



# EECS E6893 Big Data Analytics

## Intro to Big Data Analytics on GCP

Rui Chu, [rc3414@columbia.edu](mailto:rc3414@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/>

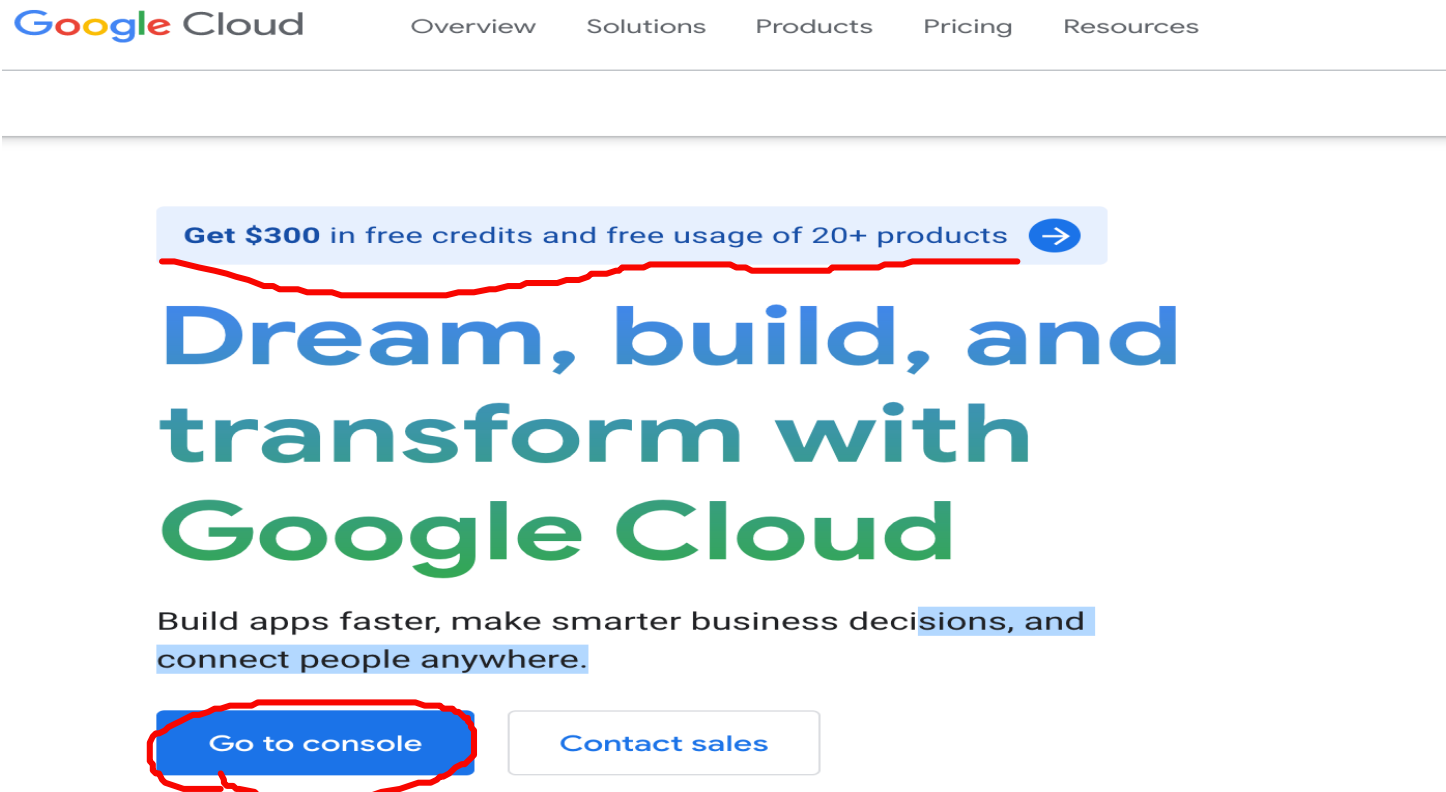


# Google Cloud Platform (GCP)



# 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





Start your Free Trial with \$300 in credit. Don't worry—you won't be charged if you run out of credits. [Learn more](#)

DISMISS

ACTIVATE



Google Cloud

Select a project ▼



Search Products, resources, docs (/)



R

Access support tools quickly

Find live and self-service support, docs, and tutorials in this menu

Close callout



# Welcome

Create or select a project to get started with Google Cloud. [Learn more about projects](#)

[Dashboard](#)

[Recommendations](#)

[+ Create a VM](#)

[+ Run a query](#)

Quick access ?

