



EECS E6893 Big Data Analytics

Intro to Big Data Analytics on GCP

Qingcheng Yu, qy2281@columbia.edu

Agenda

- GCP
 - Setup
 - Interaction
- Services
 - Cloud Storage
 - BigQuery
 - Dataproc (Spark)
- HW0

GCP

- Cloud computing platform
 - Flexibility: on-demand and scale as you want
 - Efficiency: no need to maintain infra
- Services (relevant to this assignment)
 - Compute
 - Compute Engines: VMs / Servers (automatically created by Dataproc)
 - Big data products
 - BigQuery: Data warehouse for analytics
 - Dataproc: Hadoop and Spark
 - Storage
 - Cloud Storage: Object storage system
 - Much much more at <https://cloud.google.com/products/>



Google Cloud Platform (GCP)

GCP Setup

- Create a google account
- Apply for \$300 credit for the first year: <https://cloud.google.com/free/>
- Go to [Console dashboard](#) -> Billing to check credit is there

Google Cloud

Overview

Solutions

Products

Pricing

Resources

Get \$300 in free credits and free usage of 20+ products →

**Dream, build, and
transform with
Google Cloud**

Build apps faster, make smarter business decisions, and
connect people anywhere.

Go to console

Contact sales

Build what's next. Better software. Faster.

- ✓ Use Google's core infrastructure, data analytics, and machine learning
- ✓ Protect your data and apps with the same security technology Google uses
- ✓ Avoid vendor lock-in and run your apps on open source solutions

[Get started for free](#)[Contact sales](#)

Start running workloads for free


New customers get [\\$300 in free credits](#) to run, test, and deploy workloads. All customers can use [25+ products for free](#). up to monthly usage

Built by developers, for developers

[Start your proof of concept](#) with Google Cloud's easy-to-use platform, tools, and APIs. Explore [pre-built solution templates](#) that you

Estimate your costs

Understand how your costs vary by location, workloads, and other variables with our [pricing calculator](#). Estimate your cloud migration costs

 Try Google Cloud for free

Step 1 of 2 Account Information



Yu Jacky

jackyyu2021111@gmail.com

[SWITCH ACCOUNT](#)

Country

United States

What best describes your organization or needs?

Please select

Other

Terms of Service

I have read and agree to the [Google Cloud Platform Terms of Service](#), [Supplemental Free Trial Terms of Service](#), and the terms of service of [any applicable services and APIs](#).

Required to continue

[CONTINUE](#)

Access to all Google Cloud products

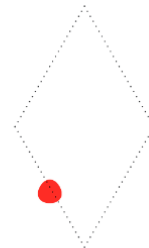
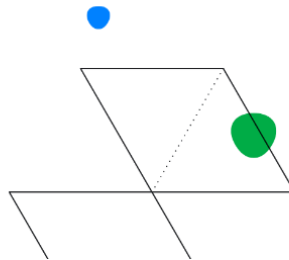
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.


\$300 credit for free

Put Google Cloud to work with \$300 in credit to spend over the next 90 days.

No autocharge after free trial ends

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



 Try Google Cloud for free

Step 2 of 2 Payment Information Verification

Your payment information helps us reduce fraud and abuse. If using a credit or debit card, you won't be charged until you manually activate your account.

Payments profile + ?

[Create new payments profile](#)

SUBMIT

Access to all Google Cloud products

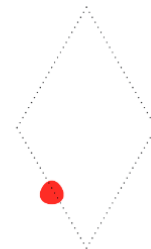
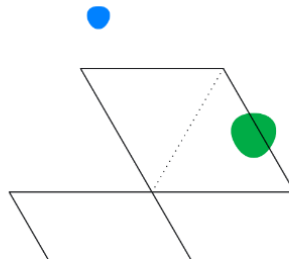
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.


\$300 credit for free

Put Google Cloud to work with \$300 in credit to spend over the next 90 days.

No autocharge after free trial ends

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



 Try Google Cloud for free

Step 2 of 2 Payment Information

Your payment information helps us reduce fraud and a debit card, you won't be charged until you manually activate it.

Payments profile

[Create new payments profile](#)

SUBMIT

Create new payments profile

Only Business accounts can have multiple users. You cannot change the account type after signing up. In some countries, this selection affects your tax options. [Learn more about payments profiles](#). If you choose Individual as your account type, you agree that use of your account is for your trade, business, craft, or profession.

Profile type

Individual

Legal name

Qingcheng Yu

Street address

Apt, suite, etc. (optional)

City

State

Zip code

Cancel

Create

Google Cloud products


need to build and run your apps, services, including Firebase and the

Free

to work with \$300 in credit to use for 90 days.

After free trial ends


to add a credit card to make sure you are using a credit or debit card, you won't be charged until you manually upgrade to a paid

 Try Google Cloud for free

Step 2 of 2 Payment Information Verification

Your payment information helps us reduce fraud and abuse. **If using a credit or debit card, you won't be charged until you manually activate your account.**

Payments profile

Qingcheng Yu [Change](#) 


Individual • United States • ID: 6169-5513-9110

Only Business accounts can have multiple users. You cannot change the account type after signing up. In some countries, this selection affects your tax options.

[Learn more about payments profiles.](#)

If you choose Individual as your account type, you agree that use of your account is for your trade, business, craft, or profession.

Payment method

[Add payment method](#) 

SUBMIT

Access to all Google Cloud products

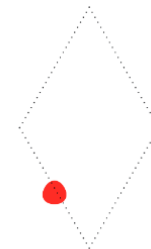
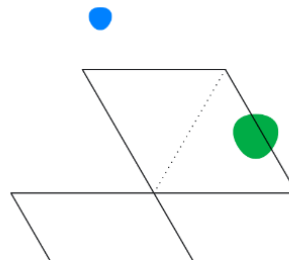
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

\$300 credit for free

Put Google Cloud to work with \$300 in credit to spend over the next 90 days.

No autocharge after free trial ends

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





Welcome, C

PREVIEW

Create or select a project to

[Learn more about projects](#)

CREATE OR SELECT A PROJECT

Explore curated resources to

Recommended based

Pre-built solution templates ?



Deploy a three-tier web app



Deploy load balanced managed VMs



Create a log analysis pipeline

Google Cloud Platform

Welcome Qingcheng YU!

Your free trial includes \$300 in credit to spend over the next 90 days. To help us serve you better, please answer 4 questions.

- ✓ What best describes your organization or needs?
- ✓ What brought you to Google Cloud?
- ✓ What are you interested in doing with Google Cloud?
- 4 What best describes your role?

Please select *

Engineer / Developer ▾

CLOSE

DONE

Welcome, Yu Jacky PREVIEW

You're in Free Trial

0 out of \$300 credits used

Expires December 7, 2023

[What happens when trial ends?](#)

[ACTIVATE FULL ACCOUNT](#)

You're working on project [My First Project](#)

Number: 500791717764 ID: high-science-398401

[Add people to your project](#)

[Set up budget alerts](#)

[Review product spend](#)

Explore curated resources to help you build and deploy your first project

[→](#)

Recommended based on your interest in Data, AI/ML, SAP

Pre-built solution templates

Create a log analysis pipeline

Automatic scaling, VM cluster, lift-and-shift, distributed traffic

Create a data warehouse with BigQuery

Data warehouse, dashboards, ETL, analytics, data analysis

Create an analytics lakehouse

Data science, IOT, streaming analytics

12

- Cloud overview >
- Products & solutions >

PINNED

- APIs & Services >
- Billing**
- IAM & Admin >
- Marketplace
- Compute Engine >
- Kubernetes Engine >
- Cloud Storage >
- BigQuery >
- VPC network >
- Cloud Run
- SQL
- Security >

Welcome, Yu Jacky PREVIEW

You're in Free Trial



0 out of \$300 credits used

Expires December 7, 2023

[What happens when trial ends?](#)

[ACTIVATE FULL ACCOUNT](#)

You're working on project [My First Project](#) ?

Number: 500791717764 ID: high-science-398401

[Add people to your project](#)

[Set up budget alerts](#)

[Review product spend](#)

Explore curated resources to help you build and deploy your first project



Recommended based on your interest in Data, AI/ML, SAP

Build solution templates

Create a log analysis pipeline

Automatic scaling, VM cluster, lift-and-shift, distributed traffic



Create a data warehouse with BigQuery

Data warehouse, dashboards, ETL, analytics, data analysis



Create an analytics lakehouse

Data science, IOT, streaming analytics

Billing

Overview

LEARN

Billing account: My Billing Account

BILLING ACCOUNT OVERVIEW | PAYMENT OVERVIEW

view report

Overview

- Reports
- Cost table
- Cost breakdown
- Commitments
- Commitment analysis
- Budgets & alerts
- Billing export
- Pricing
- Documents
- Transactions
- Payment settings
- Payment method
- Account management

Cost trend

September 1, 2020 - September 30, 2021

Average monthly total cost: \$0.00



Actual cost

view report

Check out your account health results to avoid common billing-related issues and adopt our best practice recommendations. Learn more

0 (red exclamation mark) 1 (yellow lightbulb) 1 (green checkmark)

View all health checks

Free trial credit



\$300

Free trial credit

Out of \$300



91

Days remaining

You will not be billed during your free trial. To keep your projects running after the free trial is up, upgrade to a paid account.

