



# EECS E6893 Big Data Analytics

## Intro to Big Data Analytics on GCP

Cong Han, [ch3212@columbia.edu](mailto:ch3212@columbia.edu)

# Agenda

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



# Google Cloud Platform (GCP)

# 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/>

# GCP Setup

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

# Solve real business challenges on Google Cloud

Get started for free

[Contact sales](#)

## Run workloads for free

### 20+ free products

Get free hands-on experience with popular products, including Compute Engine and Cloud Storage, [up to monthly limits](#). These free services don't expire.

### \$300 in free credits

New customers also get [\\$300 in free credits](#) to fully explore and conduct an assessment of Google Cloud Platform. You won't be charged until you choose to upgrade.

 Try Google Cloud for free

## Step 1 of 3 Account Information



Cong Han  
conghanbigdata@gmail.com

[SWITCH ACCOUNT](#)

### Country

United States ▼

### What best describes your organization or needs?

Please select  
Class project / assignment ▼

### 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

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

## Access to all Cloud Platform Products

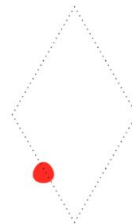
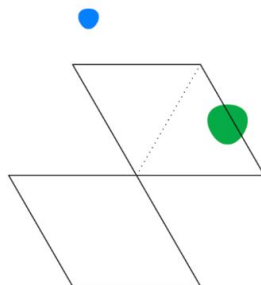
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

## \$300 credit for free

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. You won't be charged unless you manually upgrade to a paid account.



 Try Google Cloud for free

## Step 2 of 3 Identity Verification and Contact Information

Confirm where we can reach you about solutions to support your Cloud experience. Continue with the number associated with your Google account or choose a different one. ?



CONTINUE

[USE A DIFFERENT NUMBER](#)

### Access to all Cloud Platform Products

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

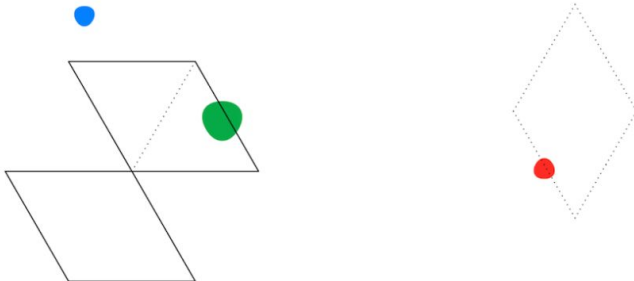
### \$300 credit for free

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. You won't be charged unless you manually upgrade to a paid account.

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





 Try Google Cloud for free

## Step 3 of 3 Payment Information Verification

Your payment information helps us reduce fraud and abuse. **You won't be charged unless you turn on automatic billing.**


 Account type 

Individual

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](#)

### Payment method

 Add credit or debit card 

Card number  
#  MM / YY CVC  
Card number is required  
Cardholder name  
Cong Han

 Billing address

When billing starts, you'll be charged automatically, typically monthly.

## Access to all Cloud Platform Products

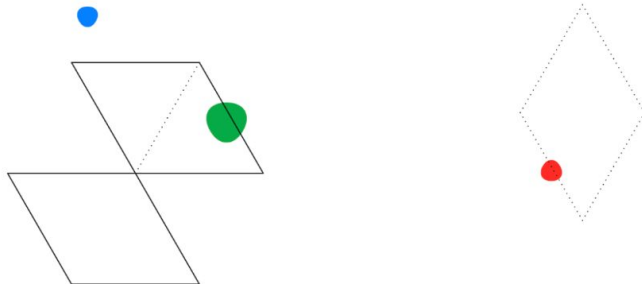
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

## \$300 credit for free

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. You won't be charged unless you manually upgrade to a paid account.





Free Trial and Free Tier | Google

Overview – Billing – My First Project

console.cloud.google.com/billing/012BBC-487AF6-EC306F?project=fiery-cabinet-325519

Google Cloud Platform

Search products and resources

1

Billing

Billing account

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

\$0

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

11

# 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

Free Trial and Free Tier | Google | Home - My First Project - Google

console.cloud.google.com/home/dashboard?folder=&organizationId=&project=fiery-cabinet-325519

Google Cloud Platform | My First Project | Search products and resources

DASHBOARD | ACTIVITY | RECOMMENDATIONS | CUSTOMIZE

Home | Recent | Pins appear here | Marketplace | Billing | APIs & Services | Support | IAM & Admin | Getting started | Compliance | Security | Anthos | COMPUTE | Compute Engine | Kubernetes Engine | VMware Engine | SERVERLESS

**Select a project**

NEW PROJECT

Search projects and folders

RECENT | STARRED | ALL

Name	ID
✓ ☆ My First Project ?	fiery-cabinet-325519

CANCEL OPEN

Google Cloud Platform status

Google Kubernetes Engine

Europe-west3, Europe-west4, US-east4, Asia-northeast1: Elevated error rates for GKE control plane

Began at 2021-09-09 (12:06:03)

All times are US/Pacific

Data provided by status.cloud.google.com

Go to Cloud status dashboard

Monitoring

Create my dashboard

Set up alerting policies

Create uptime checks

View all dashboards

Go to Monitoring

API Error Reporting

Free Trial and Free Tier | Google | Home - big data 6893 - Google

console.cloud.google.com/home/dashboard?folder=&organizationId=&project=big-data-6893-325519

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 Platform big data 6893 Search products and resources

Home Recent Pins appear here Marketplace Billing API APIS & Services Support IAM & Admin Getting started Compliance Security Anthos COMPUTE Compute Engine Kubernetes Engine VMware Engine

DASHBOARD ACTIVITY RECOMMENDATIONS CUSTOMIZE

Join us October 12-14 for Google Cloud Next. Register [here](#). DISMISS

**Project info**

Project name  
big data 6893

Project ID  
big-data-6893-325519

Project number  
881004012112

[ADD PEOPLE TO THIS PROJECT](#)

[Go to project settings](#)

**Resources**

This project has no resources

**Trace**

No trace data from the past 7 days

[Get started with Trace](#)

**API APIS**

Requests (requests/sec)

1.0  
0.8  
0.6  
0.4  
0.2  
0

2:45 3 PM 3:15 3:30

[Go to APIs overview](#)

**Google Cloud Platform status**

Google Kubernetes Engine  
europe-west3, europe-west4, us-east4, asia-northeast1: Elevated error rates for GKE control plane  
Began at 2021-09-09 (12:06:03)

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

[Go to Cloud status dashboard](#)

**Monitoring**

Create my dashboard

Set up alerting policies

Create uptime checks

[View all dashboards](#)

[Go to Monitoring](#)

# 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

# GCP: console

Search for services here

The screenshot shows the Google Cloud Platform console dashboard for the project 'big-data-ta'. The top navigation bar is blue and contains the Google Cloud Platform logo, the project name 'big-data-ta', a search bar, and several utility icons. A red rectangle highlights the search bar. Below the navigation bar, the dashboard is divided into several sections. On the left, there is a 'Project info' section showing the project name, ID, and number, and a 'Resources' section listing Compute Engine instances, Storage buckets, and BigQuery datasets. In the center, there is a 'Compute Engine' section showing a line graph of CPU utilization over time, with a red rectangle highlighting the graph. On the right, there is a 'Google Cloud Platform status' section showing 'All services normal', a 'Billing' section showing estimated charges for the billing period Sep 1 - 6, 2019, and an 'Error Reporting' section showing 'No sign of any errors. Have you set up Error Reporting?'. A red rectangle highlights the 'Billing' section. At the bottom, there is a section for 'API APIs' showing 'Requests (requests/sec)' with a red rectangle highlighting the section.