API APIs & Services

Cloud Storage

[View all products](#)

Compute Engine

Kubernetes Engine

## Google Cloud

### Welcome Rui Chu!

Create and manage your Google Cloud instances, disks, networks, and other resources in one place.

Country

United States ▼

Terms of Service

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

[AGREE AND CONTINUE](#)

# 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

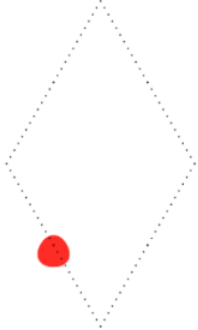
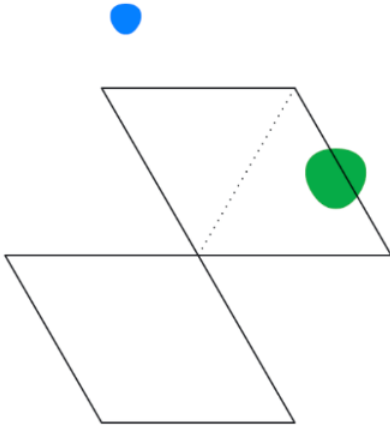
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

## \$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.





## 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.

 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
	<div>Card number is required</div>				
	Cardholder name				
	Cong Han				
	Billing address				

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

## Access to all Cloud Platform Products

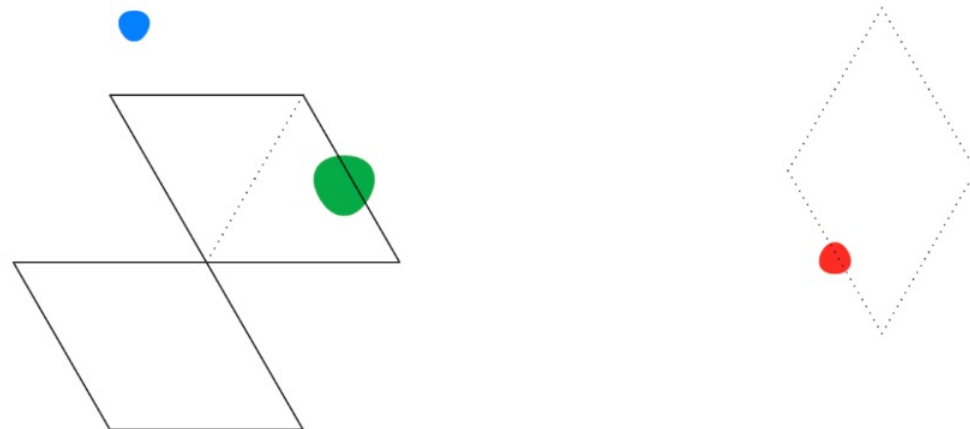
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

## \$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.

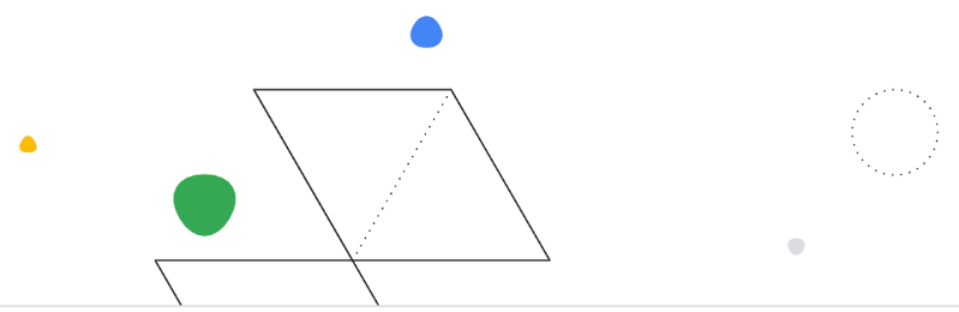


User preferences

Cloud profile

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

Welcome, Cong  
Get started with Google Cloud



Begin with the basics

Get up and running quickly by checking off common tasks

GO TO CHECKLIST

Setting up Google Cloud for scalable, production-ready enterprise workloads? Use the [Google Cloud setup checklist](#) designed for administrators.

