2020, 10:58:25 PM, Create database, Database created. The right sidebar contains 'Configuration' details (vCPUs: 1, Memory: 3.75 GB, SSD storage: 10 GB) and 'Maintenance' information (Preferred window, Updates may occur any day of the week, Order of update, Cloud SQL chooses the maintenance timing, Notifications: Off, Upcoming: No maintenance scheduled right now).

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

The screenshot shows the Google Cloud Platform Cloud Storage bucket permissions page for the bucket 'emotion_tweets'. The left sidebar contains a navigation menu with options: Browser, Monitoring, and Settings. The main content area is titled 'Bucket details' and includes tabs for OBJECTS, CONFIGURATION, PERMISSIONS (selected), RETENTION, and LIFECYCLE. The 'PERMISSIONS' tab shows 'Public access' (Not public) and 'Access control' (Uniform: No object-level ACLs enabled, 90 days left to change this setting). Below these, there is a 'Permissions' section with a table showing the bucket's permissions. The table has columns: Type, Member, Name, Role, and Inheritance. The table contains one entry: Editors of project: cloud-computing-313500, Storage Legacy Bucket Owner. The bottom of the page shows a 'Filter' section with a search bar and a table with columns: Type, Member, Name, Role, and Inheritance.

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

griffith-bucket1

Public access: Not public. This bucket is not shared publicly and uniform bucket-level access is enabled. To ensure that the bucket's data does not become public, do not add `allUsers` or `allAuthenticatedUsers` as members.

Access control: Uniform: No object-level ACLs enabled. 90 days left to change this setting. All object access is controlled by bucket permissions. Bucket permissions cannot have their own access control lists (ACLs). If you want to control object access, you can switch to fine-grained access control. [Learn more](#)

Permissions: + ADD - REMOVE

View By: MEMBERS ROLES

Filter: Enter property name or value

Type	Member	Name	Role
Group	Editors of project: glowing-bolt-314023		Storage Legacy Bucket Owner Storage Legacy Object Owner
Group	Owners of project: glowing-bolt-314023		Storage Legacy Bucket Owner Storage Legacy Object Owner
User	p60949844269-028f7c@gcp-sa-cloud-sql.iam.gserviceaccount.com		Storage Object Viewer
Group	Viewers of project: glowing-bolt-314023		Storage Legacy Bucket Reader Storage Legacy Object Reader

Edit permissions

Member: p60949844269-028f7c@gcp-sa-cloud-sql.iam.gserviceaccount.com

Resource: griffith-bucket1

Role: Storage Object Viewer

Condition: [Add condition](#)

Read access to GCS objects.

+ ADD ANOTHER ROLE

SAVE CANCEL

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**)

Google Cloud Platform Cloud computing

Search products and resources

Compute Engine

Virtual machines

- VM instances
- Instance templates
- Sole-tenant nodes
- Machine images
- TPUs
- Migrate for Compute Engi...
- Committed use discounts

Storage

- Disks
- Snapshots

VM instances CREATE INSTANCE IMPORT VM REFRESH START/RESUME STOP

INSTANCES INSTANCE SCHEDULE

Filter: Enter property name or value

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP
✓	instance-1	australia-southeast1-b			10.152.0.2 (nic0)	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.

Google Cloud Platform Cloud computing Search products and resources

SQL

PRIMARY INSTANCE

- Overview
- Connections
- Users
- Databases
- Backups
- Replicas
- Operations

Release Notes

Connections

☐ Private IP
Requires additional APIs and permissions, which may require your system admin. Can't be disabled once enabled. [Learn more](#)

☒ Public IP
Authorise a network or use [Cloud SQL Proxy](#) to connect to this instance. [Learn more](#)

Authorised networks

New network

Name
MyVM

Use [CIDR notation](#)

Network *
34.87.235.255
Example: 199.27.25.0/24

CANCEL DONE

ADD NETWORK

SAVE DISCARD CHANGES

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

Google Cloud Platform Select a project

Register your application for Cloud SQL Admin API in Google Cloud Platform

Google Cloud Platform allows you to manage your application and monitor API usage.

Select a project where your application will be registered
You can use one project to manage all of your applications, or you can create a different project for each application.