UPGRADE

LEARN MORE

GCP: Create project



- Project: basic unit for creating, enabling, and using all GCP services
 - managing APIs, billing, permissions
 - adding and removing collaborators
- Visit console dashboard or [cloud resource manager](#)
- Click on “create project / new project” and complete the flow
- Ensure billing is pointing to the \$300 credit

Select a project

 **NEW PROJECT**

Search projects and folders

RECENT STARRED ALL

	Name	ID
✓ ☆ 	My First Project 	high-science-398401

CANCEL OPEN

New Project



You have 10 projects remaining in your quota. Request an increase or delete projects. [Learn more](#)

[MANAGE QUOTAS](#)

Project name *
EECS6893

Project ID: eecs6893-398401. It cannot be changed later. [EDIT](#)

Location *
No organization [BROWSE](#)

Parent organization or folder

[CREATE](#) [CANCEL](#)

Notifications

✔ Create Project: EECS6893
[SELECT PROJECT](#)

Just now

Project info

Project name
EECS6893
Project number
755979763552
Project ID
eecs6893-398401

[ADD PEOPLE TO THIS PROJECT](#)

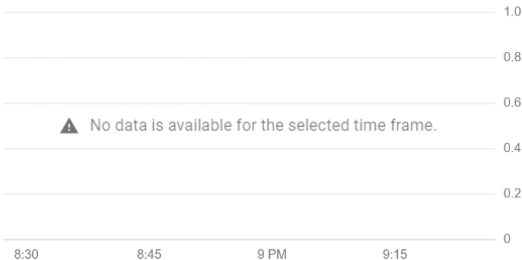
[Go to project settings](#)

Resources

- BigQuery**
Data warehouse/analytics
- SQL**
Managed MySQL, PostgreSQL, SQL Server
- Compute Engine**
VMs, GPUs, TPUs, Disks
- Storage**
Multi-class multi-region object storage
- Cloud Functions**
Event-driven serverless functions
- App Engine**

API APIs

Requests (requests/sec)



[Go to APIs overview](#)

Google Cloud Platform status

Multiple Products
We are investigating an Issue with Vertex AI Search
Began at 2023-09-07 (08:14:54)
All times are US/Pacific
Data provided by status.cloud.google.com

[Go to Cloud status dashboard](#)

Billing

Estimated charges USD \$0.00
For the billing period Sep 1 – 7, 2023

[Take a tour of billing](#)

[View detailed charges](#)

Monitoring

- [Create my dashboard](#)
- [Set up alerting policies](#)
- [Create uptime checks](#)

GCP: Interaction

- [Graphical UI / console](#): Useful to create VMs, set up clusters, provision resources, manage teams, etc
- [Command line tools / Cloud SDK](#): Useful for interacting from local host and using the resources once provisioned. E.x. ssh into instances, submit jobs, copy files, etc
- [Cloud Shell](#): Same as command line, but web-based and pre-installed with SDK and tools

Search in Google: GCP console

Google Cloud

Overview

Solutions

Products

Pricing

Resources



Docs

Support



Console



Contact Us

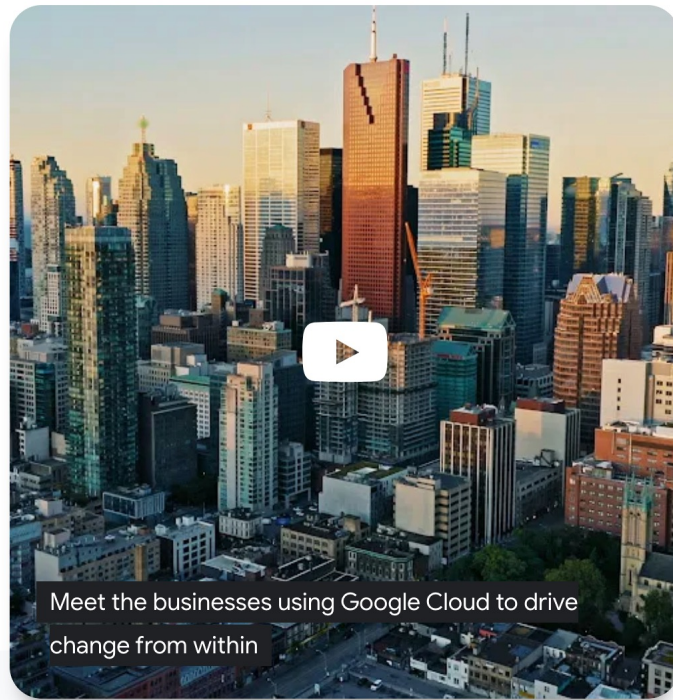
Get \$300 in free credits and free usage of 20+ products [→](#)

Dream, build, and transform with Google Cloud

Build apps faster, make smarter business decisions, and connect people anywhere.

[Go to console](#)

[Contact sales](#)





Free trial status: \$300.00 credit and 91 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

DISMISS

ACTIVATE

Google Cloud

EECS6893

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

Search



Cloud overview >

Dashboard

Products & solutions >

Activity

Recommendations

PINNED

APIs & Services >

Billing >

IAM & Admin >

Marketplace >

Compute Engine >

Kubernetes Engine >

Cloud Storage >

BigQuery >

VPC network >

Cloud Run >

SQL >

Home, Yu Jacky **PREVIEW**

You're in Free Trial

0 out of \$300 credits used

Expires December 7, 2023

What happens when trial ends?

ACTIVATE FULL ACCOUNT

You're working on project [EECS6893](#)

Number: 755979763552 ID: eecs6893-398401

[Add people to your project](#)

[Set up budget alerts](#)

[Review product spend](#)

Explore curated resources to help you build and deploy your first project

Recommended based on your interest in Data, AI/ML, SAP

Build solution templates

Create a log analysis pipeline

Automatic scaling, VM cluster, lift-and-shift,



Create a data warehouse with BigQuery

Data warehouse, dashboards, ETL, analytics,



Create an analytics lakehouse

Data science, IOT, streaming analytics

https://console.cloud.google.com/home/dashboard?hl=en&project=eecs6893-398401

GCP: console

Free trial status: \$300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

DISMISS [ACTIVATE](#)

Google Cloud EECS6893

Search (/) for resources, docs, products, and more [Search](#)

[1](#) [?](#) [Yu](#)

[DASHBOARD](#) [ACTIVITY](#) [RECOMMENDATIONS](#)

[CUSTOMIZE](#)

Search for services here

Project info

Project name
EECS6893

Project number
755979763552

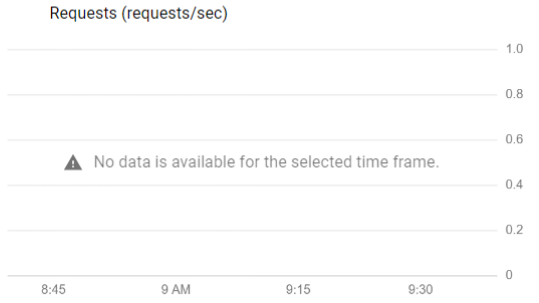
Project ID
eecs6893-398401

[ADD PEOPLE TO THIS PROJECT](#)

[Go to project settings](#)

API APIs

Requests (requests/sec)



[Go to APIs overview](#)

Google Cloud Platform status

Multiple Products

We are investigating an Issue with Vertex AI Search
Began at 2023-09-07 (08:14:54)

All times are US/Pacific
Data provided by status.cloud.google.com

[Go to Cloud status dashboard](#)

- ### Resources
- BigQuery
Data warehouse/analytics
 - SQL
Managed MySQL, PostgreSQL, SQL Server
 - Compute Engine
VMs, GPUs, TPUs, Disks
 - Storage
Multi-class multi-region object storage
 - Cloud Functions

Billing

Estimated charges USD \$0.00

For the billing period Sep 1 – 8, 2023

[Take a tour of billing](#)

[View detailed charges](#)

Monitoring

[Create my dashboard](#)

[Set up alerting policies](#)

Manage / Enable APIs

GCP: Cloud SDK

- Install the SDK that is suitable for your local environment:
<https://cloud.google.com/sdk/docs/quickstarts>
- Some testing after installation:
 - `gcloud info`
 - `gcloud auth list`
 - `gcloud components list`
- Change default config:
 - `gcloud init`

Filter

gcloud CLI

Product overview

gcloud CLI overview

gcloud CLI cheat sheet

Quickstart

Install the Google Cloud CLI

How-to guides

All how-to guides

▶ Installing the gcloud CLI

▶ Setting up the gcloud CLI

Managing gcloud CLI components

Scripting gcloud CLI commands

Enabling accessibility features

Using gcloud interactive shell ▲

Uninstalling the gcloud CLI

Installing the latest gcloud CLI version (445.0.0)

★ **Note:** If you are behind a proxy/firewall, see the [proxy settings](#) page for more information on installation.