Google Cloud Platform big-data-ta

DASHBOARD ACTIVITY

CUSTOMIZE

**Project info**

Project name  
big-data-ta

Project ID  
logical-host-251101

Project number  
312759131343

[ADD PEOPLE TO THIS PROJECT](#)

[Go to project settings](#)

**Resources**

- Compute Engine  
1 instance
- Storage  
2 buckets
- BigQuery  
1 dataset

**Compute Engine**

CPU (%)

0.0300

0.0225

0.0150

1:30 1:45 2 PM 2:15

instance/cpu/utilization: 0.016

[Go to Compute Engine](#)

**Google Cloud Platform status**

All services normal

[Go to Cloud status dashboard](#)

**Billing**

Estimated charges  
For the billing period Sep 1 - 6, 2019

USD \$0.00

[View detailed charges](#)

**Error Reporting**

No sign of any errors. Have you set up Error Reporting?

[Learn how to set up Error Reporting](#)

**API APIs**

Requests (requests/sec)

0.26

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`

Free Trial and Free Tier | Google | Home - big data 6893 - Google | Quickstart: Getting started with Cloud SDK

cloud.google.com/sdk/docs/quickstart

Google Cloud

Why Google | Solutions | Products | Pricing | Getting Started

Docs | Support

English | Console

Cloud SDK: Command Line Interface

Overview | Guides | Reference | Support | Resources

Cloud SDK

Product overview

gcloud CLI overview

gcloud CLI cheat sheet

Quickstarts

All quickstarts

Getting started with Cloud SDK

How-to guides

All how-to guides

Installing the SDK

Setting up the SDK

Managing SDK components

Scripting guidelines

Enabling accessibility features

Using gcloud interactive shell

Uninstalling the Cloud SDK

Installing the latest Cloud SDK version (356.0.0)

★ 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

1. Cloud SDK requires Python:

Supported versions are Python 3 (**3.7 recommended**) and Python 2 (2.7.9 or higher).

Modern versions of macOS include the appropriate version of Python required for the Cloud SDK. To check your current Python version, run `python -V`.

For Cloud SDK release version 352.0.0 and above, the main install script offers to install CPython's Python 3.7 on Intel-based Macs.

For more information on how to choose and configure your Python interpreter, refer to [gcloud topic startup](#).

2. Download one of the following:

★ Note: To determine your machine hardware name, run `uname -m` from your command line.

Platform	Package	Size	SHA256 Checksum
macOS 64-bit (x86_64)	<a href="#">google-cloud-sdk-356.0.0-darwin-x86_64.tar.gz</a>	88.1 MB	98f9353538cca55fe43f4bc2d75237f827bca986661c0d8d46fc34852492b940
macOS 64-bit (arm64, Apple M1 silicon)	<a href="#">google-cloud-sdk-356.0.0-darwin-arm.tar.gz</a>	88.0 MB	9372bf69982f40aeb0ca91cc47a579e56f04381471ef32bd72c5936205ddf13b
macOS 32-bit	<a href="#">google-cloud-sdk-356.0.0-darwin-i386.tar.gz</a>	91.8 MB	5ef09ff44bbaadh8f5c9bc705ba2e320had87h

Table of contents

[Installing the latest Cloud SDK version \(356.0.0\)](#)

Optional: Install the latest Google Cloud Client Libraries

Initializing the Cloud SDK

Running core commands

What's next

18

```
(base) conghan@Congs-MacBook-Pro:~/Downloads$ clear  
(base) conghan@Congs-MacBook-Pro:~/Downloads$ ./google-cloud-sdk/install.sh
```

Installed	Cloud Storage Command Line Tool	gsutil	4.3 MiB
-----------	---------------------------------	--------	---------

To install or remove components at your current SDK version [356.0.0], run:

```
$ gcloud components install COMPONENT_ID
$ gcloud components remove COMPONENT_ID
```

To update your SDK installation to the latest version [356.0.0], run:

```
$ gcloud components update
```

Modify profile to update your \$PATH and enable shell command completion?

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

The Google Cloud SDK installer will now prompt you to update an rc file to bring the Google Cloud CLIs into your environment.

Enter a path to an rc file to update, or leave blank to use

[/Users/conghan/.bash\_profile]:

Backing up [/Users/conghan/.bash\_profile] to [/Users/conghan/.bash\_profile.backup].

[/Users/conghan/.bash\_profile] has been updated.

=> Start a new shell for the changes to take effect.

Cloud SDK works best with Python 3.7 and certain modules.

Download and run Python 3.7 installer? (Y/n)? y

Running Python 3.7 installer, you may be prompted for sudo password...

Password:

installer: Package name is Python

installer: Upgrading at base path /

installer: The upgrade was successful.

Setting up virtual environment

Creating virtualenv...

Installing modules...

```
| 89 kB 4.4 MB/s
| 3.9 MB 9.1 MB/s
| 2.0 MB 8.6 MB/s
| 145 kB 9.6 MB/s
| 176 kB 24.4 MB/s
| 112 kB 23.2 MB/s
```

Running setup.py install for crcmod ... done

Virtual env enabled.

For more information on how to get started, please visit:

<https://cloud.google.com/sdk/docs/quickstarts>

\*(base) conghan@Congs-MacBook-Pro:~/Downloads\$

```
Download and run Python 3.7 installer? (Y/n)? y
```

```
Running Python 3.7 installer, you may be prompted for sudo password...
```

```
Password:
```

```
installer: Package name is Python
```

```
installer: Upgrading at base path /
```

```
installer: The upgrade was successful.
```

```
Setting up virtual environment
```

```
Creating virtualenv...
```

```
Installing modules...
```

