



# EECS E6893 Big Data Analytics

## Intro to Big Data Analytics on GCP

Apurva Patel, [amp2365@columbia.edu](mailto:amp2365@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



apurvagcp2@gmail.com

[SWITCH ACCOUNT](#)

Country

United States

By using this application, you agree to the [Google Cloud Platform](#), [Supplemental Free Trial](#), and [any applicable services and APIs](#) Terms of Service.

[AGREE & 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 activate your full 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

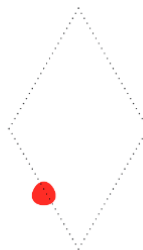
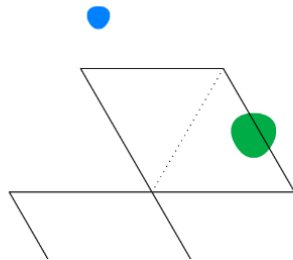
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

## \$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 protect your account. If you use a credit or debit card, you won't be charged until you manually activate your full account.

Payments profile

[Create new payments profile](#)

Your payment information is saved in a payments profile linked to your Google Account and shared across Google services. [Learn more about payments profile](#)

Payment method

[Add payment method](#)

Please complete the previous sections before continuing.

START FREE

[Privacy policy](#) | [FAQs](#)

### Create a payments profile

Only **Organization** profiles can have multiple users. If you select an **Individual** profile, you agree that use of your profile is for your trade, business, craft, or profession. In some countries, this selection affects your tax options. Your profile type can't be changed after signing up. [Learn more about payments profile](#)

Profile type

Individual

Legal name

Apurva Patel

Street address

Apt, suite, etc. (optional)

City

State

Zip code

[Cancel](#)

[Create](#)

### Google Cloud products

You need to build and run your apps, services, and other Google Cloud products, including Firebase and the

### or free

to work with \$300 in credit to get started. It expires at 90 days.

### after free trial ends

Use your credit card to make sure you are always charged. If you use a credit or debit card, you won't be charged until you manually activate your full account.

## 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 full account.

Payments profile

Apurva Patel

Individual • United States • ID: 9412-3345-6487



Your payment information is saved in a payments profile, which is associated with your Google Account and shared across Google services. [Learn more about payments profile](#)

Payment method

[Add payment method](#)



START FREE

## Access to all Google Cloud products

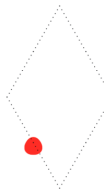
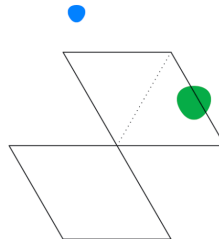
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

## \$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 activate your full account.




Google Cloud

My First Project

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

Search

1

Welcome

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

Number: 633183201907 ID: ome


[Add people to your project](#)

[Set up budget alerts](#)

[Review product spend](#)


### Recommended based


Pre-built solution templates


Deploy a three-tier web app  
Web app, rich media site, e-commerce database-backed website


[View all Solutions](#)

### Products

Create a VM  
Compute Engine

Train and host ML models  
Vertex AI

Create a database  
Cloud SQL

Analyze and manage data  
BigQuery

Google Cloud Platform

## Welcome!

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

1

What best describes your organization or needs?

Please select \*

Other

Please specify (optional)

Student

NEXT

2

What brought you to Google Cloud?

3

What are you interested in doing with Google Cloud?

4

What best describes your role?

CLOSE


DONE

11

← → ↻ 🏠 console.cloud.google.com/welcome/new?project=omega-iterator-434817-t2&authuser=6 ☆ 📁 📌 📄 | 👤 New Chrome available ⋮

🌐 📁 GH 📁 📁 Columbia 📁 HDGC 📁 📁 Courses 📁 Quant 📁 Zoom 📁 GitHub - ML 📁 Preparation Hub [...] 📁 Repositories 📁 Exploring with Spe... 📁 Big Data Analytics 📁 https://contactout.... | 📁 All Bookmarks

☰ Google Cloud 👤 My First Project 🔍 Search (/) for resources, docs, products, and more 🔍 Search ✨ 📄 1 ? ⋮ 👤



# Welcome

You're working on project [My First Project](#)  
Number: 633183201907 ID: omega-iterator-434817-t2

[Add people to your project](#)  
[Set up budget alerts](#)  
[Review product spend](#)

## Recommended based on your interests


### Pre-built solution templates

▶ Deploy a three-tier web application  
Web app, rich media site, ecommerce site, database-backed website

📄 [View all Solutions](#)

### Products

📄 Create a VM 📄 Train and host ML models 📄 Create a database 📄 Analyze and manage data



## Google Cloud Platform

### Welcome!

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?  
Websites Mobile apps Storage / backup **✓ Data analytics**  
**✓ Artificial intelligence / machine learning** Game development  
Containerization **✓ Data management**  
**✓ Virtual machines (VMs)** Google Maps  
**✓ Other APIs (e.g., Text-to-Speech, Speech-to-Text, Vision)**  
Google Photos or Google Workspace Other I'm not sure yet  
[NEXT](#)
- What best describes your role?  
CLOSE [DONE](#)

### Try our most advanced model: Gemini 1.5 Pro

Try Gemini →

a data warehouse with  
ry  
house, dashboards, ETL, analytics,  
ysis



← → ↺ 🏠 console.cloud.google.com/welcome/new?project=omega-iterator-434817-t2&authuser=6 ☆ 🗂️ 📌 📄 | 👤 New Chrome available ⋮

in 🌐 GH + 📁 Columbia 📁 HDFC 📁 Courses 📁 Quant 📁 Zoom 📁 GitHub - ML 📁 Preparation Hub [...] 📁 Repositories 📁 Exploring with Spe... 📁 Big Data Analytics 📁 https://contactout.... | 📁 All Bookmarks

☰ Google Cloud 👤 My First Project 🔍 Search (/) for resources, docs, products, and more 🔍 Search