What's covered

- Reviewing billing, credits, and projects
- Finding products and APIs
- Adding resources to a project
- Understanding and calculating pricing

Top products VIEW ALL

Compute products



Compute Engine  
Made by Google  
Scalable, high-performance virtual machines

Other popular compute options

[Kubernetes Engine](#)  
One-click Kubernetes clusters, managed by Google

[App Engine](#)  
A platform to build web and mobile apps that scale automatically



Free Trial and Free Tier | Google

Overview – Billing – My First Pr

+

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

Google Cloud Platform

Search products and resources

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

LEARN

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

Free Trial and Free Tier | Goog x Home - My First Project - Goog x

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

Google Cloud Platform My First Project Search products and resources

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

DASHBOARD ACTIVITY RECOMMENDATIONS CUSTOMIZE

### 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

Navigation menu



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

[MANAGE QUOTAS](#)

Project name \*

big data 6893



Project ID: big-data-6893-362015. It cannot be changed later. [EDIT](#)

Location \*

 No organization

[BROWSE](#)

Parent organization or folder


CREATE

CANCEL

DISMISS

ACTIVATE




Notifications



Create Project: big data 6893

[SELECT PROJECT](#)

Just now



[SEND FEEDBACK](#)

Free Trial and Free Tier | Goog x Home – big data 6893 – Googl x

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

APIs & Services

Support

IAM & Admin

Getting started

Compliance

Security

Anthos

COMPUTE

Compute Engine

Kubernetes Engine

VMware Engine

DASHBOARD

ACTIVITY

RECOMMENDATIONS

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

APIs

Requests (requests/sec)

No data is available for the selected time frame.