Linux

Debian/Ubuntu

Red Hat/Fedora/CentOS

macOS

Windows

Chromebook

The Google Cloud CLI works on Windows 8.1 and later and Windows Server 2012 and later.

1. Download the [Google Cloud CLI installer](#).

Alternatively, open a PowerShell terminal and run the following PowerShell commands:

```
(New-Object Net.WebClient).DownloadFile("https://dl.google.com/dl/cloudsdk/channels/rapid/cmdline_windows_x86.exe") & $env:Temp\GoogleCloudSDKInstaller.exe
```

2. Launch the installer and follow the prompts. The installer is signed by Google LLC.

If you're using a screen reader, check the **Turn on screen reader mode** checkbox. This option configures `gcloud` to use status trackers instead of unicode spinners, display progress as a percentage, and flatten tables. For more information, see the [Accessibility features guide](#).

3. Cloud SDK requires Python; supported versions are Python 3 (3.5 to 3.9). By default, the Windows version of Cloud SDK comes bundled with Python 3. To use Cloud SDK, your operating system must be able to run a

On this page

Before you begin

[Installing the latest gcloud CLI version \(445.0.0\)](#)

Optional: Install the latest Cloud Client Libraries

Initializing the gcloud CLI

Running core commands

Clean up

What's next

```
Pick configuration to use:  
[1] Re-initialize this configuration [default] with new settings  
[2] Create a new configuration  
Please enter your numeric choice:
```

```
Choose the account you would like to use to perform operations for this configuration:  
[1] qy2281@columbia.edu  
[2] Log in with a new account  
Please enter your numeric choice:
```

```
Pick cloud project to use:  
[1] eecs6893-398401  
[2] high-science-398401  
[3] resonant-time-398400  
[4] Enter a project ID  
[5] Create a new project  
Please enter numeric choice or text value (must exactly match list item):
```

Follow the instruction on the website. If you have a previous account, please select the correct account and project

```
C:\Users\11518\AppData\Local\Google\Cloud SDK>gcloud info
```

```
Google Cloud SDK [402.0.0]
```

```
Platform: [Windows, x86_64] uname_result(system='Windows', node='LAPTOP-LJ07H8BA', release='10', version='10.0.19044', machine='AMD64')
```

```
Locale: ('zh_CN', 'cp1252')
```

```
Python Version: [3.9.12 (tags/v3.9.12:b28265d, Mar 23 2022, 23:52:46) [MSC v.1929 64 bit (AMD64)]]
```

```
Python Location: [C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\platform\bundlepython\python.exe]
```

```
OpenSSL: [OpenSSL 1.1.1n 15 Mar 2022]
```

```
Requests Version: [2.25.1]
```

```
urllib3 Version: [1.26.9]
```

```
Site Packages: [Disabled]
```

```
Installation Root: [C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk]
```

```
Installed Components:
```

```
  beta: [2022.09.12]
```

```
  bq: [2.0.75]
```

```
  core: [2022.09.12]
```

```
  gcloud-crc32c: [1.0.0]
```

```
  gsutil: [5.13]
```

```
System PATH: [C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\bin\..\bin\sdk;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\bin;C:\Program Files\Common Files\Oracle\Java\javapath;C:\Program Files (x86)\Common Files\Oracle\Java\javapath;E:\manager wizard\ChemScript\Lib;C:\Program Files (x86)\Intel\iCLS Client;C:\Program Files\Intel\iCLS Client;C:\Windows\system32;C:\Windows;C:\Windows\System32\Wbem;C:\Windows\System32\WindowsPowerShell\v1.0;C:\Program Files (x86)\NVIDIA Corporation\PhysX\Common;C:\Program Files (x86)\Intel\Intel(R) Management Engine Components\DAL;C:\Program Files\Intel\Intel(R) Management Engine Components\DAL;C:\Program Files (x86)\Intel\Intel(R) Management Engine Components\IPT;C:\Program Files\Intel\Intel(R) Management Engine Components\IPT;C:\WINDOWS\system32;C:\WINDOWS;C:\WINDOWS\System32\Wbem;C:\WINDOWS\System32\WindowsPowerShell\v1.0;C:\WINDOWS\System32\OpenSSH;C:\Program Files (x86)\Wolfram Research\WolframScript;E:\MATLAB_R2020a\bin;C:\Program Files\Java\jdk1.8.0_202\bin;E:\Hadoop\hadoop-3.2.4\hadoop-3.2.4\bin;E:\Hadoop\hadoop-3.2.4\hadoop-3.2.4\sbin;E:\Git\cmd;E:\Git\Git LFS;E:\Node.js;C:\ProgramData\chocolatey\bin;E:\Program Files (x86)\Eclipse\Sumo\bin;E:\Program Files (x86)\Eclipse\Sumo\tools;C:\Program Files\MySQL\MySQL Shell 8.0\bin;E:\python\python3.9.1\Scripts;E:\python\python3.9.1;C:\Users\11518\AppData\Local\Microsoft\WindowsApps;E:\?????MikTeX\miktex\bin\x64;C:\Program Files\MPICH2\bin;E:\LAMMPS 64-bit 14May2021\bin;E:\Pycharm\PyCharm Community Edition 2022.1.4\bin;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\bin;C:\Users\11518\AppData\Roaming\TinyTeX\bin\win32;E:\Microsoft VS Code\bin;C:\Users\11518\AppData\Local\GitHubDesktop\bin;E:\Fiddler;C:\Users\11518\AppData\Local\Pandoc;C:\Users\11518\AppData\Roaming\npm;]
```

```
Python PATH: [C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\lib\third_party;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\lib;E:\LAMMPS 64-bit 14May2021\Python;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\platform\bundlepython\python39.zip;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\platform\bundlepython\lib;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\platform\bundlepython]
```

```
Cloud SDK on PATH: [True]
```

```
Kubectl on PATH: [False]
```

```
Installation Properties: [C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\properties]
```

```
User Config Directory: [C:\Users\11518\AppData\Roaming\gcloud]
```

```
Active Configuration Name: [default]
```

```
Active Configuration Path: [C:\Users\11518\AppData\Roaming\gcloud\configurations\config_default]
```

```
Account: [jackyyu2021111@gmail.com]
```

```
Project: [eecs6893-398401]
```

```
Current Properties:
```

```
  [accessibility]
```

```
  screen_reader: [False] (property file)
```

```
  [core]
```

```
  account: [jackyyu2021111@gmail.com] (property file)
```

```
C:\Users\11518\AppData\Local\Google\Cloud SDK>gcloud auth list
```

```
Credentialed Accounts
```

```
ACTIVE ACCOUNT
* jackkyu2021111@gmail.com
  qy2281@columbia.edu
```

```
To set the active account, run:
```

```
$ gcloud config set account `ACCOUNT`
```

```
C:\Users\11518\AppData\Local\Google\Cloud SDK>gcloud components list
```

```
Your current Google Cloud CLI version is: 402.0.0
```

```
The latest available version is: 445.0.0
```

Components			
Status	Name	ID	Size
Update Available	BigQuery Command Line Tool	bq	1.6 MiB
Update Available	Cloud Storage Command Line Tool	gsutil	11.3 MiB
Update Available	Google Cloud CLI Core Libraries	core	21.7 MiB
Update Available	gcloud Beta Commands	beta	< 1 MiB
Not Installed	App Engine Go Extensions	app-engine-go	4.6 MiB
Not Installed	Appctl	appctl	18.7 MiB
Not Installed	Artifact Registry Go Module Package Helper	package-go-module	< 1 MiB
Not Installed	Cloud Bigtable Command Line Tool	cbt	11.4 MiB
Not Installed	Cloud Bigtable Emulator	bigtable	7.0 MiB
Not Installed	Cloud Datastore Emulator	cloud-datastore-emulator	36.2 MiB
Not Installed	Cloud Firestore Emulator	cloud-firestore-emulator	42.5 MiB
Not Installed	Cloud Pub/Sub Emulator	pubsub-emulator	61.2 MiB
Not Installed	Cloud Run Proxy	cloud-run-proxy	12.0 MiB
Not Installed	Cloud SQL Proxy	cloud_sql_proxy	7.4 MiB
Not Installed	Google Container Registry's Docker credential helper	docker-credential-gcr	1.8 MiB
Not Installed	Log Streaming	log-streaming	12.4 MiB
Not Installed	Minikube	minikube	34.5 MiB
Not Installed	Skaffold	skaffold	22.8 MiB
Not Installed	Terraform Tools	terraform-tools	66.2 MiB
Not Installed	anthos-auth	anthos-auth	20.5 MiB
Not Installed	config-connector	config-connector	56.9 MiB
Not Installed	enterprise-certificate-proxy	enterprise-certificate-proxy	6.5 MiB
Not Installed	gcloud Alpha Commands	alpha	< 1 MiB
Not Installed	gcloud app Java Extensions	app-engine-java	65.1 MiB
Not Installed	gcloud app PHP Extensions	app-engine-php	19.1 MiB
Not Installed	gcloud app Python Extensions	app-engine-python	8.5 MiB
Not Installed	gcloud app Python Extensions (Extra Libraries)	app-engine-python-extras	27.3 MiB
Not Installed	gke-gcloud-auth-plugin	gke-gcloud-auth-plugin	8.0 MiB
Not Installed	kubectl	kubectl	< 1 MiB
Not Installed	kubectl-oidc	kubectl-oidc	20.5 MiB