🌐 Welcome

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

Number: 633183201907 📄 ID: omega-iterator-434817-t2

[Add people to your project](#)

[Set up budget alerts](#)

[Review product spend](#)

Try our most advanced model: Gemini 1.5 Pro

Try Gemini →

**Recommended based on your use case**

Pre-built solution templates

Deploy a three-tier web application

Web app, rich media site, e-commerce site, database-backed website

[View all Solutions](#)

**Products**

Create a VM

Train and host ML models

Create a database

Analyze and manage data

Google Cloud Platform

**Welcome!**

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

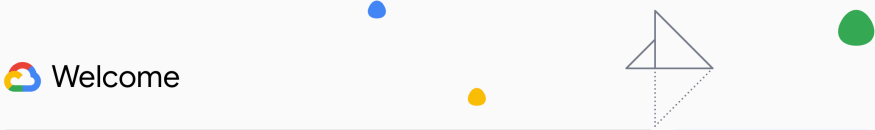
 Google Cloud


My First Project

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

Search





 **Welcome**

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

Number: 633183201907 ID: omega-iterator-43481742

[Add people to your project](#)

[Set up budget alerts](#)


[Review product spend](#)


Try our most advanced model: Gemini 1.5 Pro


Try Gemini →

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

### Pre-built solution templates


 Summarize large documents using Generative AI  
Generative AI, summarization, machine learning


 Create a data warehouse with BigQuery  
Data warehouse, dashboards, ETL, analytics, data analysis


 Create an analytics lakehouse  
Data science, IoT, streaming analytics


 [View all Solutions](#)

### Products

 Train and host ML models  
Vertex AI

 View and use notebook data  
Vertex AI Workbench

 Use foundation models  
Vertex AI Studio

 Analyze and manage data  
BigQuery



Google Cloud

My First Project

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

Search

◆

📄

1

?

⋮

👤

Billing

Overview

FREE TRIAL ACCOUNT

MANAGE BILLING ACCOUNT

LEARN

Billing account

My Billing Account

Overview

Cost management

Reports

Cost table

Cost breakdown

Budgets & alerts

Billing export

Cost optimization

FinOps hub

Committed use discounts...

CUD analysis

Pricing

Cost estimation

Credits

Payments

Documents

Transactions

Release Notes

<|

BILLING ACCOUNT OVERVIEW

PAYMENT OVERVIEW

→ View details on Reports

Top projects

September 1, 2023 – September 30, 2024

\$1

\$0.50

\$0

→ View report

Top services

September 1, 2023 – September 30, 2024

\$1

\$0.50

\$0

→ View report

→ View all health checks

Free trial credit

\$300

Free trial credit

Out of \$300

91

Days remaining

Ends December 6, 2024

You are not billed during your Free Trial. When the Free Trial ends, all resources you created during the trial are stopped and you will not be charged, unless you upgrade to a paid Cloud Billing account.

Please note: The Free Trial period cannot be paused or extended. The Free Trial ends after all available credits are consumed, or at the end of the Free Trial period, whichever happens first.

ACTIVATE

LEARN MORE

Google Cloud Platform

Search products and resources

1

Billing

Overview

My Billing Account

Overview

Reports

Cost table

Cost breakdown

Commitments

Commitment analysis

Budgets & alerts

Billing export

Pricing

Documents

Transactions

Payment settings

Payment method

Account management

Release Notes

Overview

BILLING ACCOUNT OVERVIEW

PAYMENT OVERVIEW

view report

Cost trend

September 1, 2020 – September 30, 2021

Average monthly total cost

\$0.00

Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep

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

1

1

View all health checks

Free trial credit

\$300

Free trial credit

Out of \$300

91

Days remaining

Ends December 9, 2021

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

Google Cloud

My First Project

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

Search

1

Billing

Overview

FREE TRIAL ACCOUNT

MANAGE BILLING ACCOUNT

LEARN

Billing account

My Billing Account

BILLING ACCOUNT OVERVIEW

PAYMENT OVERVIEW

View details on Reports

Top projects

September 1, 2023 – September

\$1

\$0.50

\$0

View report

Top services

September 1, 2023 – September

\$1

\$0.50

\$0

View report

Cost management

Reports

Cost table

Cost breakdown

Budgets & alerts

Billing export

Cost optimization

FinOps hub

Committed use discounts...

CUD analysis

Pricing

Cost estimation

Credits

Payments

Documents

Transactions

Release Notes

Select a project

NEW PROJECT

Search projects and folders

RECENT

STARRED

ALL

Name	ID
✓ ☆ ☰ My First Project ?	omega-iterator-434817-12

CANCEL

View all health checks

Free trial credit

\$300

Free trial credit

Out of \$300

91

Days remaining

Ends December 6, 2024

are not billed during your Free Trial. When the Free Trial ends, all resources you ed during the trial are stopped and you will not be charged, unless you upgrade said Cloud Billing account.

se note: The Free Trial period cannot be paused or extended. The Free Trial after all available credits are consumed, or at the end of the Free Trial period, ever happens first.

IVATE

LEARN MORE

## New Project



You have 11 projects remaining in your quota. Request an increase or delete projects. [Learn more](#)

[MANAGE QUOTAS](#)

Project name \*

EECS6893



Project ID: eecs6893-434817. It cannot be changed later. [EDIT](#)

Location \*

No organization

[BROWSE](#)

Parent organization or folder

CREATE

CANCEL



Google Cloud

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

Search

LEARN

Tutorial

Billing

My Billing Account

Overview

Cost management

Reports

Cost table

Cost breakdown

Budgets & alerts

Billing export

Cost optimization

FinOps hub

Committed use discounts...

CUD analysis

Pricing

Cost estimation

Credits

Payments

Documents

Transactions

Release Notes

Overview

Free trial account

Billing account overview

Payment overview

\$0.00

Costs take a few hours to show up, and might take longer than 24 hours. [Learn more](#) or [get an alert](#) when usage cost is available.