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

16

# 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



r

Contact Us

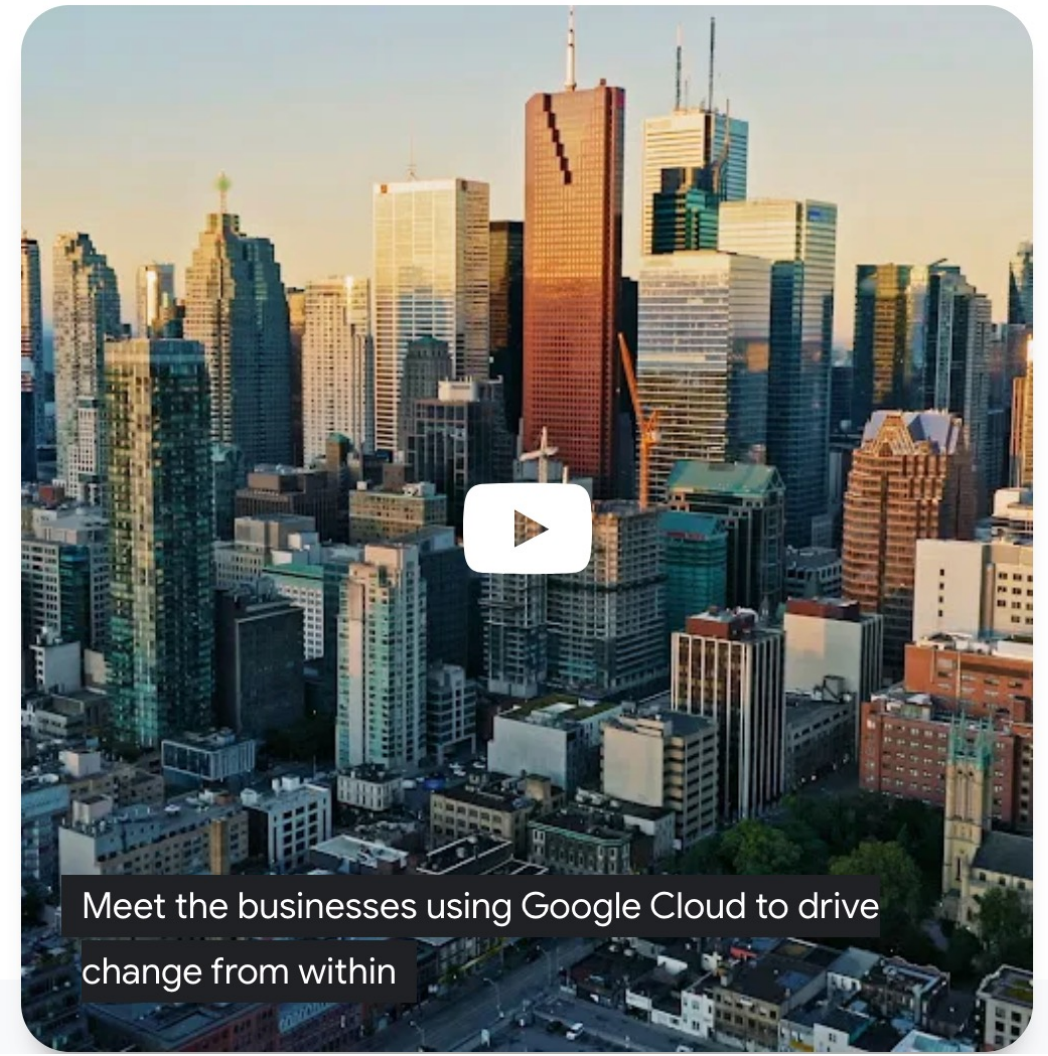
Get \$300 in free credits and free usage of 20+ products →

## 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





# Welcome

You're working in [big data 6893](#)

Project number: 943139336510



Project ID: big-data-6893-362015



[Dashboard](#)

[Recommendations](#)

[+ Create a VM](#)

[+ Run a query in BigQuery](#)

[+ Create a GKE cluster](#)

[+ Create a storage bucket](#)

## Quick access ?

big data 6893

Search – big data 6893

My First Project

API API/Service Details – APIs &...

My First Project

API Cloud Resource Manager A...

My First Project

Cloud profile – User prefere...

big data 6893

Cloud Storage Browser

My First Project

Language & region – User p...