	89 kB	4.4 MB/s
	3.9 MB	9.1 MB/s
	2.0 MB	8.6 MB/s
	145 kB	9.6 MB/s
	176 kB	24.4 MB/s
	112 kB	23.2 MB/s

```
Running setup.py install for crcmod ... done
```

```
Virtual env enabled.
```

```
For more information on how to get started, please visit:
```

```
https://cloud.google.com/sdk/docs/quickstarts
```

```
(base) conghan@Cong's-MacBook-Pro:~/Downloads$ ./google-cloud-sdk/bin/gcloud init
```

```
Welcome! This command will take you through the configuration of gcloud.
```

```
Your current configuration has been set to: [default]
```

```
You can skip diagnostics next time by using the following flag:
```

```
gcloud init --skip-diagnostics
```

```
Network diagnostic detects and fixes local network connection issues.
```

```
Checking network connection...done.
```

```
Reachability Check passed.
```

```
Network diagnostic passed (1/1 checks passed).
```

```
You must log in to continue. Would you like to log in (Y/n)? y
```

```
Your browser has been opened to visit:
```

```
https://accounts.google.com/o/oauth2/auth?response\_type=code&client\_id=32555940559.apps.googleusercontent.com&redirect\_uri=http%3A%2F%2Flocalhost%3A8085%2F&scope=openid+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fuserinfo.email+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fcloud-platform+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fappengine.admin+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fcompute+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Faccounts.reauth&state=ml7Vnevs9DKqLLSRDyWzsfDgcRgYBW&access\_type=offline&code\_challenge=Y\_hSRd9TakgmBNRj1qkLJghIlcIum9mbQs9jjdk3KXI&code\_challenge\_method=S256
```

```
You are logged in as: [conghanbigdata@gmail.com].
```

```
Pick cloud project to use:
```

```
[1] big-data-6893-325519
```

```
[2] fiery-cabinet-325519
```

```
[3] Create a new project
```

```
Please enter numeric choice or text value (must exactly match list
```

```
item): 1
```

```

*(base) conghan@Congss-MacBook-Pro:~$ gcloud config list
[core]
account = conghanbigdata@gmail.com
disable_usage_reporting = False
project = big-data-6893-325519

Your active configuration is: [default]
*(base) conghan@Congss-MacBook-Pro:~$ gcloud info
Google Cloud SDK [356.0.0]

Platform: [Mac OS X, x86_64] uname_result(system='Darwin', node='Congss-MacBook-Pro.local', release='20.6.0', version='Darwin Kernel Version 20.6.0: Wed Jun 23 00:26:31 PDT 2021; root:xnu-7195.141.2~5/RELEASE_ARM_T8020'
machine='x86_64', processor='i386')
Locale: ('en_US', 'UTF-8')
Python Version: [3.7.9 (v3.7.9:13c94747c7, Aug 15 2020, 01:31:08)] [Clang 6.0 (clang-600.0.57)]
Python Location: [/Users/conghan/.config/gcloud/virtenv/bin/python3]
Site Packages: [Enabled]

Installation Root: [/Users/conghan/Downloads/google-cloud-sdk]
Installed Components:
  gsutil: [4.67]
  core: [2021.09.03]
  bq: [2.0.71]
System PATH: [/Users/conghan/.config/gcloud/virtenv/bin:/Users/conghan/Downloads/google-cloud-sdk/bin:/Users/conghan/anaconda3/bin:/Users/conghan/anaconda3/condabin:/Users/conghan/anaconda3/Library/Frameworks/Python.framework/Versions/3.6/bin:/Library/Frameworks/Python.framework/Versions/3.5/bin:/usr/local/bin:/usr/bin:/bin:/usr/sbin:/sbin:/Library/TeX/texbin]
Python PATH: [/Users/conghan/Downloads/google-cloud-sdk/lib/third_party:/Users/conghan/Downloads/google-cloud-sdk/lib/Library/Frameworks/Python.framework/Versions/3.7/lib/python37.zip:/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7:/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/lib-dynload:/Users/conghan/.config/gcloud/virtenv/lib/python3.7/site-packages]
Cloud SDK on PATH: [True]
Kubectl on PATH: [False]

Installation Properties: [/Users/conghan/Downloads/google-cloud-sdk/properties]
User Config Directory: [/Users/conghan/.config/gcloud]
Active Configuration Name: [default]
Active Configuration Path: [/Users/conghan/.config/gcloud/configurations/config_default]

Account: [conghanbigdata@gmail.com]
Project: [big-data-6893-325519]

Current Properties:
[core]
  account: [conghanbigdata@gmail.com]
  disable_usage_reporting: [False]
  project: [big-data-6893-325519]

Logs Directory: [/Users/conghan/.config/gcloud/logs]
Last Log File: [/Users/conghan/.config/gcloud/logs/2021.09.09/16.00.44.581670.log]

git: [xcrun: error: invalid active developer path (/Library/Developer/CommandLineTools), missing xcrun at: /Library/Developer/CommandLineTools/usr/bin/xcrun]
ssh: [OpenSSH_8.1p1, LibreSSL 2.7.3]

*(base) conghan@Congss-MacBook-Pro:~$

```

# GCP: Cloud Shell



The screenshot shows the Google Cloud Platform dashboard for the project 'big-data-ta'. The top navigation bar is blue and contains the Google Cloud Platform logo, the project name, a search bar, and several utility icons. The 'Activate Cloud Shell' button, represented by a terminal icon, is highlighted with a red rectangle. Below the navigation bar, the dashboard is divided into three main sections. The left section, titled 'Project info', displays the project name 'big-data-ta', Project ID 'logical-host-251101', and Project number '312759131343', with a link to 'Go to project settings'. The middle section, titled 'API APIs', shows a line graph for 'Requests (requests/sec)' with a y-axis ranging from 0.5 to 2.5. The right section contains two cards: 'Google Cloud Platform status' showing 'All services normal' with a link to 'Go to Cloud status dashboard', and 'Billing' showing 'Estimated charges' of 'USD \$0.00' for the billing period 'Sep 1 - 12, 2019'.

Google Cloud Platform big-data-ta

DASHBOARD ACTIVITY

Activate Cloud Shell CUSTOMIZE

**Project info**

- Project name: big-data-ta
- Project ID: logical-host-251101
- Project number: 312759131343

→ Go to project settings

**API APIs**

Requests (requests/sec)

Google Cloud Platform status

All services normal

→ Go to Cloud status dashboard

**Billing**

Estimated charges: USD \$0.00

For the billing period Sep 1 - 12, 2019

persistent home directory :)