View details on Reports

Top projects

September 1, 2023 – September 30, 2024

View report

Notifications

Create Project: EEC56893

SELECT PROJECT

Create Project: My First Project

SELECT PROJECT

SEE ALL ACTIVITIES

Billing health checks

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

011

View all health checks

Free trial credit

\$300

Out of \$300

91

Days remaining

Ends December 6, 2024

You are not billed during your Free Trial. When the Free Trial ends, all resources you created during the trial are stopped and you will not be charged, unless you upgrade to a paid Cloud Billing account.

Please note: The Free Trial period cannot be paused or extended. The Free Trial

Recommended for you

Cloud Billing overview

Help document

Understand how resources, billing accounts, and projects work together.

Google Cloud Billing Tour

Tutorial10 min

Introduces the billing section of the console and some of the reports available to you.

Billing Reports Tutorial

Tutorial30 min

Familiarize yourself with billing reports and learn how to answer cost management questions.

Analyze Cloud Billing data with BigQuery

Tutorial20 min

Learn how to export Cloud Billing data to BigQuery and query it.

View your billing reports and cost trends

Help document

Learn how to use billing reports to gain visibility into your costs.

Overview of Cloud Billing access control

Help document

Learn about permissions and access management for Cloud Billing resources.

Make changes to your Cloud Billing account

21

# 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



English ▾

Console



Contact Us

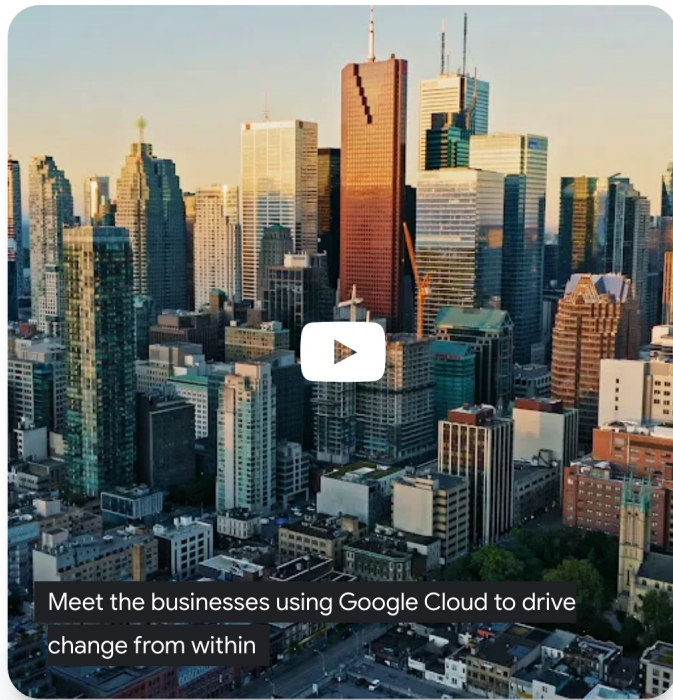
Get \$300 in free credits and free usage of 20+ products →

## Dream, build, and transform with Google Cloud

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

Go to console

Contact sales



Meet the businesses using Google Cloud to drive change from within

Google Cloud

EECS6893

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

Search

Cloud overview

Solutions

Dashboard

Recommendations

PINNED PRODUCTS

API

APIs & Services

Billing

IAM & Admin

Marketplace

Vertex AI

Compute Engine

Kubernetes Engine

Cloud Storage

BigQuery

VPC Network

Cloud Run

SQL

Security

Google Maps Platfor...

VIEW ALL PRODUCTS

Welcome

You're in Free Trial

0 out of \$300 credits used

Expires December 6, 2024

What happens when trial ends?

ACTIVATE FULL ACCOUNT

You're working on project EECS6893

Number: 927384071737 ID: eecs6893-434817

Add people to your project

Set up budget alerts

Review product spend

Try our most advanced model: Gemini 1.5 Pro

Try Gemini

Recommended based on your interest in Data, AI/ML

Pre-built solution templates

Summarize large documents using Generative AI

Generative AI, summarization, machine learning

Create a data warehouse with BigQuery

Data warehouse, dashboards, ETL, analytics, data analysis

Create an analytics lakehouse

Data science, IOT, streaming analytics

View all Solutions

Products

Train and test ML

Monitor and use network

Use foundation models

Analyze and manage

24

Google Cloud

EECS6893

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

Search

Cloud overview

Solutions

PINNED PRODUCTS

APIs & Services

Billing

IAM & Admin

Marketplace

Vertex AI

Compute Engine

Kubernetes Engine

Cloud Storage

BigQuery

VPC Network

Cloud Run

SQL

Security

Google Maps Platfor...

VIEW ALL PRODUCTS

DASHBOARD

ACTIVITY

RECOMMENDATIONS

CUSTOMIZE

Project info

Project name  
EECS6893

Project number  
927384071737

Project ID  
eecs6893-434817

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 Run functions  
Event-driven serverless functions

Cloud Run  
Serverless for containerized applications

Getting Started

APIs

Requests (requests/sec)

No data is available for the selected time frame.