GCP: Cloud Shell

The screenshot shows the Google Cloud Platform dashboard. At the top, there is a navigation bar with the Google Cloud logo, a search bar, and a project selector set to 'EECS6893'. A red box highlights a notification icon in the top right corner. To the right of the notification icon are buttons for 'DISMISS' and 'ACTIVATE', along with a notification count of '1'. Below the navigation bar are tabs for 'DASHBOARD', 'ACTIVITY', and 'RECOMMENDATIONS'. The main content area is divided into three panels: 'Project info' (showing project name, number, and ID), 'API APIs' (showing a graph for 'Requests (requests/sec)' with a warning that no data is available), and 'Google Cloud Platform status' (showing a message about an issue with Vertex AI Search).

Free trial status: \$300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

Google Cloud EECS6893 Search (/) for resources, docs, products, and more Search

DASHBOARD ACTIVITY RECOMMENDATIONS

DISMISS ACTIVATE 1 ? Yu

CUSTOMIZE

Project info

- Project name: EECS6893
- Project number: 755979763552
- Project ID: eeCS6893-398401

[ADD PEOPLE TO THIS PROJECT](#)

API APIs

Requests (requests/sec)

No data is available for the selected time frame.

Google Cloud Platform status

Multiple Products

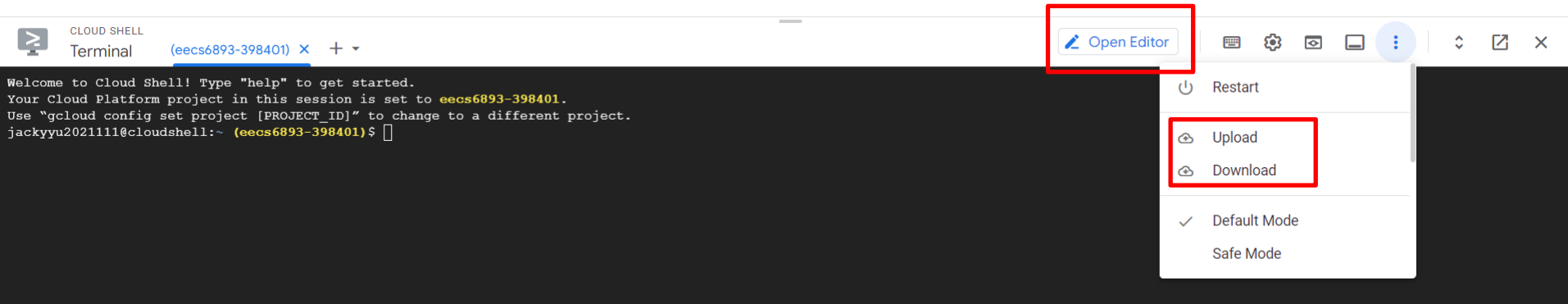
We are investigating an Issue with Vertex AI Search
Began at 2023-09-07 (08:14:54)

All times are US/Pacific
Data provided by status.cloud.google.com

[Go to Cloud status dashboard](#)

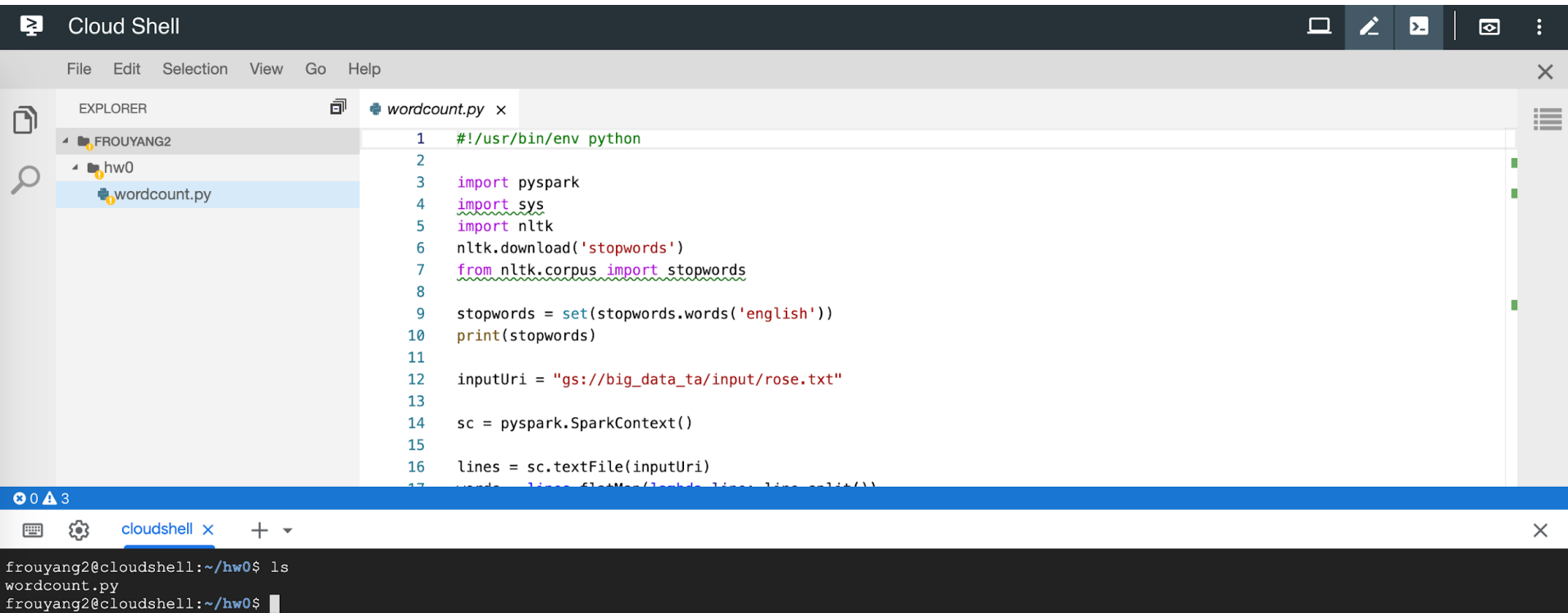
persistent home directory :). The most useful way to complete the HW0

GCP: Cloud Shell



Files can be uploaded through Cloud Storage, which will be introduced later

GCP: Cloud Shell Code Editor



The screenshot displays the Google Cloud Shell environment. At the top, the title bar reads "Cloud Shell" with standard window controls. Below it is a menu bar with "File", "Edit", "Selection", "View", "Go", and "Help". The main workspace is divided into two panes. The left pane, titled "EXPLORER", shows a file tree with a folder named "FROUYANG2" containing a sub-folder "hw0" and a file "wordcount.py". The right pane shows the code editor for "wordcount.py" with the following Python code:

```
1  #!/usr/bin/env python
2
3  import pyspark
4  import sys
5  import nltk
6  nltk.download('stopwords')
7  from nltk.corpus import stopwords
8
9  stopwords = set(stopwords.words('english'))
10 print(stopwords)
11
12 inputUri = "gs://big_data_ta/input/rose.txt"
13
14 sc = pyspark.SparkContext()
15
16 lines = sc.textFile(inputUri)
17 words = lines.flatMap(lambda line: line.split())
```

At the bottom, a terminal window shows the command prompt "frouyang2@cloudshell:~/hw0\$ ls" and the output "wordcount.py".