big data 6893

API APIs & Services

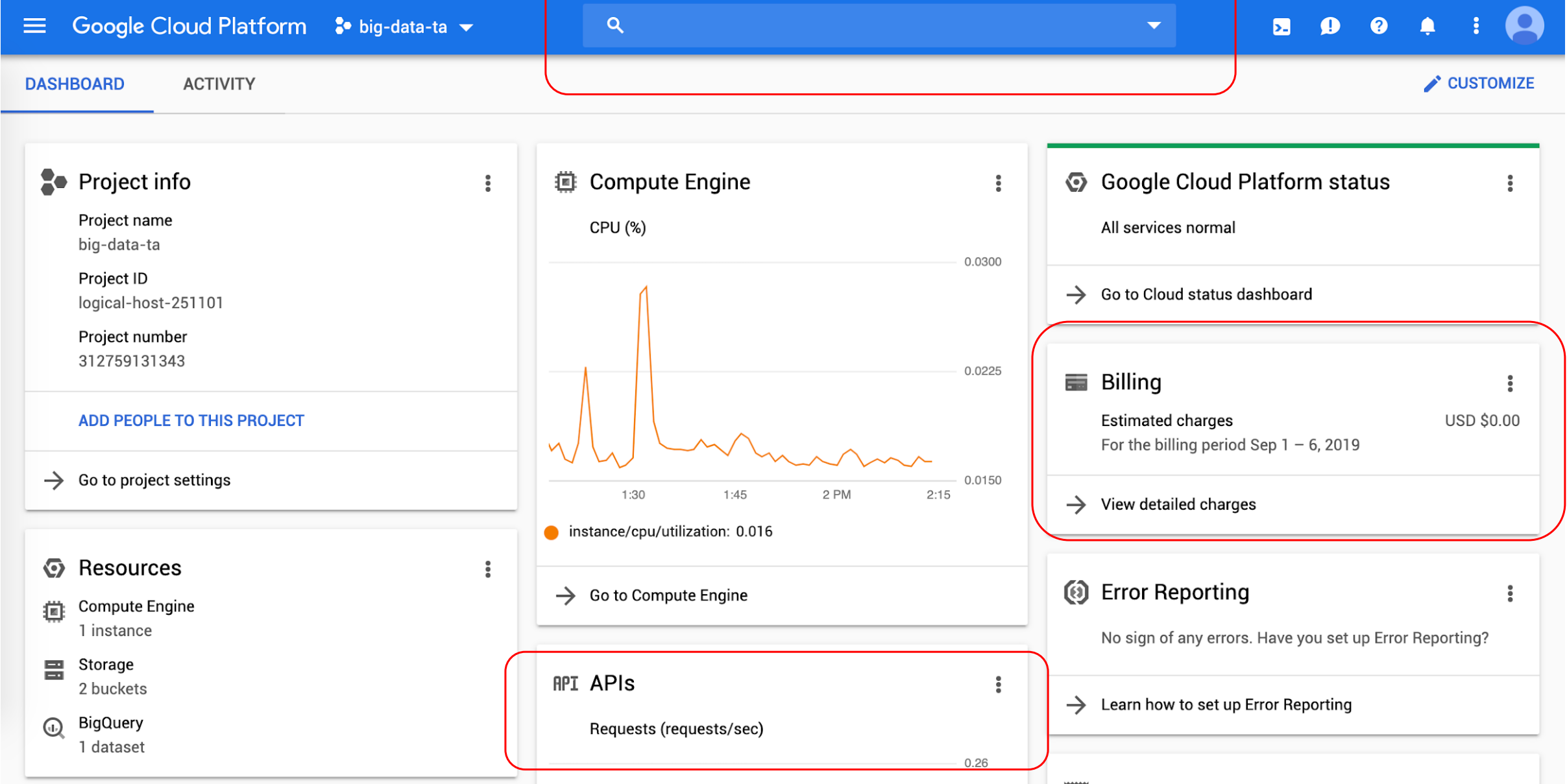
big data 6893

IAM & Admin

[View all products](#)

# GCP: console

Search for services here



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 | Goog x | Home – big data 6893 – Googl x | Quickstart: Getting started with x +

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

Contact Us

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-</a>	91.8 MB	5ef09ff44bbaadh8f5c9bc705ba2e320had87b

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

22

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

```
Choose the account you would like to use to perform operations for this configuration:  
[1] jerry.r.chu@gmail.com  
[2] rc3414@columbia.edu  
[3] Log in with a new account  
Please enter your numeric choice:
```

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

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>

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=m17Vnevs9DKqLLSRDyW2sFDgcRgYBW&access\\_type=offline&code\\_challenge=Y\\_hSRd9TakgmBNRj1qk1JghIlcIum9mBqS9jjdk3KXI&code\\_challenge\\_method=S256](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=m17Vnevs9DKqLLSRDyW2sFDgcRgYBW&access_type=offline&code_challenge=Y_hSRd9TakgmBNRj1qk1JghIlcIum9mBqS9jjdk3KXI&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@Congs-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@Congs-MacBook-Pro:~$ gcloud info
Google Cloud SDK [356.0.0]