# GCP: Cloud Shell

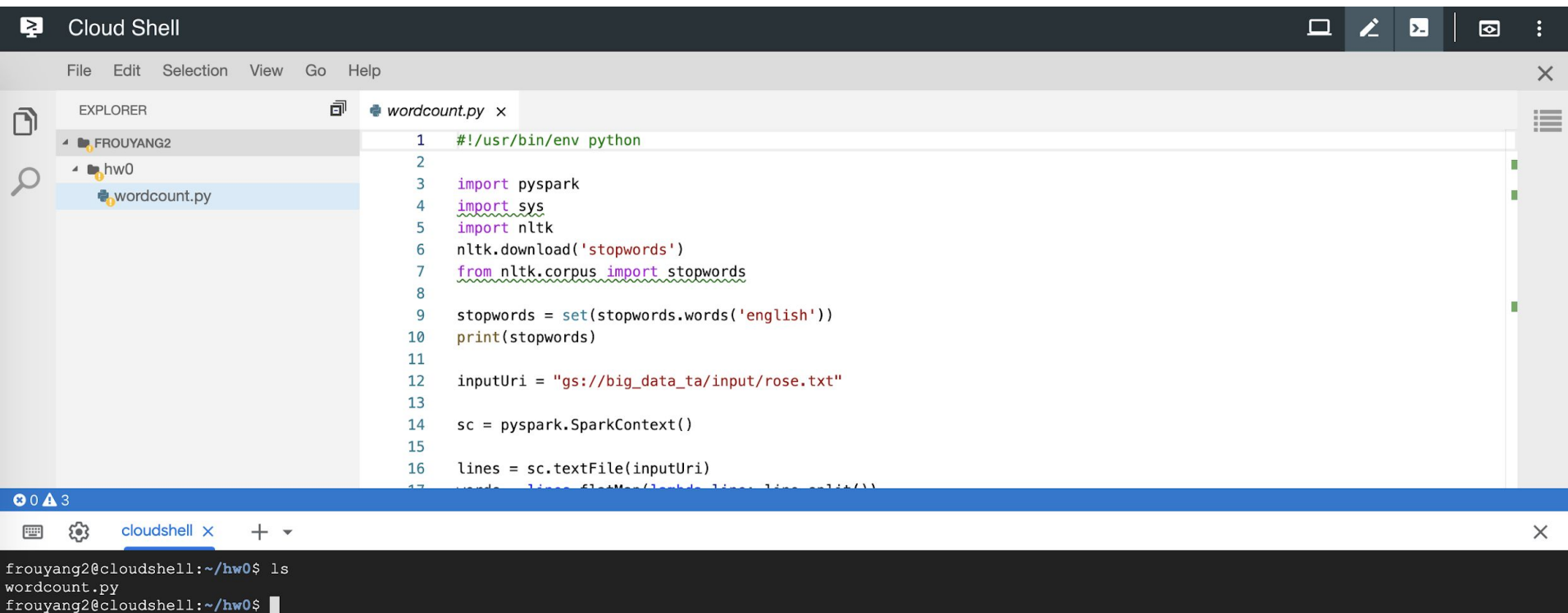
The screenshot displays the Google Cloud Platform (GCP) dashboard interface. At the top, a blue header bar contains the GCP logo, the text "Google Cloud Platform", a dropdown menu showing "big-data-ta", a search bar, and several utility icons (notifications, help, user profile). Below the header, a navigation bar shows "DASHBOARD" and "ACTIVITY" tabs, with a "CUSTOMIZE" link on the right. The main content area is divided into three panels:

- Project info:** Displays details for the project "big-data-ta", including the Project ID "logical-host-251101" and the Project number.
- API APIs:** Shows a line graph for "Requests (requests/sec)" with a y-axis ranging from 2.0 to 2.5. A single data point is visible at approximately 2.1 requests/sec.
- Google Cloud Platform status:** Indicates that "All services normal" and provides a link to "Go to Cloud status dashboard".

At the bottom, a terminal window titled "cloudshell" is open, showing the command prompt "frouyang2@cloudshell:~\$". The user has entered "ls" and "README-cloudshell.txt" is listed. A "Launch code editor BETA" button is visible in the bottom right corner of the terminal area.



# 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`

# Cloud Storage

The screenshot shows the Google Cloud Platform console interface. On the left, a navigation menu is open, highlighting the 'STORAGE' section. Within this section, 'Filestore', 'Cloud Storage', and 'Data Transfer' are listed. 'Cloud Storage' is selected, and a sub-menu is visible with options: 'Browser', 'Monitoring', and 'Settings'. The main content area displays a 'Google Cloud Platform status' card indicating 'All services normal', a 'Monitoring' card with options to create dashboards and set up alerting, and an 'API Error Reporting' card. A red rectangle highlights the 'Cloud Storage' menu item and its sub-menu.

Free Trial and Free Tier | Google Cloud Platform

Home - big data 6893 - Google Cloud Platform

console.cloud.google.com/home/dashboard?folder=&organizationId=&project=big-data-6893-325519

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 Platform big data 6893

Search products and resources

Home

Recent

SERVERLESS

Cloud Run

Cloud Functions

App Engine

STORAGE

Filestore

Cloud Storage

Data Transfer

DATABASES

Bigtable

Datastore

Database Migration

Firestore

Memorystore

RECOMMENDATIONS

cloud Next. Register [here](#). DISMISS

RPI 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

Monitoring

Create my dashboard

Set up alerting policies

Create uptime checks

View all dashboards

Go to Monitoring

RPI Error Reporting

No sign of any errors. Have you set up Error Reporting?

<https://console.cloud.google.com/storage?project=big-data-6893-325519>

# Cloud Storage

Free Trial and Free Tier | Google Cloud

Browser - Cloud Storage - big data 6893

console.cloud.google.com/storage/browser?cloudshell=false&project=big-data-6893-325519&prefix=

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](#)

LEARN Home

Google Cloud Platform big data 6893 Search products and resources

Cloud Storage

Browser

Monitoring

Settings

Browser

Filter Filter buckets

CREATE BUCKET DELETE REFRESH SHOW INFO PANEL

Name	Created	Location type	Location	Default storage class	Updated	Public access
No rows to display						

Store and retrieve your data

Get started by creating a bucket — a container where you can organize and control access to your data and files in Cloud Storage.

CREATE BUCKET TAKE QUICKSTART

Recommended for you

Create a storage bucket

Create a cloud storage bucket and learn about storage location, class, and access control. [Tutorials](#) 5 min

Transfer data into Cloud Storage

Move, back up, or archive data from another cloud provider or storage service. [Tutorials](#)

Host website content

Learn how to set up a bucket to serve content for a static website. [Tutorials](#)

You might also like

Tutorials

Walkthroughs and guides

Concepts

Deep dive explanations

API & references

API and command-line resources

Resources

Pricing, release notes, and tools

Access control

Permissions and privacy tools

All product documentation

Not seeing what you need? [Give feedback](#)

# Cloud Storage

Free Trial and Free Tier | Google

Create a bucket - Cloud Storage

console.cloud.google.com/storage/create-bucket?cloudshell=false&project=big-data-6893-325519

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 Platform

big data 6893

Search products and resources

Cloud Storage

Create a bucket

Monthly cost estimate

Browser

Monitoring

Settings

✓

Name your bucket

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

6893\_data

Tip: Don't include any sensitive information

CONTINUE

• Choose where to store your data

• Choose a default storage class for your data

• Choose how to control access to objects

• Advanced settings (optional)