Cloud Storage

Cloud Storage

- Online file storage system
- Graphical UI through console
- Command line tool: `gsutil`

```
(base) dyn-160-39-199-154:~ xinjianzhanghu$ gsutil
Usage: gsutil [-D] [-DD] [-h header]... [-i service_account] [-m] [-o section:flag=value]... [-q] [-u user_project] [command [opts...] args...]
Available commands:
acl           Get, set, or change bucket and/or object ACLs
autoclass     Configure auto-class feature
bucketpolicyonly  Configure uniform bucket-level access
cat           Concatenate object content to stdout
compose       Concatenate a sequence of objects into a new composite object.
config        Obtain credentials and create configuration file
cors          Get or set a CORS JSON document for one or more buckets
cp            Copy files and objects
defacl        Get, set, or change default ACL on buckets
defstorageclass  Get or set the default storage class on buckets
du            Display object size usage
hash          Calculate file hashes
help          Get help about commands and topics
hmac          CRUD operations on service account HMAC keys.
iam           Get, set, or change bucket and/or object IAM permissions.
kms           Configure Cloud KMS encryption
label         Get, set, or change the label configuration of a bucket.
lifecycle     Get or set lifecycle configuration for a bucket
logging       Configure or retrieve logging on buckets
ls            List providers, buckets, or objects
mb            Make buckets
mv            Move/rename objects
notification  Configure object change notification
pap           Configure public access prevention
perfdiag     Run performance diagnostic
rb            Remove buckets
requesterpays  Enable or disable requester pays for one or more buckets
retention     Provides utilities to interact with Retention Policy feature.
rewrite       Rewrite objects
rm            Remove objects
rpo           Configure replication
rsync         Synchronize content of two buckets/directories
setmeta       Set metadata on already uploaded objects
signurl       Create a signed URL
stat          Display object status
test         Run gsutil unit/integration tests (for developers)
ubla         Configure Uniform bucket-level access
update        Update to the latest gsutil release
```

Cloud Storage

The image shows a screenshot of the Google Cloud console interface. At the top, there is a navigation bar with the Google Cloud logo, a project dropdown menu set to 'EECS6893', a search bar with the text 'Search (/) for resources, docs, products, and more', and a search button. On the right side of the navigation bar, there are icons for notifications (a green circle with '1'), help, and a user profile icon labeled 'Yu'. Below the navigation bar is a sidebar menu with the following items: 'Cloud overview', 'Products & solutions', 'PINNED', 'APIs & Services', 'Billing', 'IAM & Admin', 'Marketplace', 'Compute Engine', 'Kubernetes Engine', 'Cloud Storage' (highlighted with a red underline), 'BigQuery', 'VPC network', 'Cloud Run', and 'SQL'. A sub-menu is open for 'Cloud Storage', showing 'Buckets', 'Monitoring', and 'Settings'. The main content area is currently blank. At the bottom of the page, the URL is visible: <https://console.cloud.google.com/storage/browser?hl=en&project=eecs6893-398401>.

Cloud Storage

Google Cloud EEC56893 Search (/) for resources, docs, products, and more

Cloud Storage Buckets **+ CREATE** REFRESH HELP ASSISTANT LEARN

Transfer New

Power near real-time analytics and replication with event-driven transfers

You can now capture changes faster at your Google Cloud Storage and Amazon S3 sources via event-driven transfers, enabling you to act on your data in near real time. To get started, create a transfer job with a Pub/Sub- or AWS SQS-based event stream configured to send event notifications when objects are created or updated.

[CREATE TRANSFER JOB](#) [LEARN MORE](#)

Analytics New

Preview the new Cloud Storage monitoring dashboard

Check out the new Cloud Storage monitoring dashboard and bucket observability pages! Powered by Cloud Operations, you can customize these dashboards for each project.

[TRY NOW](#)

Filter Filter buckets

<input type="checkbox"/>	Name ↑	Created	Location type	Location	Default storage class ?	Last modified	Public access ?	Access control ?	Prote
No rows to display									

Cloud Storage

Free trial status: \$300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

DISMISS

ACTIVATE

Google Cloud

EECS6893

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

Search



Name your own bucket

Cloud Storage

Create a bucket

HELP ASSISTANT

- Buckets
- Monitoring
- Settings

Name your bucket

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

Tip: Don't include any sensitive information

LABELS (OPTIONAL)

CONTINUE

Choose where to store your data

Location: us (multiple regions in United States)

Location type: Multi-region

Choose a storage class for your data

Default storage class: Standard

Choose how to control access to objects

Public access prevention: On

Access control: Uniform

Choose how to protect object data

Protection tools: None

Good to know

Location pricing

Storage rates vary depending on the storage class of your data and location of your bucket. [Pricing details](#)

Current configuration: Multi-region / Standard

Item	Cost
us (multiple regions in United States)	\$0.026 per GB-month
With default replication	\$0.020 per GB written

ESTIMATE YOUR MONTHLY COST

Marketplace

Release Notes

<

Cloud Storage

Free trial status: \$300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

DISMISS

ACTIVATE

Google Cloud

EECS6893

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

Search



1



Yu

Cloud Storage

Create a bucket

HELP ASSISTANT

Buckets

Monitoring

Settings

Marketplace

Release Notes

<

✓ Name your bucket

Name: 6893_ta

• Choose where to store your data

This choice defines the geographic placement of your data and affects cost, performance, and availability. Cannot be changed later. [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

us-east1 (South Carolina)

CONTINUE

• Choose a storage class for your data

Default storage class: Standard

• Choose how to control access to objects

Public access prevention: On

Good to know

Location pricing

Storage rates vary depending on the storage class of your data and location of your bucket. [Pricing details](#)

Current configuration: Region / Standard

Item	Cost
us-east1 (South Carolina)	\$0.020 per GB-month

ESTIMATE YOUR MONTHLY COST

Cloud Storage

Free trial status: \$300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

DISMISS

ACTIVATE

Google Cloud

EECS6893

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

Search



1



Yu

Cloud Storage

Create a bucket

HELP ASSISTANT

Buckets

Monitoring

Settings

Marketplace

Release Notes

<

Choose a storage class for your data

A storage class sets costs for storage, retrieval, and operations, with minimal differences in uptime. Choose if you want objects to be managed automatically or specify a default storage class based on how long you plan to store your data and your workload or use case. [Learn more](#)

Autoclass
Automatically transitions each object to hotter or colder storage based on object-level activity, to optimize for cost and latency. Recommended if usage frequency may be unpredictable. Can be changed to a default class at any time. [Pricing details](#)

Set a default class
Applies to all objects in your bucket unless you manually modify the class per object or set object lifecycle rules. Best when your usage is highly predictable. Can't be changed to Autoclass once the bucket is created.

Standard
Best for short-term storage and frequently accessed data

Nearline
Best for backups and data accessed less than once a month

Coldline
Best for disaster recovery and data accessed less than once a quarter

Archive
Best for long-term digital preservation of data accessed less than once a year

CONTINUE

Choose how to control access to objects

Public access prevention: On

Item	Cost
us-east1 (South Carolina)	\$0.020 per GB-month

ESTIMATE YOUR MONTHLY COST

Cloud Storage

Free trial status: \$300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

DISMISS **ACTIVATE**

Google Cloud

EECS6893

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

Search



1



Yu

Cloud Storage

Create a bucket

HELP ASSISTANT

Buckets

Monitoring

Settings

Default storage class: Standard

us-east1 (South Carolina)

\$0.020 per GB-month

Choose how to control access to objects

Prevent public access

Restrict data from being publicly accessible via the internet. Will prevent this bucket from being used for web hosting. [Learn more](#)

Enforce public access prevention on this bucket

Access control

Uniform

Ensure uniform access to all objects in the bucket by using only bucket-level permissions (IAM). This option becomes permanent after 90 days. [Learn more](#)

Fine-grained

Specify access to individual objects by using object-level permissions (ACLs) in addition to your bucket-level permissions (IAM). [Learn more](#)

CONTINUE

ESTIMATE YOUR MONTHLY COST

Choose how to protect object data

Protection tools: None

Data encryption: Google-managed

CREATE

CANCEL

<

Cloud Storage

Free trial status: \$300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform. DISMISS ACTIVATE

Google Cloud EECS6893 Search (/) for resources, docs, products, and more Search

Cloud Storage ← Bucket details REFRESH HELP ASSISTANT LEARN

6893_ta

Location: us-east1 (South Carolina) | Storage class: Standard | Public access: Not public | Protection: None

OBJECTS | CONFIGURATION | PERMISSIONS | PROTECTION | LIFECYCLE | OBSERVABILITY | INVENTORY REPORTS

Buckets > 6893_ta

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	
data_citibike_stations.csv	114.3 KB	text/csv	Sep 8, 2023, 10:14:43 AM	Standard	Sep 8, 2023, 10:14:43 AM	Not public	—	

dataset provided in HW0 details

Click on the uploaded dataset file

Cloud Storage

Free trial status: \$300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

DISMISS **ACTIVATE**

Google Cloud EEC56893 Search (/) for resources, docs, products, and more Search

Cloud Storage Object details HELP ASSISTANT LEARN

Buckets > 6893_ta > data_citibike_stations.csv

LIVE OBJECT VERSION HISTORY

DOWNLOAD EDIT METADATA EDIT ACCESS DELETE

Overview

Type	text/csv
Size	114.3 KB
Created	Sep 8, 2023, 10:14:43 AM
Last modified	Sep 8, 2023, 10:14:43 AM
Storage class	Standard
Custom time	—
Public URL	Not applicable
Authenticated URL	https://storage.cloud.google.com/6893_ta/data_citibike_stations.csv
gsutil URI	gs://6893_ta/data_citibike_stations.csv