Platform: [Mac OS X, x86_64] uname_result(system='Darwin', node='Congs-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')
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:/anaconda3/bin:/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@Congs-MacBook-Pro:~$ █

```



# GCP: Cloud Shell

The screenshot shows the Google Cloud Platform dashboard for 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 rounded rectangle highlights the 'Activate Cloud Shell' button, which is represented by a terminal icon and a question mark icon. Below the navigation bar, the dashboard is divided into three main sections: 'Project info', 'API APIs', and 'Google Cloud Platform status'. The 'Project info' section on the left lists the project name 'big-data-ta', project ID 'logical-host-251101', and project number '312759131343', with a link to 'Go to project settings'. The 'API APIs' section in the center shows a line graph for 'Requests (requests/sec)' with a y-axis ranging from 0.5 to 2.5. The 'Google Cloud Platform status' section on the right indicates 'All services normal' and provides a link to 'Go to Cloud status dashboard'. Below this, the 'Billing' section shows '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 :). The most useful way to complete the HW0

# GCP: Cloud Shell

The screenshot shows the Google Cloud Platform (GCP) dashboard for the project 'big-data-ta'. The dashboard includes sections for Project info, API APIs, and Google Cloud Platform status. At the bottom, the Cloud Shell terminal is open, showing the command 'ls' and its output 'README-cloudshell.txt'. A red arrow points from the filename in the terminal to the text below. In the top right of the dashboard, a 'Launch code editor BETA' button is circled in red.

Google Cloud Platform big-data-ta

DASHBOARD ACTIVITY CUSTOMIZE

**Project info**

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

**API APIs**

Requests (requests/sec)

2.5

2.0

**Google Cloud Platform status**

All services normal

→ Go to Cloud status dashboard

cloudshell x +

```
frouyang2@cloudshell:~$ ls
hw0 README-cloudshell.txt
frouyang2@cloudshell:~$
```

Launch code editor BETA

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

- FROUYANG2
  - hw0
    - wordcount.py

```
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$
```



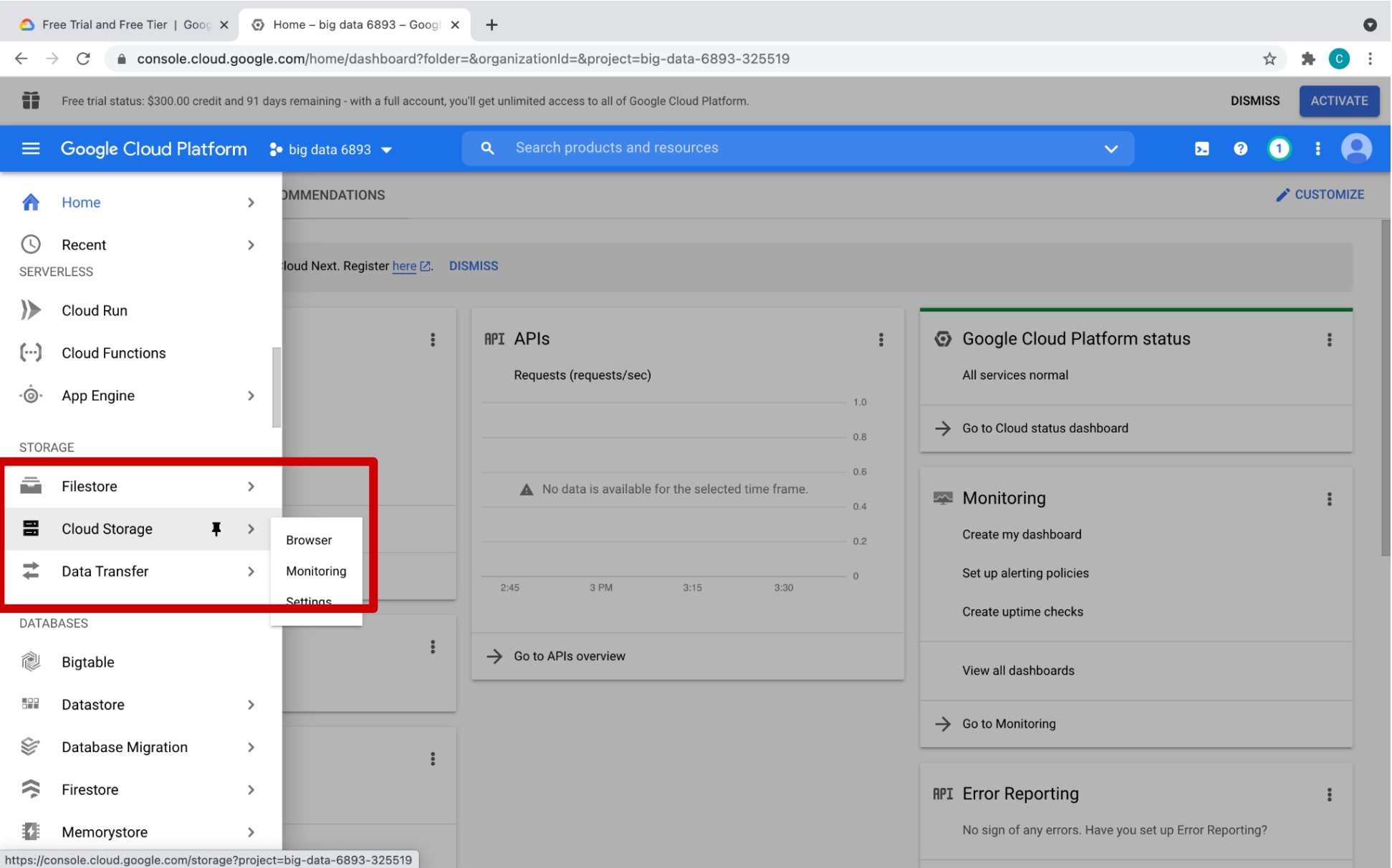
# Cloud Storage

# 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
requesterpays Enable or disable requester pays for one or more buckets
retention     Provides utilities to interact with Retention Policy feature.
rewrite       Rewrite objects
rm            Remove objects
rpo           Configure replication
rsync         Synchronize content of two buckets/directories
setmeta       Set metadata on already uploaded objects
signurl       Create a signed URL
stat          Display object status
test          Run gsutil unit/integration tests (for developers)
ubla          Configure Uniform bucket-level access
update        Update to the latest gsutil release
```