CREATE

CANCEL

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

Storage and retrieval

Storage size

GB

\$0.026 per GB-month

Data retrieval size

GB

Free

Operations

Class A operations

per-month

\$0.005 per 1,000 ops

Class B operations

per-month

\$0.0004 per 1,000 ops

Availability SLA: 99.95%

Monthly cost: \$0.00

Currency: US Dollar (\$) ▼

# Cloud Storage

Free Trial and Free Tier | Google Cloud

Create a bucket - Cloud Storage

console.cloud.google.com/storage/create-bucket?cloudshell=false&project=big-data-6893-325519

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 Platform

big data 6893

Search products and resources

Cloud Storage

Browser

Monitoring

Settings

Create a bucket

Name your bucket

Choose where to store your data

This permanent choice defines the geographic placement of your data and affects cost, performance, and availability. [Learn more](#)

Location type

☐ Multi-region

Highest availability across largest area

☐ Dual-region

High availability and low latency across 2 regions

☒ Region

Lowest latency within a single region

Location

us-east1 (South Carolina)

CONTINUE

Choose a default storage class for your data

Choose how to control access to objects

Advanced settings (optional)

CREATE

CANCEL

Monthly cost estimate

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

Storage and retrieval

Storage size

GB

\$0.020 per GB-month

Data retrieval size

GB

Free

Operations

Class A operations

per-month

\$0.005 per 1,000 ops

Class B operations

per-month

\$0.0004 per 1,000 ops

Availability SLA: 99.9%

Monthly cost: \$0.00

Currency: US Dollar (\$) ▼

# Cloud Storage

Free Trial and Free Tier | Google

Create a bucket - Cloud Storage

console.cloud.google.com/storage/create-bucket?cloudshell=false&project=big-data-6893-325519

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 Platform

big data 6893

Search products and resources

Cloud Storage

Create a bucket

Monthly cost estimate

Browser

Monitoring

Settings

✓

Name your bucket

✓

Choose where to store your data

•

Choose a default storage class for your data

A storage class sets costs for storage, retrieval, and operations. Pick a default storage class based on how long you plan to store your data and how often it will be accessed. [Learn more](#)

☒ 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

•

Advanced settings (optional)

CREATE

CANCEL

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

Storage and retrieval

Storage size

GB

\$0.020 per GB-month

Data retrieval size

GB

Free

Operations ⓘ

Class A operations

per-month

\$0.005 per 1,000 ops

Class B operations

per-month

\$0.0004 per 1,000 ops

Availability SLA: 99.9%

Monthly cost: \$0.00

Currency: US Dollar (\$) ▼

Release Notes

&lt;|



# Cloud Storage

Free Trial and Free Tier | Google

Create a bucket - Cloud Storage

+

console.cloud.google.com/storage/create-bucket?cloudshell=false&project=big-data-6893-325519

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 Platform

big data 6893

Search products and resources

Cloud Storage

Create a bucket

Monthly cost estimate

Browser

Monitoring

Settings

✓

Name your bucket

✓

Choose where to store your data

✓

Choose a default storage class for your data

•

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

• Advanced settings (optional)

CREATE

CANCEL

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

Storage and retrieval

Storage size

GB

\$0.020 per GB-month

Data retrieval size

GB

Free

Operations

Class A operations

per-month

\$0.005 per 1,000 ops

Class B operations

per-month

\$0.0004 per 1,000 ops

Availability SLA: 99.9%

Monthly cost: \$0.00

Currency: US Dollar (\$) ▼

33

# Cloud Storage

Free Trial and Free Tier | Google Cloud

Create a bucket - Cloud Storage

console.cloud.google.com/storage/create-bucket?cloudshell=false&project=big-data-6893-325519

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 Platform

big data 6893

Search products and resources

Cloud Storage

Create a bucket

Browser

Monitoring

Settings

✓ Choose a default storage class for your data

✓ Choose how to control access to objects

• Advanced settings (optional)

Encryption

☒ Google-managed encryption key  
No configuration required

☐ Customer-managed encryption key (CMEK)  
Manage via Google Cloud Key Management Service

Retention policy

Set a retention policy to specify the minimum duration that this bucket's objects must be protected from deletion or modification after they're uploaded. You might set a policy to address industry-specific retention challenges. [Learn more](#)

☐ Set a retention policy

Labels

Labels are key:value pairs that allow you to group related buckets together or with other Cloud Platform resources. [Learn more](#)

CREATE

CANCEL

Monthly cost estimate

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

Storage and retrieval

Storage size

GB

\$0.020 per GB-month

Data retrieval size

GB

Free

Operations

Class A operations

per-month

\$0.005 per 1,000 ops

Class B operations

per-month

\$0.0004 per 1,000 ops

Availability SLA: 99.9%

Monthly cost: \$0.00

Currency: US Dollar (\$) ▼

# Cloud Storage

Free Trial and Free Tier | Google

data - big-data-6893 - Bucket

+

console.cloud.google.com/storage/browser/big-data-6893/data;tab=objects?cloudshell=false&project=big-data-6893-325519&pageState={"StorageObjectListTable":{"f":"...}}

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 Platform

big data 6893

Search products and resources

2

Cloud Storage

Browser

Monitoring

Settings

Bucket details

REFRESH

LEARN

big-data-6893

OBJECTS

CONFIGURATION

PERMISSIONS

RETENTION

LIFECYCLE

Buckets > big-data-6893 > data

UPLOAD FILES

UPLOAD FOLDER

CREATE FOLDER

MANAGE HOLDS



DOWNLOAD

DELETE

Filter by name prefix only

Filter

Filter objects and folders

<input type="checkbox"/>	Name	Size	Type	Created time	Storage class	Last modified	Public access	Encryption	Retention
<input type="checkbox"/>	 data_citibike_stations.csv	114.3 KB	text/csv	Sep 9, 2021, 4:...	Standard	Sep 9, 202...	Not public	Google-managed key	— 

Release Notes

# Cloud Storage

The screenshot shows the Google Cloud Platform console interface. The browser address bar displays the URL: `console.cloud.google.com/storage/browser/_details/big-data-6893/data/data_citibike_stations.csv?cloudshell=false&project=big-data-6893-325519`. The console header includes the Google Cloud Platform logo, the project name 'big data 6893', a search bar, and navigation icons. The left sidebar shows the 'Cloud Storage' section with options for 'Browser', 'Monitoring', and 'Settings'. The main content area displays the 'Object details' for the object 'data\_citibike\_stations.csv' in the bucket 'big-data-6893'. The 'Overview' section lists various metadata fields. The 'Authenticated URI' and 'gsutil URI' are highlighted with a red box.