Go to APIs overview

Google Cloud Platform status

All services normal

Go to Cloud status dashboard

Billing

Estimated charges  
For the billing period Sep 1 – 6, 2024

USD \$0.00

Take a tour of billing

View detailed charges

Monitoring

Create my dashboard

Set up alerting policies

Create uptime checks

View all dashboards

Go to Monitoring

API Error Reporting

25

# 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/  
& $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\bundledpython\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;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\WPICHI2\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\bundledpython\python39.zip;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\platform\bundledpython\DLLs;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\platform\bundledpython\lib;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\platform\bundledpython]
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
* jackyyu2021111@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	kubectrl	kubectrl	< 1 MiB
Not Installed	kubectrl-oidc	kubectrl-oidc	20.5 MiB

# GCP: Cloud Shell

The screenshot shows the Google Cloud Platform dashboard. At the top, there's a header with the Google Cloud logo, a search bar, and a notification about a free trial status. Below the header, there are tabs for DASHBOARD, ACTIVITY, and RECOMMENDATIONS. The main content area is divided into three columns. The left column shows 'Project info' for project 'EECS6893', including its name, number, and ID. The middle column shows 'API APIs' with a table of requests per second, but it indicates 'No data is available for the selected time frame.' The right column shows 'Google Cloud Platform status', which includes a message about an issue with Vertex AI Search and a link to the Cloud status dashboard. A red box highlights a terminal icon in the top right corner of the dashboard.

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

CUSTOMIZE

**Project info**

- Project name: EECS6893
- Project number: 755979763552
- Project ID: eeecs6893-398401

[ADD PEOPLE TO THIS PROJECT](#)

**API APIs**

Requests (requests/sec)

1.0
0.8
0.6
0.4
0.2

No data is available for the selected time frame.

**Google Cloud Platform status**

**Multiple Products**

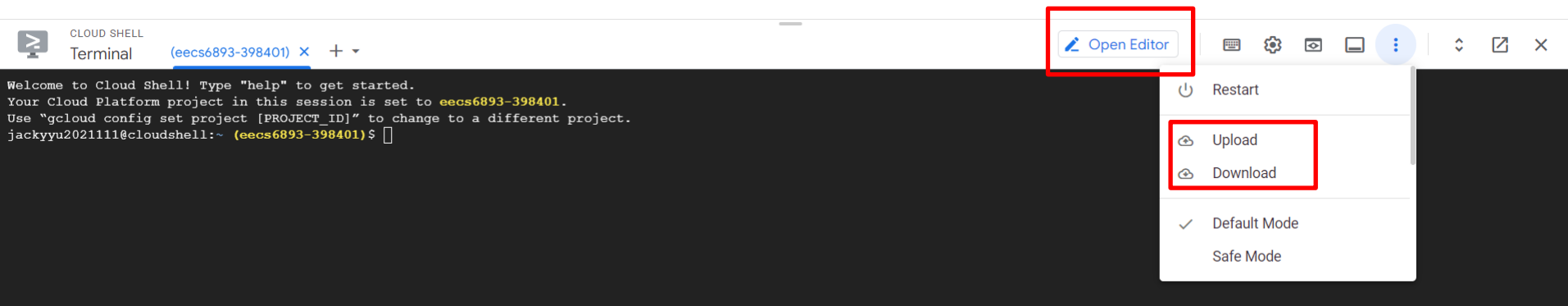
We are investigating an Issue with Vertex AI Search  
Began at 2023-09-07 (08:14:54)

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

[Go to Cloud status dashboard](#)

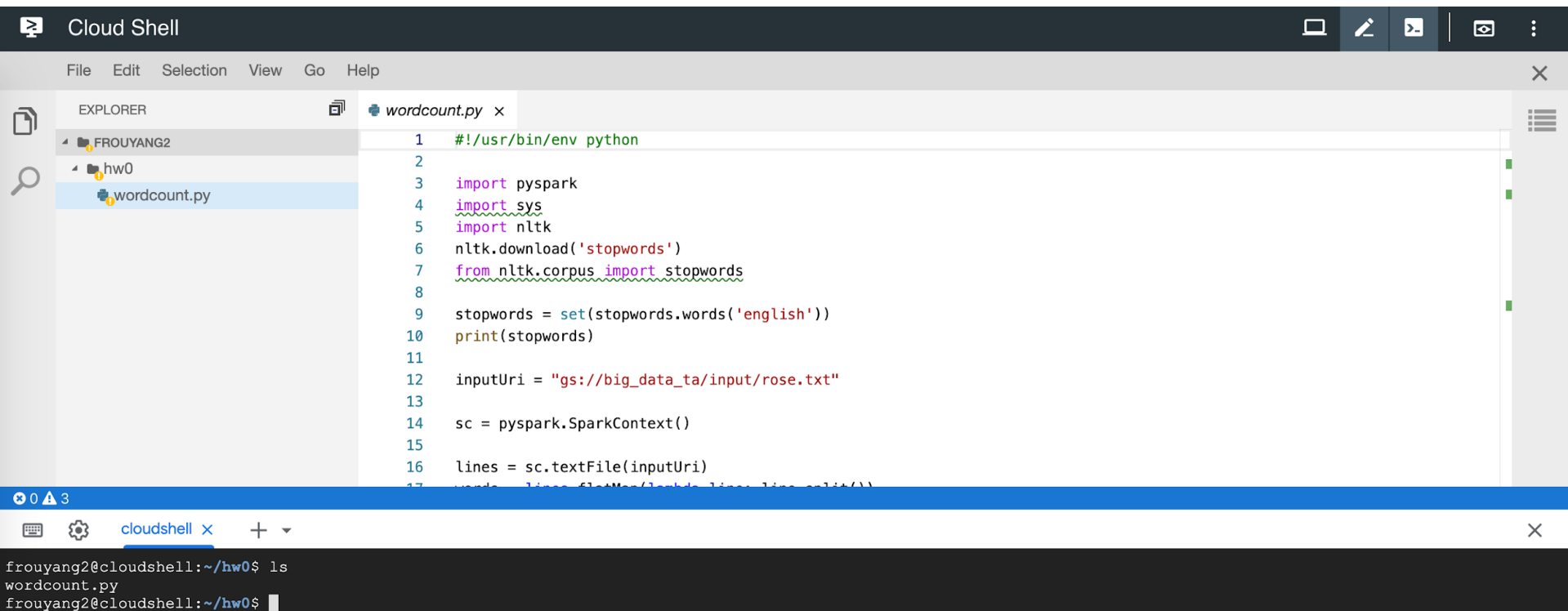
persistent home directory :). The most useful way to complete the HW0