Permissions

Public access	Not public
---------------	------------

Protection

Version history	—
Retention policy	None
Hold status	None
Encryption type	Google-managed

Uniform Resource Identifier, like a *filepath* on GCP, use this in your program

Cloud Storage - gsutil

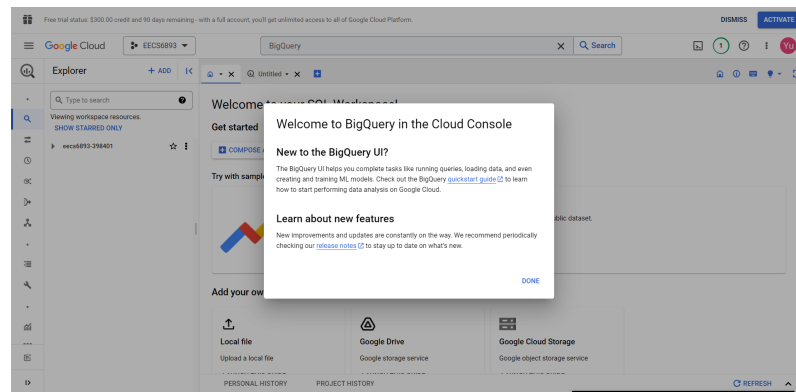
- Interact with Cloud Storage through command line
- Works similar to unix command line
- Useful commands:
 - Concatenate object content to stdout:
`gsutil cat [-h] url...`
 - Copy file:
`gsutil cp [OPTION]... src_url dst_url`
 - List files:
`gsutil ls [OPTION]... url...`
- Explore more at <https://cloud.google.com/storage/docs/gsutil>



BigQuery

BigQuery

- Data warehouse for analytics
- SQL-like languages to interact with DB
- RESTful APIs / client libraries for programmatic access
- Graphical UI



search for BigQuery and go for it

BigQuery

Free trial status: \$300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

DISMISS **ACTIVATE**

Google Cloud EECS6893 BigQuery Search

Explorer + ADD

Welcome to your SQL Workspace!

Get started

COMPOSE A NEW QUERY + ADD

Create dataset

Try the Google Trends Demo Query

Try the Google Trends Demo Query

This simple query generates the top search terms in the US from the Google Trends public dataset.

OPEN THIS QUERY VIEW DATASET

Add your own data

Local file Upload a local file

Google Drive Google storage service

Google Cloud Storage Google object storage service

PERSONAL HISTORY PROJECT HISTORY REFRESH

BigQuery

The screenshot shows the Google Cloud BigQuery Explorer interface. At the top, there's a navigation bar with the Google Cloud logo, the account ID 'EECS6893', and the 'BigQuery' service name. Below this is an 'Explorer' sidebar with a search bar and a list of workspace resources, including the project 'eecs6893-398401'. The main content area is titled 'Welcome to your SQL Workspace!' and features a 'Get started' section with buttons for 'COMPOSE A NEW QUERY' and 'ADD'. Below that is a 'Try with sample data' section with a 'Try the Google Trends Demo Query' card, which includes a colorful bar chart icon and buttons for 'OPEN THIS QUERY' and 'VIEW DATASET'. At the bottom, there's an 'Add your own data' section with options for 'Local file' and 'Google Drive'. The interface also shows 'PERSONAL HISTORY' and 'PROJECT HISTORY' tabs at the very bottom.

Create dataset

Project ID
eecs6893-398401

[CHANGE](#)

Dataset ID *
dataset1

Letters, numbers, and underscores allowed

Location type

Region

Specify a region to colocate your datasets with other Google Cloud services.

Multi-region

Allow BigQuery to select a region within a group to achieve higher quota limits.

Multi-region *

US (multiple regions in United States)

Default table expiration

Enable table expiration

Default maximum table age

Days

Advanced options

[CREATE DATASET](#)

[CANCEL](#)

BigQuery

Free trial status: \$300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

DISMISS **ACTIVATE**

Google Cloud

EECS6893

BigQuery

Search

Explorer

+ ADD

⌂

⌂ - x

Untitled - x

+

⌂

1

?

⋮

Yu

Type to search

Viewing workspace resources.

SHOW STARRED ONLY

eeecs6893-398401

dataset1

SHOW MORE

☆ ⋮

☆ ⋮

Open

Open in

Create table

Share

Copy ID

Refresh contents

Delete

Welcome to your SQL Workspace!

Get started

+ COMPOSE A NEW QUERY

+ ADD

Open in

Open in

Create table

Share

Copy ID

Refresh contents

Delete

Try the Google Trends Demo Query

This simple query generates the top search terms in the US from the Google Trends public dataset.

OPEN THIS QUERY

VIEW DATASET

Local file

Upload a local file



Google Drive



Google Cloud Storage

Google object storage service

"dataset1" created.

GO TO DATASET

×

PERSONAL HISTORY

PROJECT HISTORY

REFRESH

^

BigQuery

Free trial status: \$300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

DISMISS [ACTIVATE](#)

Google Cloud EECS6893 BigQuery Search

Explorer + ADD K dataset1

dataset1

CREATE TABLE SHARING COPY DELETE REFRESH

Dataset info

[EDIT DETAILS](#)

Dataset ID	eeecs6893-398401.dataset1
Created	Sep 8, 2023, 10:19:11 AM UTC-4
Default table expiration	Never
Last modified	Sep 8, 2023, 10:19:11 AM UTC-4
Data location	US
Description	
Default collation	
Default rounding mode	ROUNDING_MODE_UNSPECIFIED
Time travel window	7 days
Case insensitive	false
Labels	
Tags	

dataset1 created. [GO TO DATASET](#) X

PERSONAL HISTORY PROJECT HISTORY [REFRESH](#)

BigQuery

Create table

Source

Create table from

Empty table

Google Cloud Storage

Upload

Drive

Google Bigtable

Amazon S3

Azure Blob Storage

Project *

eece6893-398401

BROWSE

Dataset *

dataset1

Table *

Unicode letters, marks, numbers, connectors, dashes or spaces allowed.

Table type

Native table

Schema

CREATE TABLE

CANCEL

Create table

Source

Create table from

Google Cloud Storage

Select file from GCS bucket or [use a URI pattern](#)

File format

Avro

Source Data Partitioning

Destination

Project *

eece6893-398401

Dataset *

dataset1

Table *

Unicode letters, marks, numbers, connectors, dashes or spaces allowed.

Table type

Native table

Schema

CREATE TABLE

CANCEL

Choose a file

6893_1a

data_citibike_stations.csv

Filename

SELECT

CANCEL

BigQuery

Create table



Destination

Project *
eecs6893-398401 BROWSE

Dataset *
dataset1

Table *
bike_table

Unicode letters, marks, numbers, connectors, dashes or spaces allowed.

Table type
Native table ▼ ?

Schema

Auto detect

i Schema will be automatically generated.

Partition and cluster settings

Partitioning
No partitioning ▼ ?

Clustering order ?

Clustering order determines the sort order of the data. Clustering can be used on both partitioned and non-partitioned tables.

CREATE TABLE

CANCEL

BigQuery

Free Trial and Free Tier | Google Cloud | SQL workspace - BigQuery - | | big-data-6893 - Bucket details | +

consolidated.cloud.google.com/bigquery?referrer=search&cloudshell=false&project=big-data-6893-325519&ws=1m514m11m31sbig-data-6893-325519|2sbquj

Free trial status: \$300.00 credit and 91 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

Google Cloud Platform | big data 6893 | bigque

FEATURES & INFO | SHORTCUT | DISABLE EDITOR TABS

Explorer | + ADD DATA | *UNSAVE... | X

Q Type to search

Viewing pinned projects.

- big-data-6893-325519
 - dataset1
 - bike_data

MORE RESULTS

```
1 SELECT * FROM `big-data-6893-325519.dataset1.bike_data`
2 WHERE region_id=70
3 LIMIT 5
```

Query results | SAVE RESULTS | EXPLORE DATA

Query complete (0.3 sec elapsed, 108.5 KB processed)

Job information | Results | JSON | Execution details

Row	station_id	name	short_name	latitude	longitude	region_id	rental_methods	capacity	eightd_has_key_dispenser	num_bikes_availab
1	3206	Hilltop	JC019	40.7311689	-74.0575736	70	KEY,CREDITCARD	26	false	
2	3195	Sip Ave	JC056	40.73089709786179	-74.06391263008118	70	KEY,CREDITCARD	34	false	
3	3640	Journal Square	JC103	40.73367	-74.0625	70	KEY,CREDITCARD	18	false	
4	3481	York St	JC096	40.71649	-74.04105	70	KEY,CREDITCARD	22	false	

JOB HISTORY | QUERY HISTORY | SAVED QUERIES

Free trial status: \$300.00 credit and 91 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

Google Cloud | big data 6893 | Search bigque

Explorer | + ADD DATA | | bike_data | X | Editor 2 | X | +

Q Type to search

Viewing pinned projects.