Overview	
Type	text/csv
Size	114.3 KB
Created	Sep 9, 2021, 4:35:06 PM
Last modified	Sep 9, 2021, 4:35:06 PM
Storage class	Standard
Custom time	—
Public URL	Not applicable
Authenticated URI	<a href="https://storage.cloud.google.com/big-data-6893/data/data_citibike_stations.csv">https://storage.cloud.google.com/big-data-6893/data/data_citibike_stations.csv</a>
gsutil URI	<a href="gs://big-data-6893/data/data_citibike_stations.csv">gs://big-data-6893/data/data_citibike_stations.csv</a>

Permissions	
Public access	Not public

Protection	
Hold status	None
Retention policy	None
Encryption type	Google-managed key

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

# BigQuery

Free Trial and Free Tier | Google Cloud Platform

SQL workspace - BigQuery - big-data-6893 - Bucket details

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

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 Platform big data 6893

bigque

FEATURES & INFO SHORTCUT DISABLE EDITOR TABS

Explorer + ADD DATA

EDITOR

COMPOSE NEW QUERY

RUN SAVE SCHEDULE MORE

Type a query to get started

Type to search

Viewing pinned projects.

big-data-6893-325519

Open

Create dataset

JOB HISTORY QUERY HISTORY SAVED QUERIES



# BigQuery

The screenshot shows the Google Cloud Platform console interface. The top navigation bar includes the Google Cloud logo, a free trial status message, and a search bar. The left sidebar contains the 'Explorer' panel with a search bar and a list of projects, including 'big-data-6893-325519'. The main content area is titled 'Create dataset' and contains the following fields and options:

- Dataset ID:** A text input field containing 'dataset1'. Below it, a note states: 'Letters, numbers, and underscores allowed'.
- Data location:** A dropdown menu set to 'Default'.
- Default table expiration:** A section with a checkbox 'Enable table expiration' (unchecked) and a sub-field 'Default maximum table age' set to 'Days'.
- Encryption:** A section with two radio button options:
  - Google-managed encryption key:** Selected by default. Subtext: 'No configuration required'.
  - Customer-managed encryption key (CMEK):** Subtext: 'Manage via Google Cloud Key Management Service'.

At the bottom of the form are two buttons: 'CREATE DATASET' (in blue) and 'CANCEL'.

# BigQuery

The screenshot displays the Google Cloud Platform BigQuery console. At the top, a browser window shows the URL `console.cloud.google.com/bigquery?referrer=search&cloudshell=false&project=big-data-6893-325519`. Below the browser, a notification bar indicates a free trial status with a \$300.00 credit and 91 days remaining. The main interface features a blue header with the Google Cloud Platform logo, the project name 'big data 6893', and a search bar containing 'bigque'. The left sidebar contains navigation links for 'FEATURES & INFO', 'SHORTCUT', and 'DISABLE EDITOR TABS'. The 'Explorer' panel on the left shows a search bar, a list of pinned projects, and a tree view of the project 'big-data-6893-325519' containing a dataset named 'dataset1'. A context menu is open over 'dataset1' with 'Open' and 'Delete' options. The main editor area has tabs for 'EDITOR' and 'COMPOSE NEW QUERY', and a prompt to 'Type a query to get started'. At the bottom, a notification bar states 'dataset1 created.' with a 'GO TO DATASET' link. The footer includes links for 'JOB HISTORY', 'QUERY HISTORY', and 'SAVED QUERIES'.

Free Trial and Free Tier | Google Cloud

BigQuery - big data 6893 - Google Cloud

big-data-6893 - Bucket details

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

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 Platform big data 6893

bigque

FEATURES & INFO SHORTCUT DISABLE EDITOR TABS

Explorer + ADD DATA

EDITOR X

COMPOSE NEW QUERY

RUN SAVE SCHEDULE MORE

Type a query to get started

Type to search

Viewing pinned projects.

big-data-6893-325519

dataset1

Open Delete

"dataset1" created. GO TO DATASET

JOB HISTORY QUERY HISTORY SAVED QUERIES

# BigQuery

The screenshot shows the Google Cloud Platform BigQuery console. The browser address bar displays the URL: `console.cloud.google.com/bigquery?referrer=search&cloudshell=false&project=big-data-6893-325519&d=dataset1&p=big-data-6893-325519&page=dataset&ws=!1m4!1m...`. The Google Cloud Platform header shows the project name 'big data 6893' and a search bar with 'bigque'. The left sidebar contains the 'Explorer' panel with a search bar and a list of projects, including 'big-data-6893-325519' and 'dataset1'. The main panel shows the details for 'big-data-6893-325519:dataset1'. A red box highlights the '+ Create table' button. Below the button are links for 'SHARE DATASET', 'AUTHORIZE ROUTINES', 'COPY DATASET', and 'DELETE DATASET'. The 'Description' and 'Labels' sections both show 'None'. The 'Dataset info' section contains a table with the following data:

Dataset ID	big-data-6893-325519:dataset1
Created	Sep 9, 2021, 7:02:31 PM
Default table expiration	Never
Last modified	Sep 9, 2021, 7:02:31 PM
Data location	US

At the bottom, a notification bar states: "dataset1" created. GO TO DATASET. Below the notification bar are tabs for 'JOB HISTORY', 'QUERY HISTORY', and 'SAVED QUERIES'.

# BigQuery

Free Trial status: \$300.00 credit and 91 days remaining - with a full account, you'll get unlimited access

Google Cloud Platform

big data 6893

big data 6893 - Bucket details

console.cloud.google.com/bigquery?referrer=search&cloudshell=false&project=big-data-6893-325519&d=dataset1&p=big-data-6893-325519&page=dataset&ws=1m4!1m...

### Create table

Source

Create table from:

- Empty table
- Google Cloud Storage
- Upload
- Drive
- Google Cloud Bigtable

Destination

Search for a dataset

Project name: big data 6893

Dataset name: dataset1

Table type: Native table

Table name: Letters, numbers, and underscores allowed

Schema

Edit as text

+ Add field

Partition and cluster settings

Partitioning: No partitioning

Clustering order (optional):

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

Comma-separated list of fields to define clustering order (up to 4)

"dataset1" created. GO TO DATASET

# BigQuery

The screenshot shows the Google Cloud Platform BigQuery console. On the left, the Explorer pane displays the project 'big-data-6893-325519' and a dataset named 'dataset1'. The main area shows the 'Create table' dialog. The 'Source' section is set to 'Google Cloud Storage' with the file 'big-data-6893/data\_citibike\_stations.csv' selected. The 'File format' is 'CSV'. The 'Destination' section has 'Search for a project' selected, with 'Project name' set to 'big data 6893', 'Dataset name' set to 'dataset1', and 'Table type' set to 'Native table'. The 'Table name' is 'bike\_data'. The 'Schema' section has 'Auto detect' checked, and a message states 'Schema will be automatically generated.' The 'Partitioning' section is set to 'No partitioning'. The 'Clustering order (optional)' section is empty. At the bottom, the 'Advanced options' section is collapsed. A red box highlights the 'Create table' button at the bottom right of the dialog.