# Cloud Storage



# Cloud Storage

Free Trial and Free Tier | Google

Browser – Cloud Storage – big

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

Google Cloud Platform

big data 6893

Search products and resources

Cloud Storage

Browser

Monitoring

Settings

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

LEARN Home

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](#)

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

33

# 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

Browser

Monitoring

Settings

Create a bucket

✓

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

Monthly cost estimate

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

Storage and retrieval

Storage size

GB

\$0.026 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 (\$) ▼

Name your own bucket

# 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 Platformbig data 6893Search products and resources

Cloud Storage

Browser

Monitoring

Settings

Release Notes

<1

Create a bucket

Name your bucket

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

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

Browser

Monitoring

Settings

Create a bucket

✓

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

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 Platformbig data 6893Search products and resources

Cloud Storage

Browser

Monitoring

Settings

Release Notes

<|

←

Create a bucket

✓

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

×

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.

DISM

Google Cloud Platformbig data 6893Search products and resources

Cloud Storage

Browser

Monitoring

Settings

Create a bucket

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](#)

ADD LABEL

CREATE

CANCEL

Monthly cost estimate

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

Storage and retrieval

Storage size  
\$0.020 per GB-month

Data retrieval size  
Free

Operations

Class A operations  
\$0.005 per 1,000 ops

Class B operations  
\$0.0004 per 1,000 ops

Availability SLA: 99.9%

Monthly cost: \$0.00

Currency: US Dollar (\$)

### 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

Your data is always protected with Cloud Storage but you can also choose from these additional data protection options to prevent data loss. Note that object versioning and retention policies cannot be used together.

### Protection tools

- ☒ None
- ☐ Object versioning (best for data recovery)  
For restoring deleted or overwritten objects. To minimize the cost of storing versions, we recommend limiting the number of noncurrent versions per object and scheduling them to expire after a number of days. [Learn more](#)
- ☐ Retention policy (best for compliance)  
For preventing the deletion or modification of the bucket's objects for a specified minimum duration of time after being uploaded. [Learn more](#)

### Data encryption

- ☒ Google-managed encryption key  
No configuration required
- ☐ Customer-managed encryption key (CMEK)  
Manage via Google Cloud Key Management Service