- big-data-6893-362015
 - dataset1
 - bike_data

MORE RESULTS

SCHEMA | DETAILS | PREVIEW

Filter Enter property name or value

Field name	Type	Mode	Collation	Default Value	Policy Tags	Description
<input type="checkbox"/> station_id	INTEGER	NULLABLE				
<input type="checkbox"/> name	STRING	NULLABLE				
<input type="checkbox"/> short_name	STRING	NULLABLE				
<input type="checkbox"/> latitude	FLOAT	NULLABLE				
<input type="checkbox"/> longitude	FLOAT	NULLABLE				
<input type="checkbox"/> region_id	INTEGER	NULLABLE				
<input type="checkbox"/> rental_methods	STRING	NULLABLE				
<input type="checkbox"/> capacity	INTEGER	NULLABLE				
<input type="checkbox"/> eightd_has_key_dispenser	BOOLEAN	NULLABLE				
<input type="checkbox"/> num_bikes_available	INTEGER	NULLABLE				
<input type="checkbox"/> num_bikes_disabled	INTEGER	NULLABLE				
<input type="checkbox"/> num_docks_available	INTEGER	NULLABLE				
<input type="checkbox"/> num_docks_disabled	INTEGER	NULLABLE				
<input type="checkbox"/> is_installed	BOOLEAN	NULLABLE				
<input type="checkbox"/> is_renting	BOOLEAN	NULLABLE				
<input type="checkbox"/> is_returning	BOOLEAN	NULLABLE				
<input type="checkbox"/> eightd_has_available_keys	BOOLEAN	NULLABLE				
<input type="checkbox"/> last_reported	TIMESTAMP	NULLABLE				

EDIT SCHEMA | VIEW ROW ACCESS POLICIES

PERSONAL HISTORY | PROJECT HISTORY | REFRESH

'bike_data' created. GO TO TABLE X



Dataproc

Dataproc

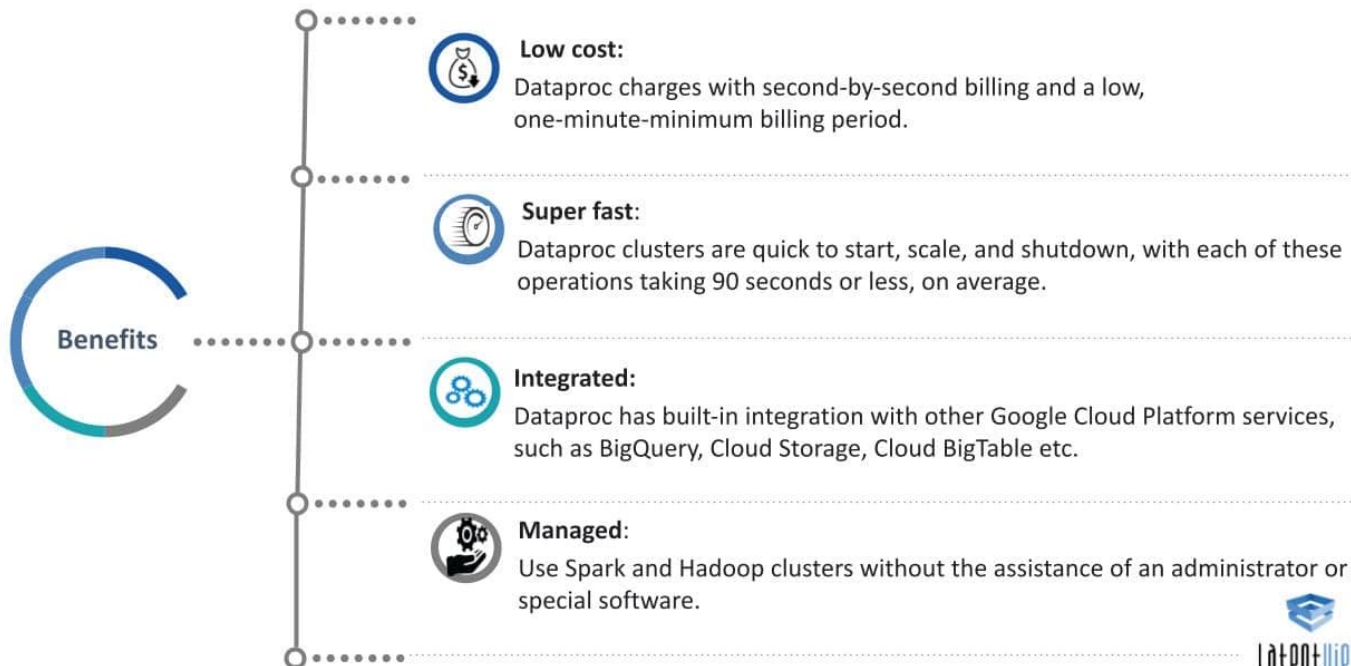
What is dataproc?

- Google Cloud Dataproc is a managed service for running **Apache Hadoop and Spark jobs**.
- Dataproc uses **Compute Engine instances** under the hood, but it takes care of the management details.
- Includes **Hadoop, Spark, Hive and Pig**.
- **Ideal for moving** existing code to GCP



Dataproc

Why dataproc?



Dataproc

search cloud data proc
click on the API link

Free trial status: \$300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform. DISMISS ACTIVATE

Google Cloud EECS6893

Product details

Cloud Dataproc API

[Google Enterprise API](#)

Manages Hadoop-based clusters and jobs on Google Cloud Platform.

ENABLE [TRY THIS API](#)

[OVERVIEW](#) [DOCUMENTATION](#) [RELATED PRODUCTS](#)

Overview

Manages Hadoop-based clusters and jobs on Google Cloud Platform.

Additional details

Type: [SaaS & APIs](#)
Last product update: 7/21/22
Category: [Google Enterprise APIs](#)
Service name: dataproc.googleapis.com

Tutorials and documentation

Dataproc - graphical UI

Free trial status: \$300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

Google Cloud EECS6893 dataproc Search

Dataproc Clusters CREATE CLUSTER REFRESH START STOP DELETE REGIONS + 5 RECOMMENDED ALERTS

Jobs on Clusters

- Clusters
- Jobs
- Workflows
- Autoscaling policies

Serverless

- Batches
- Interactive

Metastore Services

- Metastore
- Federation

Utilities

Release Notes

Cluster
Cloud Dataproc
Google Cloud Dataproc lets you provision Apache Hadoop clusters and connect to underlying analytic data stores.
There are no clusters in the currently selected Cloud Dataproc region(s). Create a cluster to get started.
CREATE CLUSTER

Go to Cloud Dataproc

Create Dataproc cluster

Select the infrastructure service that you want to use.

Cluster on Compute Engine

Create the cluster on Compute Engine.

CREATE

Cluster on GKE

Create the cluster on Google Kubernetes Engine (GKE).

CREATE

CANCEL

cloud dataproc

Create a Dataproc cluster on Compute Engine

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.

Name

Cluster Name *
cluster-6893

Location

Region *
us-east1

Zone *
us-east1-b

Cluster type

Standard (1 master, N workers)

Single Node (1 master, 0 workers)
Provides one node that acts as both master and worker. Good for proof-of-concept or small-scale processing

High Availability (3 masters, N workers)
Hadoop High Availability mode provides uninterrupted YARN and HDFS operations despite single-node failures or reboots

Autoscaling

Automates cluster resource management based on an autoscaling policy.

Policy
None

Enhanced Flexibility Mode

Autoscaling policies

- Serverless
- Batches
- Metastore Services
- Metastore
- Federation
- Utilities
- Component exchange
- Workbench

Release Notes

Customize cluster (optional)
Add cluster properties, features, and actions.

Manage security (optional)
Change access, encryption, and security settings.