Free Trial status: \$300.00 credit and 91 days remaining - with a full account, you'll get unlimited access

Google Cloud Platform big data 6893

Explorer + ADD DATA

big-data-6893-325519:dataset1

Description

None

Dataset info

Dataset ID	big-data-6893-325519
Created	Sep 9, 2024
Default table expiration	Never
Last modified	Sep 9, 2024
Data location	US

### Create table

Source

Create table from:  Select file from GCS bucket: ☒ big-data-6893/data\_citibike\_stations.csv  File format:

☐ Source Data Partitioning

Destination

☒ Search for a project ☐ Enter a project name

Project name:  Dataset name:  Table type:

Table name:

Schema

Auto detect ☒ Schema and input parameters

Partitioning and cluster settings

Partitioning:

Clustering order (optional):

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

Advanced options

# BigQuery

Free Trial and Free Tier | Google Cloud

SQL workspace - BigQuery - b | big-data-6893 - Bucket details | +

console.cloud.google.com/bigquery?referrer=search&cloudshell=false&project=big-data-6893-325519&ws=1m51m41m311sbig-data-6893-325519!2sbqjob\_3e3595fa... ☆ ⚙️ C ⋮

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 Platform big data 6893 bigque

FEATURES & INFO SHORTCUT DISABLE EDITOR TABS

Explorer + ADD DATA

🔍 Type to search

Viewing pinned projects.

big-data-6893-325519 dataset1 bike\_data

RUN SAVE SCHEDULE MORE

This query will process 108.5 KiB when run.

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

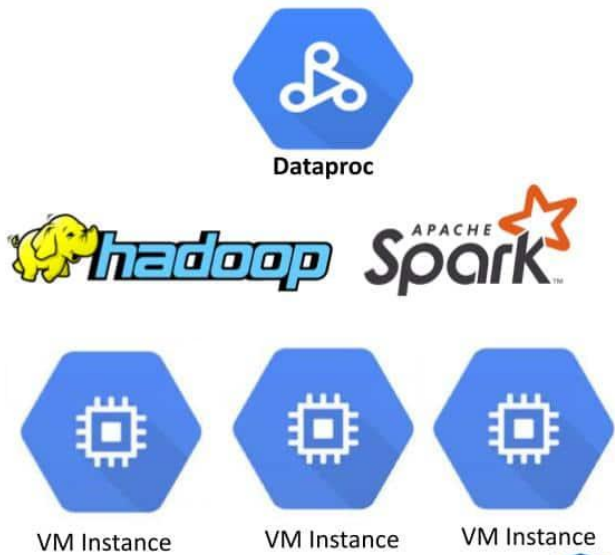


# Dataproc

# Dataproc

## What is dataproc?

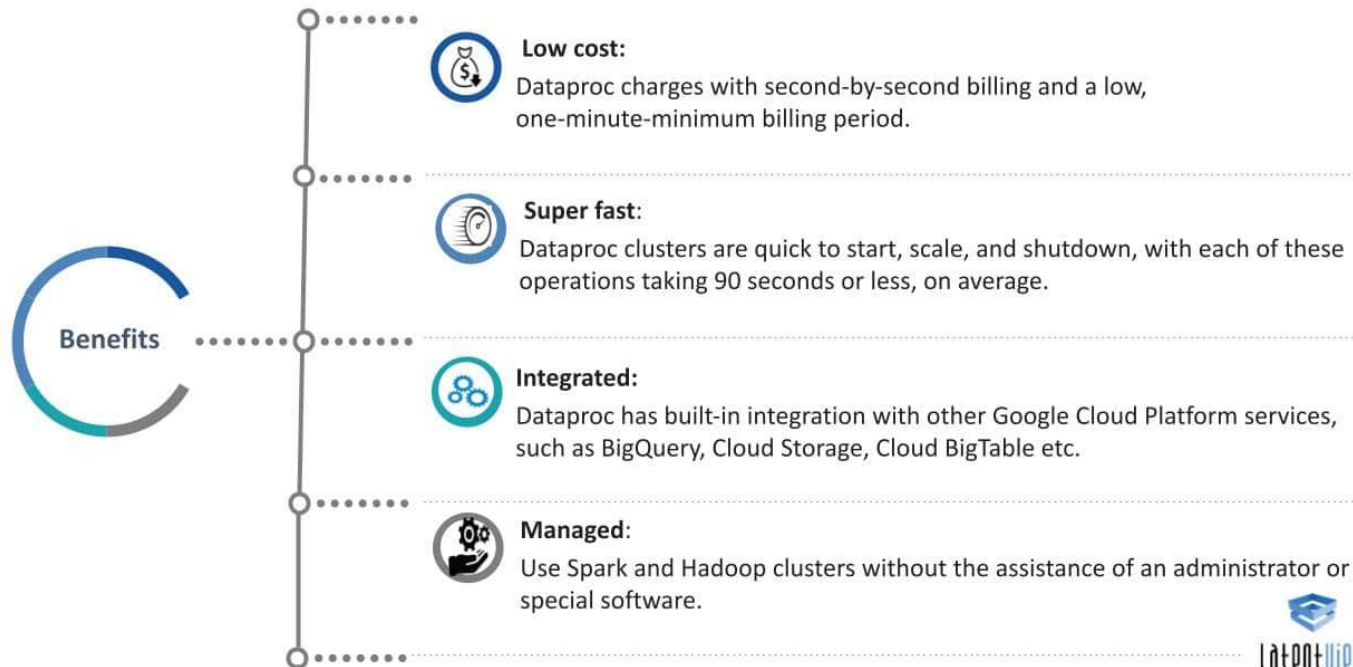
- Google Cloud Dataproc is a managed service for running **Apache Hadoop and Spark jobs**.
- Dataproc uses **Compute Engine instances** under the hood, but it takes care of the management details.
- Includes **Hadoop, Spark, Hive and Pig**.
- **Ideal for moving** existing code to GCP





# Dataproc

## Why dataproc?



# Dataproc

Free Trial and Free Tier | Google Cloud | SQL workspace - BigQuery - | big-data-6893 - Bucket detail | Cloud Dataproc API - Marketplace | +

← → ↺ console.cloud.google.com/marketplace/product/google/dataproc.googleapis.com?returnUrl=%2Fdataproc%2Fclusters%3Fcloudshell%3Dfalse%26project%3Dbig-data-6893... ☆ ⚙️ G ⋮

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 Platform ▶ big data 6893 🔍 📧 ⓘ 1 ⋮ 👤

←

## Cloud Dataproc API

Google Enterprise API

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

ENABLE TRY THIS API ↗

OVERVIEW

DOCUMENTATION

### Overview

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

### Additional details

Type: [SaaS & APIs](#)

Last updated: 7/22/21

Category: [Google Enterprise APIs](#)

Service name: dataproc.googleapis.com

### Tutorials and documentation

[Learn more ↗](#)

# Dataproc - graphical UI

The screenshot shows the Google Cloud Platform interface for the Dataproc Clusters page. The browser address bar shows the URL: `console.cloud.google.com/dataproc/clusters?cloudshell=false&project=big-data-6893-325519&folder=&organizationId=`. The page header includes the Google Cloud Platform logo, the project name "big data 6893", and a search bar containing "dataproc". The left sidebar lists navigation options: "Dataproc", "Jobs on Clusters", "Jobs", "Workflows", "Autoscaling policies", "Utilities", "Component exchange", "Metastore", and "Notebooks". The main content area is titled "Clusters" and includes buttons for "CREATE CLUSTER", "REFRESH", "START", "STOP", "DELETE", and "REGIONS". A message box states: "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." A "CREATE CLUSTER" button is visible at the bottom of the message box.