Google Cloud Learning

Continue

Follow the instruction to create the service account:

Google Cloud Platform Cloud computing Search products and resources

IAM & Admin

Create service account

1 Service account details

Service account name
emotions
Display name for this service account

Service acc...
emotions @cloud-computing-313500.iam.gserviceaccount.com

Service account description
Describe what this service account will do

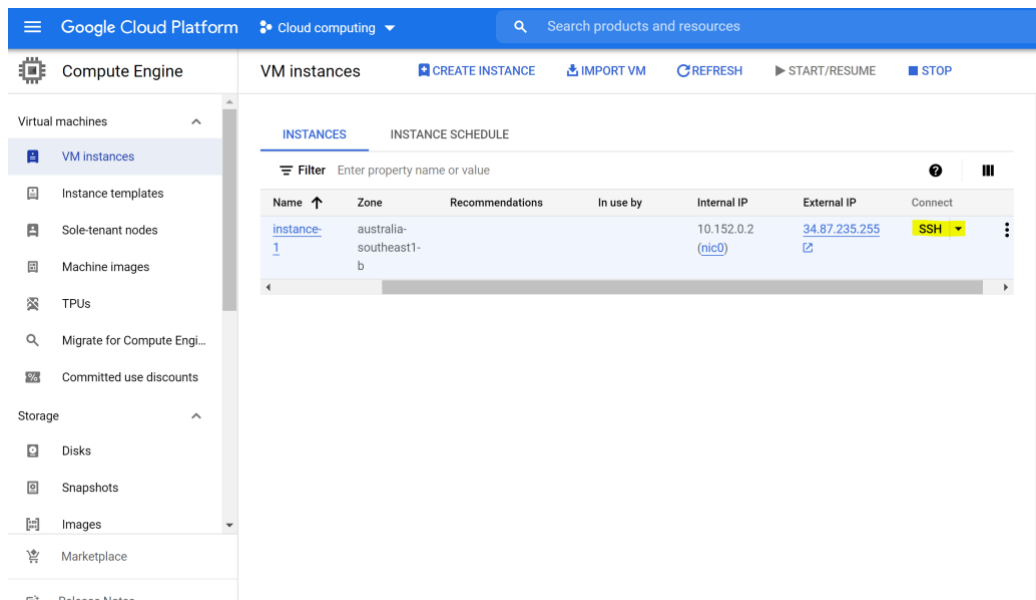
CREATE

2 Grant this service account access to the project (optional)

3 Grant users access to this service account (optional)

DONE CANCEL

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**

```
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 [10.0 kB]
Get:2 http://australia-southeast1.gce.archive.ubuntu.com/ubuntu xenial-updates/main amd64 mysql-client-core-5.7 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 5.7.25-0ubuntu0.16.04.2 [15.5 kB]
Get:4 http://australia-southeast1.gce.archive.ubuntu.com/ubuntu xenial-updates/main amd64 mysql-client-5.7 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 5.7.25-0ubuntu0.16.04.2 [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 (2.23-0ubuntu1) ...
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

```
instance-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

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

```
@instance-1:~$ gcloud auth login

You are running on a Google Compute Engine virtual machine.
It is recommended that you use service accounts for authentication.

You can run:

  $ gcloud config set account `ACCOUNT`

to switch accounts if necessary.

Your credentials may be visible to others with access to this
virtual machine. Are you sure you want to authenticate with
your personal account?

Do you want to continue (Y/n)? Y

Go to the following link in your browser:

  https://accounts.google.com/o/oauth2/auth?redirect_uri=urn%3Aietf%3Aawg%3Aoauth%3A2.0%3Aaob&prompt=select_account&
response_type=code&client_id=32555940559.apps.googleusercontent.com&scope=https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fu
serinfo.email+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fcloud-platform+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fap
engine.admin+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fcompute+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Faccounts.rea
uth&access_type=offline

Enter verification code: 4/8wCRu-cwLjjLfkD9eWbv-aINxEmkF-AiGjLxuaX9_TEn8_qL2yYMJLk
WARNING: `gcloud auth login` no longer writes application default credentials.
If you need to use ADC, see:
  gcloud auth application-default --help
```

Import data from emotion_tweets.csv file(in your Storage 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

```
@instance-1:~$ 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.