EQUIVALENT COMMAND LINE

CREATE **CANCEL**

create cluster with Jupyter

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
- Presto
- ZooKeeper
- Ranger
- HBase
- Flink
- Docker
- Solr

Dataproc - Cloud SDK

Cluster creation (using Cloud SDK): (Instead of using GUI, command line tool can also be used to create Dataproc, recommended for Linux experts)

```
(base) conghan@Cong's-MacBook-Pro:~$ gcloud dataproc clusters create example-cluster --region=us-east1
```

Dataproc - Cloud SDK

Cluster creation (using Cloud SDK):

```
(base) conghan@Cong's-MacBook-Pro:~$ gcloud dataproc clusters create example-cluster --region=us-east1
Waiting on operation [projects/big-data-6893-325519/regions/us-east1/operations/e3efb89c-f2ad-35e2-9a91-b62392477950].
Waiting for cluster creation operation...
WARNING: No image specified. Using the default image version. It is recommended to select a specific image version in production, as the default image version may change at any time.
Waiting for cluster creation operation...done.
Created [https://dataproc.googleapis.com/v1/projects/big-data-6893-325519/regions/us-east1/clusters/example-cluster] Cluster placed in zone [us-east1-c].
(base) conghan@Cong's-MacBook-Pro:~$
```

Dataproc - Cloud SDK

Submit a job - Pi calculation

```
(base) conghan@Cong's-MacBook-Pro:~$ gcloud dataproc jobs submit spark --cluster
example-cluster \
> --region=us-east1 \
> --class org.apache.spark.examples.SparkPi \
> --jars file:///usr/lib/spark/examples/jars/spark-examples.jar -- 1000
```

Dataproc - Cloud SDK

Submit a job - Pi calculation

```
urceManager at example-cluster-m/10.142.0.3:8032
21/09/10 01:32:11 INFO org.apache.hadoop.yarn.client.AHSPProxy: Connecting to App
lication History server at example-cluster-m/10.142.0.3:10200
21/09/10 01:32:12 INFO org.apache.hadoop.conf.Configuration: resource-types.xml
not found
21/09/10 01:32:12 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unabl
e to find 'resource-types.xml'.
21/09/10 01:32:13 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Su
bmitted application application_1631237290616_0001
21/09/10 01:32:14 INFO org.apache.hadoop.yarn.client.RMPProxy: Connecting to Reso
urceManager at example-cluster-m/10.142.0.3:8030
21/09/10 01:32:16 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.h
adoop.gcsio.GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonRespons
exception; verified object already exists with desired state.
Pi is roughly 3.1416210314162103
21/09/10 01:32:33 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped
spark-repl-10782abe{m/17/1.1, (tcp://1.1)}{0.0.0.0}
Job [3f9861f7e3744a5580068001cdf48bf9] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-east1-881004012112-ixdi0md0/goog
le-cloud-dataproc-metainfo/7ff01079-3cec-47b3-b2f4-ba88665d16e1/jobs/3f9861f7e37
44a5580068001cdf48bf9/
driverOutputResourceUri: gs://dataproc-staging-us-east1-881004012112-ixdi0md0/go
ogle-cloud-dataproc-metainfo/7ff01079-3cec-47b3-b2f4-ba88665d16e1/jobs/3f9861f7e
3744a5580068001cdf48bf9/driveroutput
jobUuid: e5839c28-799f-3591-8dd8-eba4f198110e
```


Dataproc

- On-demand, fully managed cloud service for running Apache Hadoop and Spark on GCP
- Cluster creation (using Cloud SDK):
 - Automatically creates VMs with Spark pre-installed

Install
Jupyter
Notebook

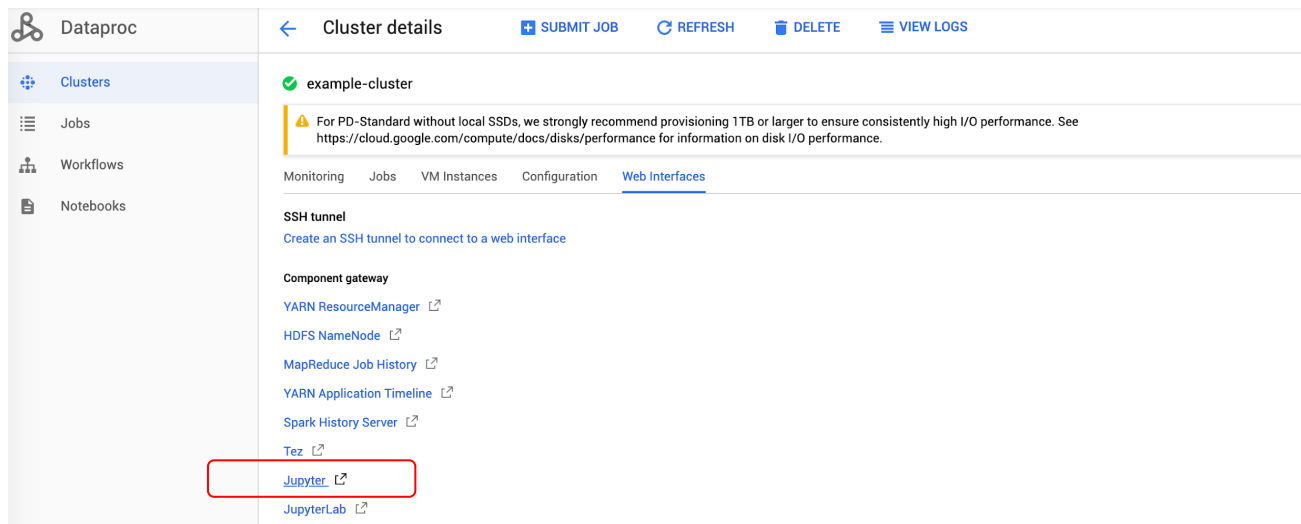
```
(base) conghan@Cong's-MacBook-Pro:~$ gcloud beta dataproc clusters create example-cluster --region=us-east1 --optional-components=ANACONDA,JUPYTER --image-version=1.3 --enable-component-gateway --bucket big-data-6893 --project big-data-6893-325519 --single-node --metadata 'PIP_PACKAGES=graphframes==0.6' --initialization-actions gs://dataproc-initialization-actions/python/pip-install.sh
```

Cloud Storage bucket: where your jupyter notebooks are saved

Works like `pip install <your package>`

Dataproc - Spark execution / submit jobs

- Jupyter notebook:



The screenshot shows the Dataproc console interface. On the left is a navigation sidebar with 'Clusters' selected. The main area displays 'Cluster details' for 'example-cluster'. A warning message is shown at the top. Below it, the 'Web Interfaces' tab is active, listing several services: SSH tunnel, Component gateway, YARN ResourceManager, HDFS NameNode, MapReduce Job History, YARN Application Timeline, Spark History Server, Tez, **Jupyter** (highlighted with a red box), and JupyterLab.

- Cloud SDK:

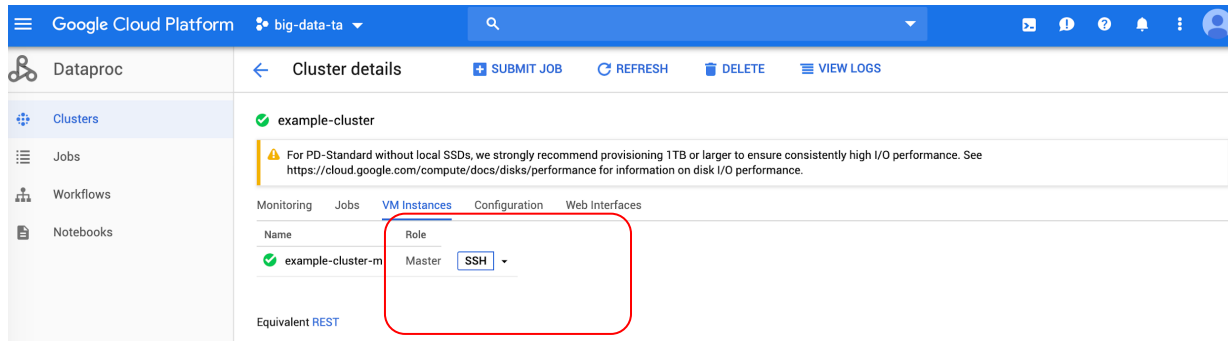
- `gcloud dataproc jobs submit pyspark <your_program.py> -- cluster=<cluster-name>`

- [View your jobs in console](#)

- Program could be Cloud Storage URI / local path / Cloud Shell path
- Data should be on Cloud storage

Dataproc - Spark execution / submit jobs (cont')

- Spark shell
 - ssh into master node



Google Cloud Platform big-data-ta

Dataproc Cluster details

example-cluster

For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See <https://cloud.google.com/compute/docs/disks/performance> for information on disk I/O performance.

Monitoring Jobs VM Instances Configuration Web Interfaces

Name	Role
example-cluster-m	Master SSH

Equivalent REST

- pyspark

```
frouyang@example-cluster-m:~$ pyspark
Python 2.7.14 [Anaconda, Inc.] (default, Dec 7 2017, 17:05:42)
[GCC 7.2.0] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
19/09/06 18:46:51 WARN org.apache.spark.scheduler.FairSchedulableBuilder: Fair Scheduler configuration file not found so jobs will be scheduled in FIFO order. To use fair scheduling, configure pools in fairscheduler.xml or set spark.scheduler.allocation.file to a file that contains the configuration.
Welcome to

  SPARK  version 2.3.3

Using Python version 2.7.14 (default, Dec 7 2017 17:05:42)
SparkSession available as 'spark'.
>>> |
```

HW0

1. Read documentations and tutorials
 - a. Setup GCP and Cloud SDK
 - b. Familiar with BigQuery
 - c. Run Spark examples on Dataproc - Pi calculation and word count
2. Two light programming questions
 - a. BigQuery
 - b. Spark program - Find top k most frequent words

Remember to delete your dataproc clusters when you finish executions to save money.