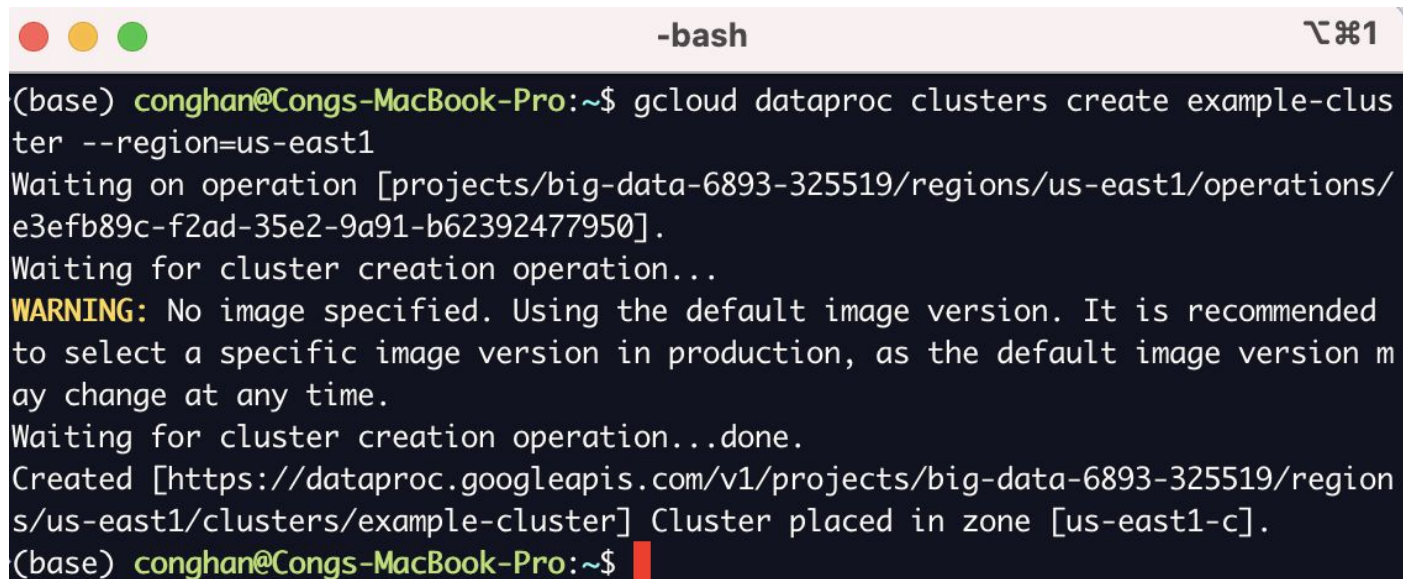
# Dataproc - Cloud SDK

Cluster creation (using Cloud SDK):

```
(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 light pink title bar containing three colored window control buttons (red, yellow, green) on the left, the text '-bash' in the center, and 'v#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/10.142.0.3:4040, (http://10.142.0.3:4040)
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
```

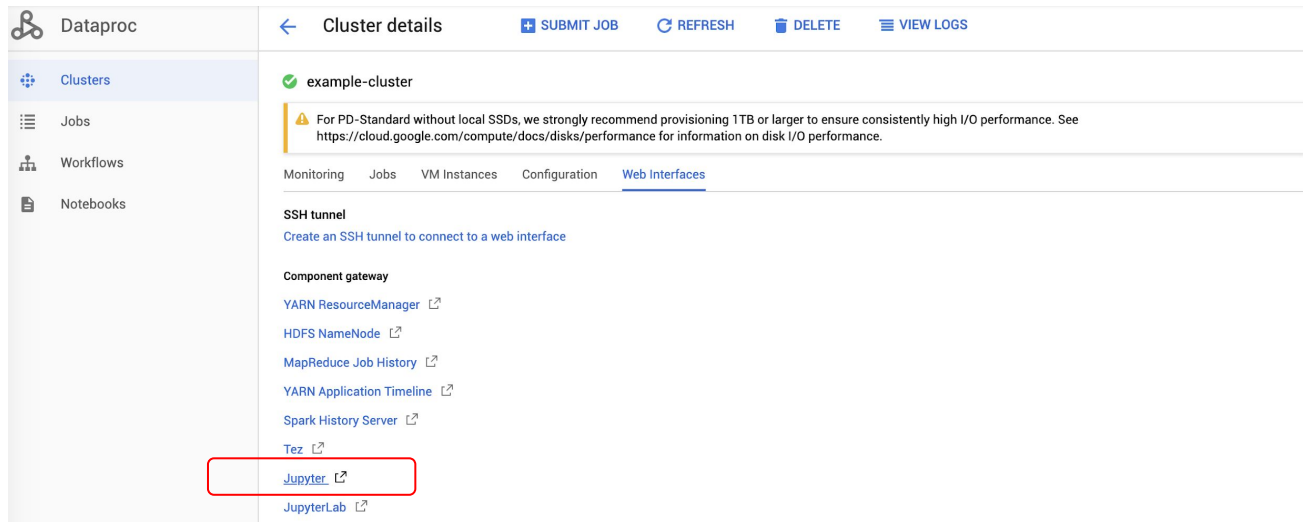
Cloud Storage  
bucket: where  
your jupyter  
notebooks are  
saved

Works like pip install <your  
package>



# Dataproc - Spark execution / submit jobs

- Jupyter notebook:



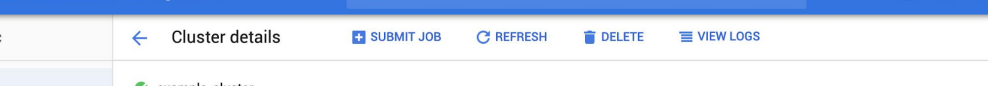
- Cloud SDK:

- `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



The screenshot shows the Google Cloud Platform Dataproc console. The left sidebar contains navigation links for Clusters, Jobs, Workflows, and Notebooks. The main content area is titled 'Cluster details' and shows information for 'example-cluster'. A warning message is displayed: '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.' Below this, there are tabs for Monitoring, Jobs, VM instances, Configuration, and Web interfaces. The 'Configuration' tab is active, showing a table with columns 'Name' and 'Role'. The table lists 'example-cluster-m' with the role 'Master'. A red box highlights the 'SSH' button next to the 'Master' role.

Google Cloud Platform

big-data-ta

Dataproc

Cluster details

SUBMIT JOB REFRESH DELETE VIEW LOGS

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

```

crouyang2@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
and 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.**