```
#[START gae_flex_mysql_env]
env_variables:
  # Replace user, password, database, and instance connection name with the values obtained
  # when configuring your Cloud SQL instance.
  SQLALCHEMY_DATABASE_URI: >-
  mysql+pymysql://root:griffith_cloud_learning@/currencies?unix_socket=/cloudsql/lexical-archery-231806:australia-southeast1:googlecloudsql
#[END gae_flex_mysql_env]

#[START gae_flex_mysql_settings]
# Replace project and instance with the values obtained when configuring your
# Cloud SQL instance.
beta_settings:
  cloud_sql_instances: lexical-archery-231806:australia-southeast1:googlecloudsql
#[END gae_flex_mysql_settings]
```

- + 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

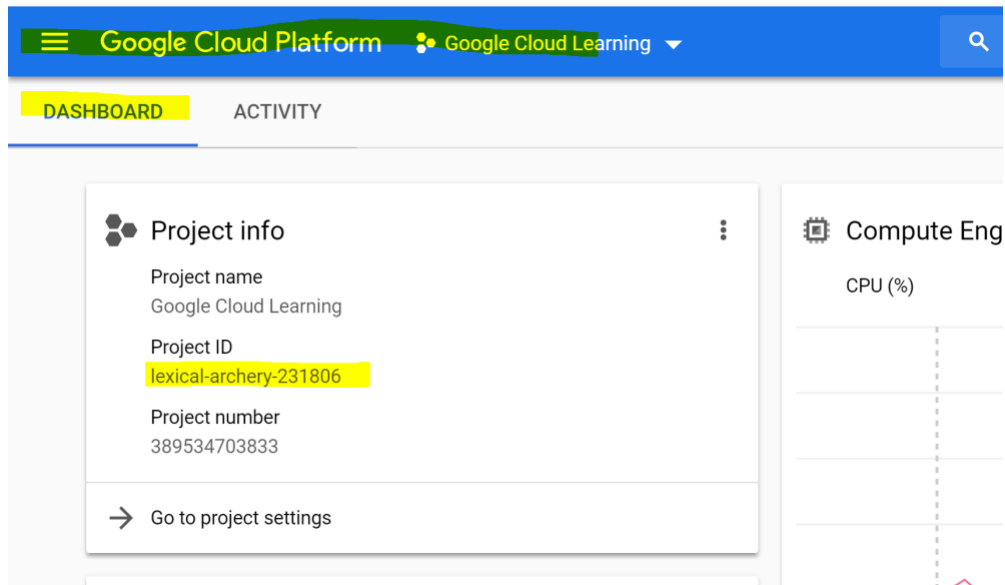
curl https://packages.cloud.google.com/apt/doc/apt-key.gpg | sudo apt-key add -

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)

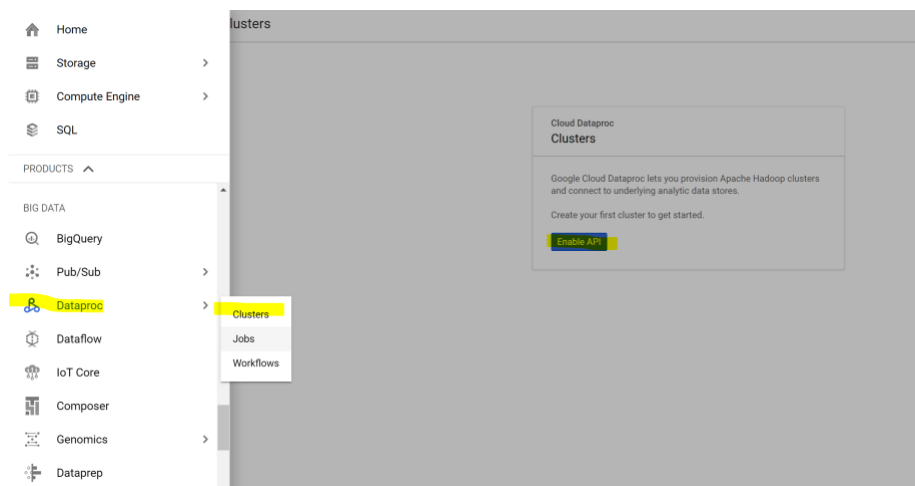
+ Browse your web in browser.

+ 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

← Create a cluster

Name [?]
cluster-spark

Region [?] australia-southeast1 Zone [?] australia-southeast1-a

Cluster mode [?]
Standard (1 master, N workers)

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

Machine type [?]
1 vCPU 3.75 GB memory [Customize](#)
[Upgrade your account](#) to create instances with up to 96 cores

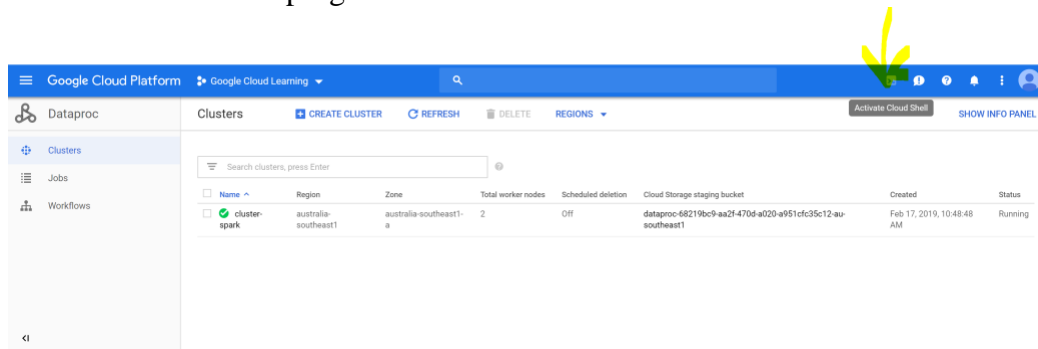
Primary disk size (minimum 10 GB) [?] 500 GB Primary disk type [?] Standard persistent disk

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

Machine type [?]
1 vCPU 3.75 GB memory [Customize](#)
[Upgrade your account](#) to create instances with up to 96 cores

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-Stanford.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:

Free trial status: \$453.55 credit and 91 days remaining. With a full account, you'll get unlimited access to all of Google Cloud Platform.