# GCP: Cloud Shell



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

# GCP: Cloud Shell Code Editor



Cloud Shell

File Edit Selection View Go Help

EXPLORER

wordcount.py x

```
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())
```

cloudshell x

```
frouyang2@cloudshell:~/hw0$ ls
wordcount.py
frouyang2@cloudshell:~/hw0$
```



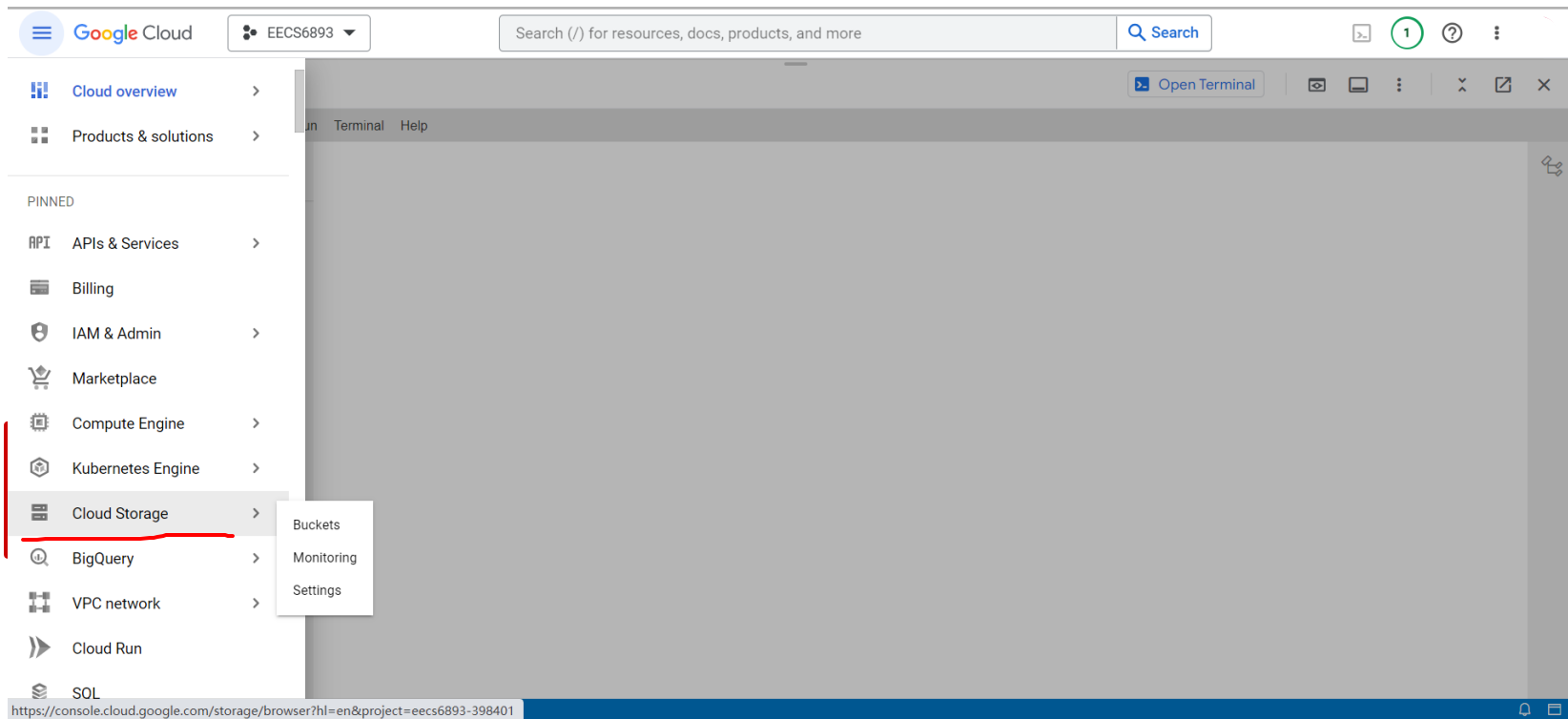
# Cloud Storage

# 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 autoclass 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
requester Pays 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





# Cloud Storage

Google Cloud

EECS6893

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									


Marketplace

Release Notes

<1


# 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.

 Google Cloud


EECS6893

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

 Search


DISMISS


ACTIVATE


 Cloud Storage


Create a bucket


HELP ASSISTANT

 Buckets

 Monitoring

 Settings

 Marketplace

 Release Notes

<1

- Name your bucket**

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

6893\_tq

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

Name your own bucket

# 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



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



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

?

:

Cloud Storage

Buckets

Monitoring

Settings

Create a bucket

us-east1 (South Carolina)

\$0.020 per GB-month

HELP ASSISTANT

Default storage class: Standard

ESTIMATE YOUR MONTHLY COST

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

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 1 ? :

Cloud Storage ← **Bucket details** REFRESH HELP ASSISTANT LEARN

**6893\_ta**

**Location** **Storage class** **Public access** **Protection**  
us-east1 (South Carolina) Standard Not public 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

<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	
<input type="checkbox"/>	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.

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

Cloud Storage Object details

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	<a href="https://storage.cloud.google.com/6893_ta/data_citibike_stations.csv">https://storage.cloud.google.com/6893_ta/data_citibike_stations.csv</a>
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

 EECS6893



[←](#) Product details



## BigQuery API

[Google Enterprise API](#)

A data platform for customers to create, manage, share and query data.

ENABLE

TRY THIS API [↗](#)

OVERVIEW

DOCUMENTATION

RELATED PRODUCTS

### Overview

A data platform for customers to create, manage, share and query data.

### Additional details

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

### Tutorials and documentation

[Learn more](#) [↗](#)

### Terms of Service

By using this product you agree to the terms and conditions of the following license: [Google Cloud Platform](#).

### Related Products

# BigQuery

The screenshot shows the Google Cloud BigQuery Studio interface. At the top, there's a Google Cloud logo and a search bar. The left sidebar contains an 'Explorer' panel with a search bar and a list of resources. The main area displays a 'Welcome to BigQuery Studio!' message with options to 'Create new' (SQL QUERY, PYTHON NOTEBOOK, DATA CANVAS) and 'Try with sample data'. A modal window is open in the center with the title 'Welcome to BigQuery in the Cloud Console'. The modal contains the following text:

**Welcome to BigQuery in the Cloud Console**

**New to the BigQuery UI?**

The BigQuery UI helps you complete tasks like running queries, loading data, and even creating and training ML models. Check out the BigQuery [quickstart guide](#) to learn how to start performing data analysis on Google Cloud.

**Learn about new features**

New improvements and updates are constantly on the way. We recommend periodically checking our [release notes](#) to stay up to date on what's new.

At the bottom of the modal is a 'DONE' button. Below the modal, there's a section 'Add your own data' with a 'Local file' option and a 'LAUNCH THIS GUIDE' button. On the right, there's a 'Try the Colab Demo Notebook' section with an 'OPEN THIS NOTEBOOK' button. At the bottom right, there's a checkbox labeled 'Show welcome page on startup' which is checked. The bottom of the screen shows a URL bar with the address: <https://cloud.google.com/bigquery/docs/quickstarts/quickstart-web-u?authuser=6>.

# BigQuery

The screenshot shows the Google Cloud BigQuery console. At the top, there's a header with the Google Cloud logo, a project selector set to 'EECS6893', and a search bar. Below the header, the 'Explorer' sidebar on the left contains a search bar and a list of resources. A red rectangle highlights the 'Create dataset' button in the Explorer sidebar. The main content area displays a 'Welcome to your SQL Workspace!' message, a 'Get started' section with 'COMPOSE A NEW QUERY' and 'ADD' buttons, and a 'Try the Google Trends Demo Query' section with an 'OPEN THIS QUERY' button. At the bottom, there's a section titled 'Add your own data' with options for 'Local file', 'Google Drive', and 'Google Cloud Storage'. The bottom of the page shows tabs for 'PERSONAL HISTORY' and 'PROJECT HISTORY', along with a 'REFRESH' button.

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

Type to search

Viewing workspace resources.  
[SHOW STARRED ONLY](#)

eeecs6893-398401

☆ ⋮

**Create dataset**  
Refresh contents

[COMPOSE A NEW QUERY](#) [ADD](#)

Get started

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 SQL Workspace. At the top, a banner indicates a free trial status with \$300.00 credit and 90 days remaining. The interface includes a top navigation bar with the Google Cloud logo, the project ID 'EECS6893', and the 'BigQuery' label. On the left, an 'Explorer' sidebar shows the project 'eecs6893-398401'. The main workspace area displays a 'Welcome to your SQL Workspace!' message, a 'Get started' section with buttons for 'COMPOSE A NEW QUERY' and 'ADD', and a 'Try with sample data' section featuring a 'Try the Google Trends Demo Query' with an 'OPEN THIS QUERY' button. Below this, the 'Add your own data' section offers options to upload a 'Local file' or connect to 'Google Drive' (Google storage service). At the bottom, tabs for 'PERSONAL HISTORY' and 'PROJECT HISTORY' are visible.

## 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



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](#)

Sample data

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](#)

Own data



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

X

Search

1

?

:



Explorer

+ ADD

K

Type to search

Viewing workspace resources.

SHOW STARRED ONLY

eeecs6893-398401

dataset1

SHOW MORE

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



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 | +

console.cloud.google.com/bigquery?referrer=search&cloudshell=false&project=big-data-6893-325519&ws=1m51m41m31sbig-data-6893-32551912sbquj

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 | bigquery

FEATURES & INFO | SHORTCUT | DISABLE EDITOR TABS

Explorer + ADD DATA

Q Type to search

Viewing pinned projects.