DISMISS

ACTIVATE

Google Cloud

My First Project

Search (/) for resources, docs, products and more

Search

Account: Ken Quach (kenq.ict@gmail.com)

REFRESH

HELP ASSISTANT

CLEAN

Cloud Storage

Buckets

Marketplace

Release notes

Bucket details

UPLOAD FILES

UPLOAD FOLDER

CREATE FOLDER

TRANSFER DATA

MANAGE HOLDS

DOWNLOAD

DELETE

Filter by name prefix only

Filter

Filter objects and folders

Show deleted data

	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	
<input type="checkbox"/>	hello_world.py	151 B	text/x-python-script	21 May 2023, 23:30:00	Standard	21 May 2023, 23:30:00	Not public	—	
<input type="checkbox"/>	pagerank.py	1.5 KB	text/x-python-script	21 May 2023, 23:46:56	Standard	21 May 2023, 23:46:56	Not public	—	
<input type="checkbox"/>	web-Stanford.txt	31.4 MB	text/plain	21 May 2023, 23:47:13	Standard	21 May 2023, 23:47:13	Not public	—	

CLOUD SHELL

Terminal

ultimate-bit-387412

Open editor

The connection to your Google Cloud Shell was lost.

Close

Reconnect

```
clusterName: cluster-spark
clusterUuid: f85e2678-3fa4-41c5-9f8c-ff401e3df4c9
pysparkJob:
  args:
    - gs://sparklearning_quach/web-Stanford.txt
    - '20'
  mainPythonFileUri: gs://sparklearning_quach/pagerank.py
reference:
  jobId: 661c5c93a0cb4904acfd9c622fc137e
  projectId: ultimate-bit-387412
status:
  state: DONE
  stateStartTime: '2023-05-21T13:55:05.524506Z'
statusHistory:
  - state: PENDING
    stateStartTime: '2023-05-21T13:48:05.960082Z'
  - state: SETUP_DONE
    stateStartTime: '2023-05-21T13:48:05.996026Z'
  - details: Agent reported job success
    state: RUNNING
    stateStartTime: '2023-05-21T13:48:06.746185Z'
yarnApplications:
  - name: PythonPageRank
    progress: 1.0
    state: FINISHED
    trackingUri: http://cluster-spark-m:8088/proxy/application_1684676504211_0002/
kenq_ict@cloudshell:~ (ultimate-bit-387412) $
```