big-data-6893-325519

dataset1

bike\_data

Q \*UNSAVE... X

RUN SAVE SCHEDULE MORE

```
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	eighthd_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 bigquery

Explorer + ADD DATA

Q Type to search

Viewing pinned projects.

big-data-6893-362015

dataset1

bike\_data

MORE RESULTS

QUERY SHARE COPY SNAPSHOT DELETE EXPORT

SCHEMA DETAILS PREVIEW

Filter Enter property name or value

Field name	Type	Mode	Collation	Default Value	Policy Tags	Description
station_id	INTEGER	NULLABLE				
name	STRING	NULLABLE				
short_name	STRING	NULLABLE				
latitude	FLOAT	NULLABLE				
longitude	FLOAT	NULLABLE				
region_id	INTEGER	NULLABLE				
rental_methods	STRING	NULLABLE				
capacity	INTEGER	NULLABLE				
eighthd_has_key_dispenser	BOOLEAN	NULLABLE				
num_bikes_available	INTEGER	NULLABLE				
num_bikes_disabled	INTEGER	NULLABLE				
num_docks_available	INTEGER	NULLABLE				
num_docks_disabled	INTEGER	NULLABLE				
is_installed	BOOLEAN	NULLABLE				
is_renting	BOOLEAN	NULLABLE				
is_returning	BOOLEAN	NULLABLE				
eighthd_has_available_keys	BOOLEAN	NULLABLE				
last_reported	TIMESTAMP	NULLABLE				

EDIT SCHEMA VIEW ROW ACCESS POLICIES

'bike\_data' created. GO TO TABLE X

PERSONAL HISTORY PROJECT HISTORY REFRESH

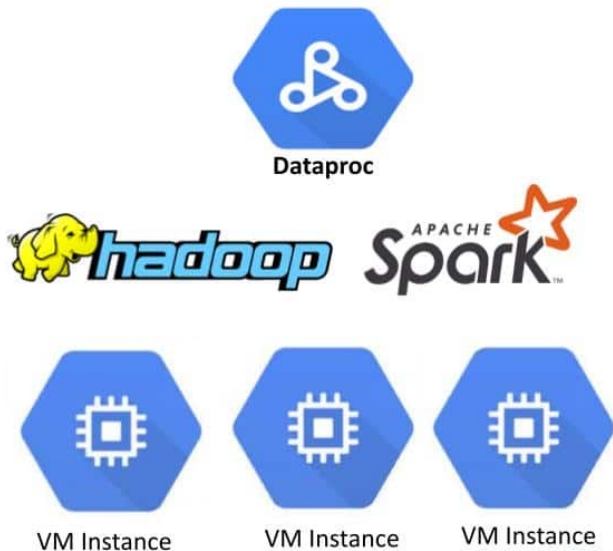


# Dataproc

# Dataproc

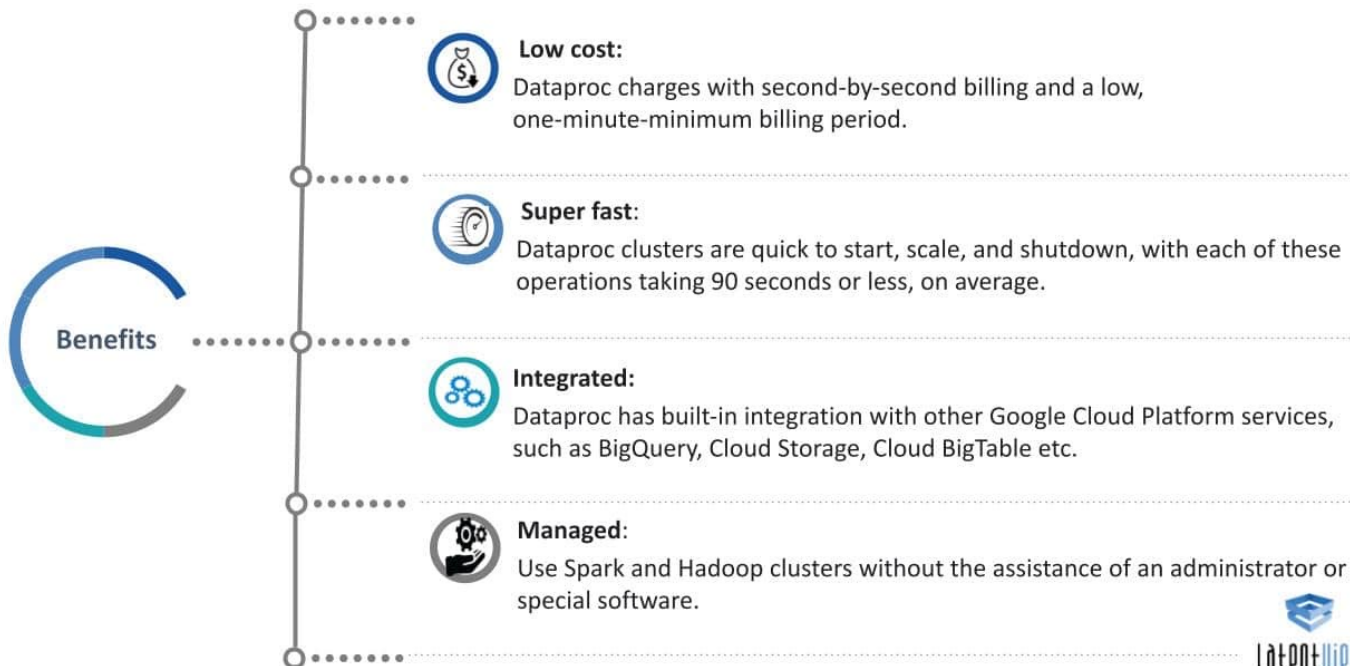
## What is dataproc?

- 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

Q 1 ? ⋮

← 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.

CREATE

CANCEL

EQUIVALENT COMMAND LINE

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.

CREATE

CANCEL

EQUIVALENT COMMAND LINE

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

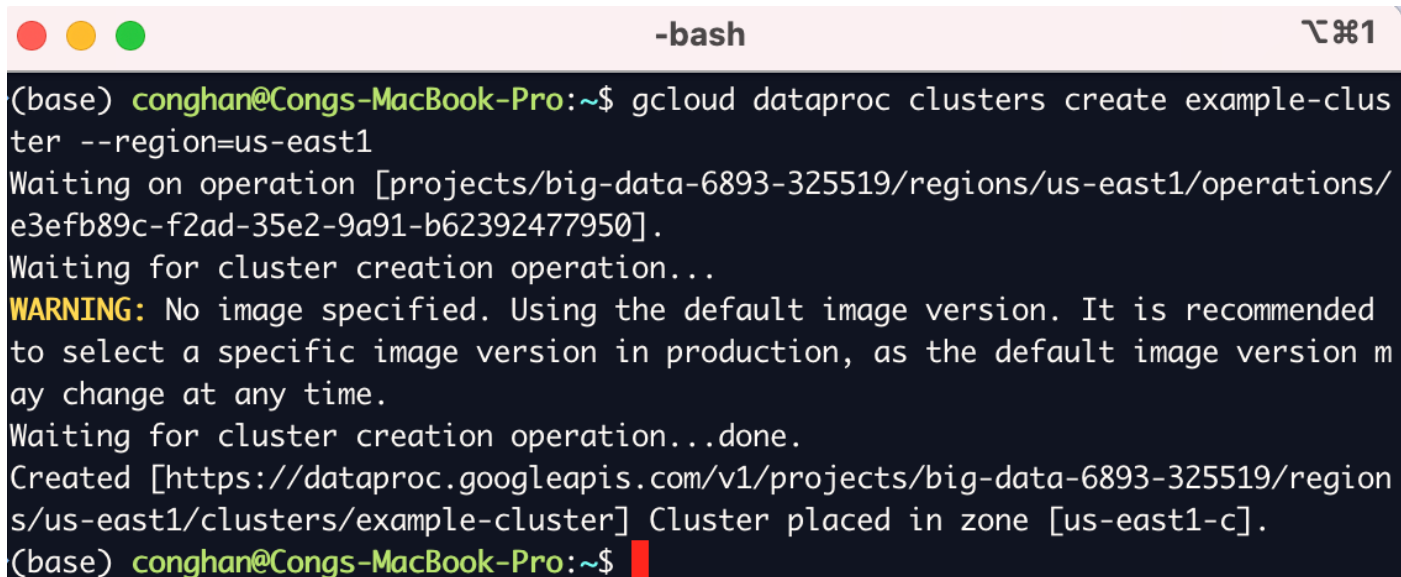
# Dataproc - Cloud SDK

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):

A terminal window with a pink title bar containing three colored window control buttons (red, yellow, green) on the left, the text '-bash' in the center, and a zoom icon followed by '#1' on the right. The terminal content shows a user running a 'gcloud dataproc clusters create' command. It displays the operation ID, a warning about the default image version, and the final cluster creation details including the URL and zone.

```
(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
eException; 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
SparkUI at example-cluster-m/10.142.0.3:8032
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-ebef198110e
```

# 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

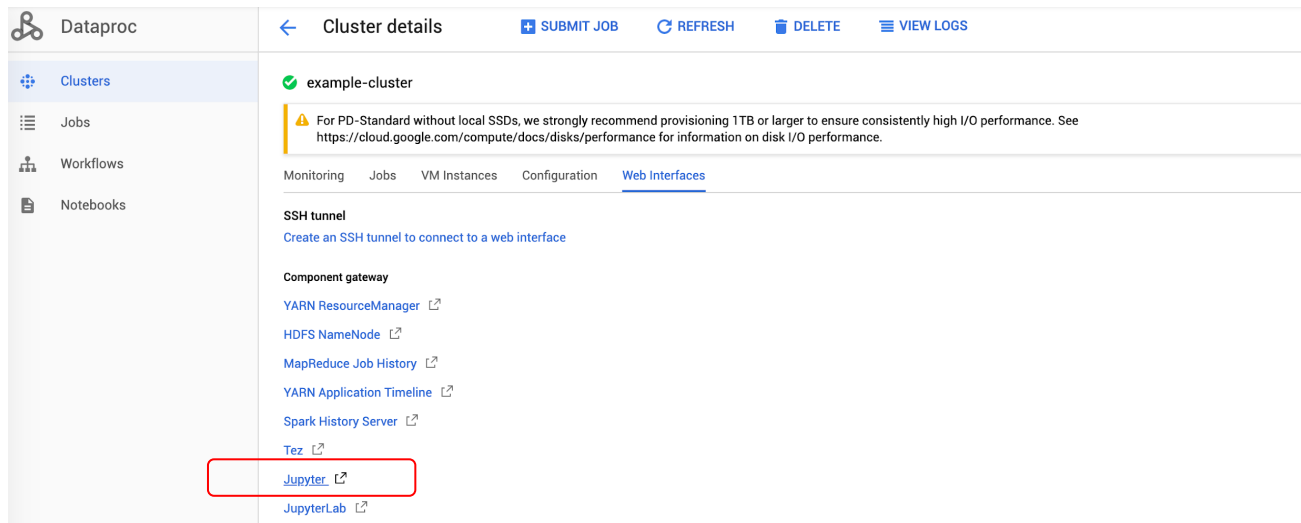
```
(base) conghan@Cong's-MacBook-Pro:~$ gcloud beta dataproc clusters create example-cluster --region=us-east1 --optional-components=ANACONDA,JUPYTER --image-version=1.3 --enable-component-gateway --bucket big-data-6893 --project big-data-6893-325519 --single-node --metadata 'PIP_PACKAGES=graphframes==0.6'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



- Cloud SDK:

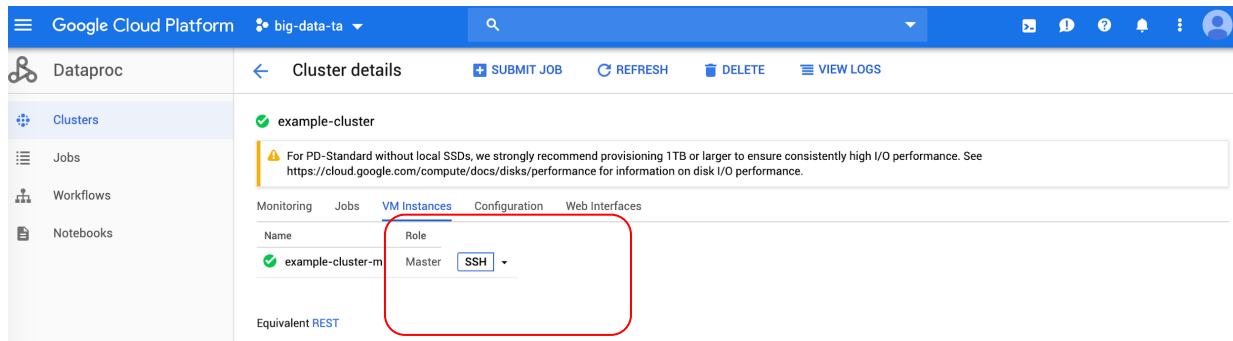
- `gcloud dataproc jobs submit pyspark <your_program.py> --cluster=<cluster-name>`

- [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



- pyspark

```
frouyang2@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

  ____  __
 / ___/ /  _ \
/ /   / /  / \
/ /___/ /  /_/ \
\____/_/  \____/

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.**