SHOW LESS

# 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

Cloud Storage

Browser

Monitoring

Settings

Release Notes

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 ?	Reten	
<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	—	⬇ ⋮

dataset provided in HW0 details

Click on the uploaded dataset file

# Cloud Storage

The screenshot shows the Google Cloud Platform console interface. The top navigation bar includes the Google Cloud Platform logo, the project name 'big data 6893', and a search bar. 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 file 'data\_citibike\_stations.csv' in the 'big-data-6893' bucket. The 'Overview' section contains a table with the following information:

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	gs://big-data-6893/data/data_citibike_stations.csv

The 'Authenticated URI' and 'gsutil URI' rows are highlighted with a red box. Below the 'Overview' section, the 'Permissions' and 'Protection' sections are visible, showing that the object is not public and has no hold status or retention policy.

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>

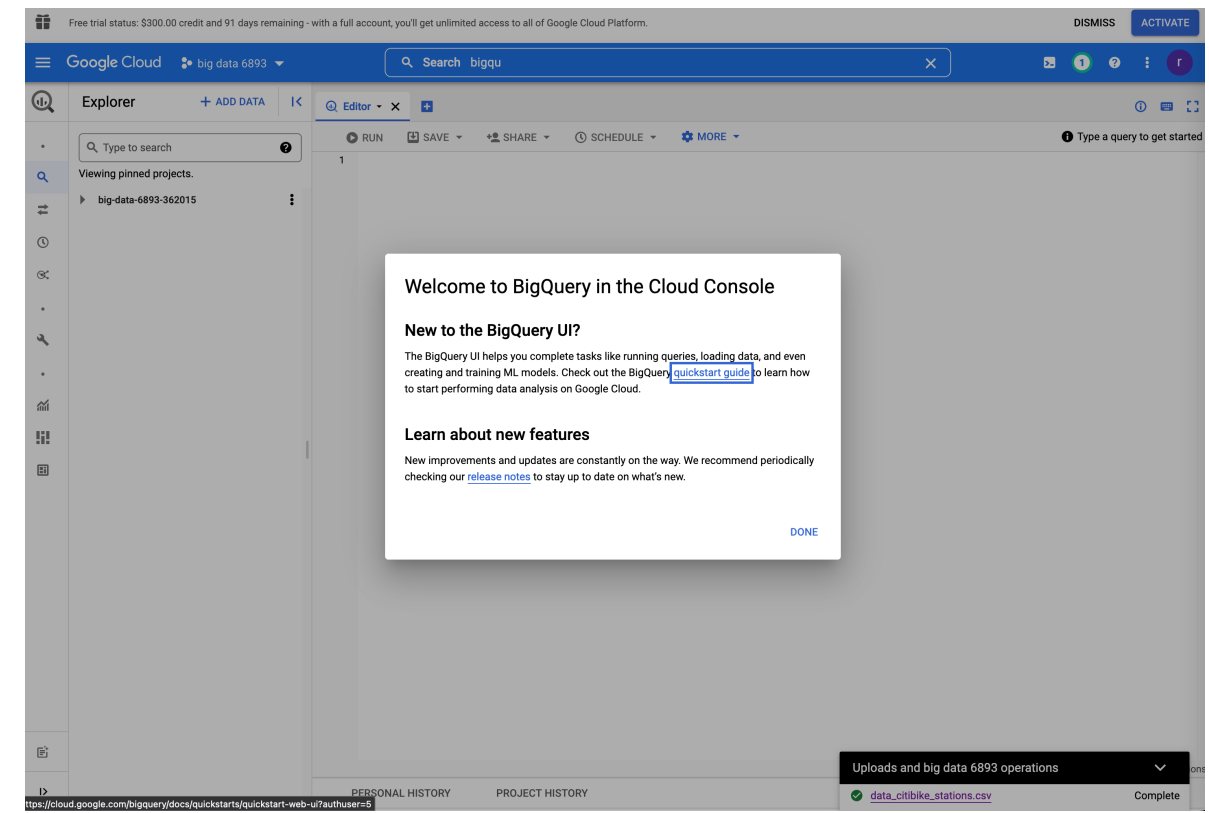


# BigQuery

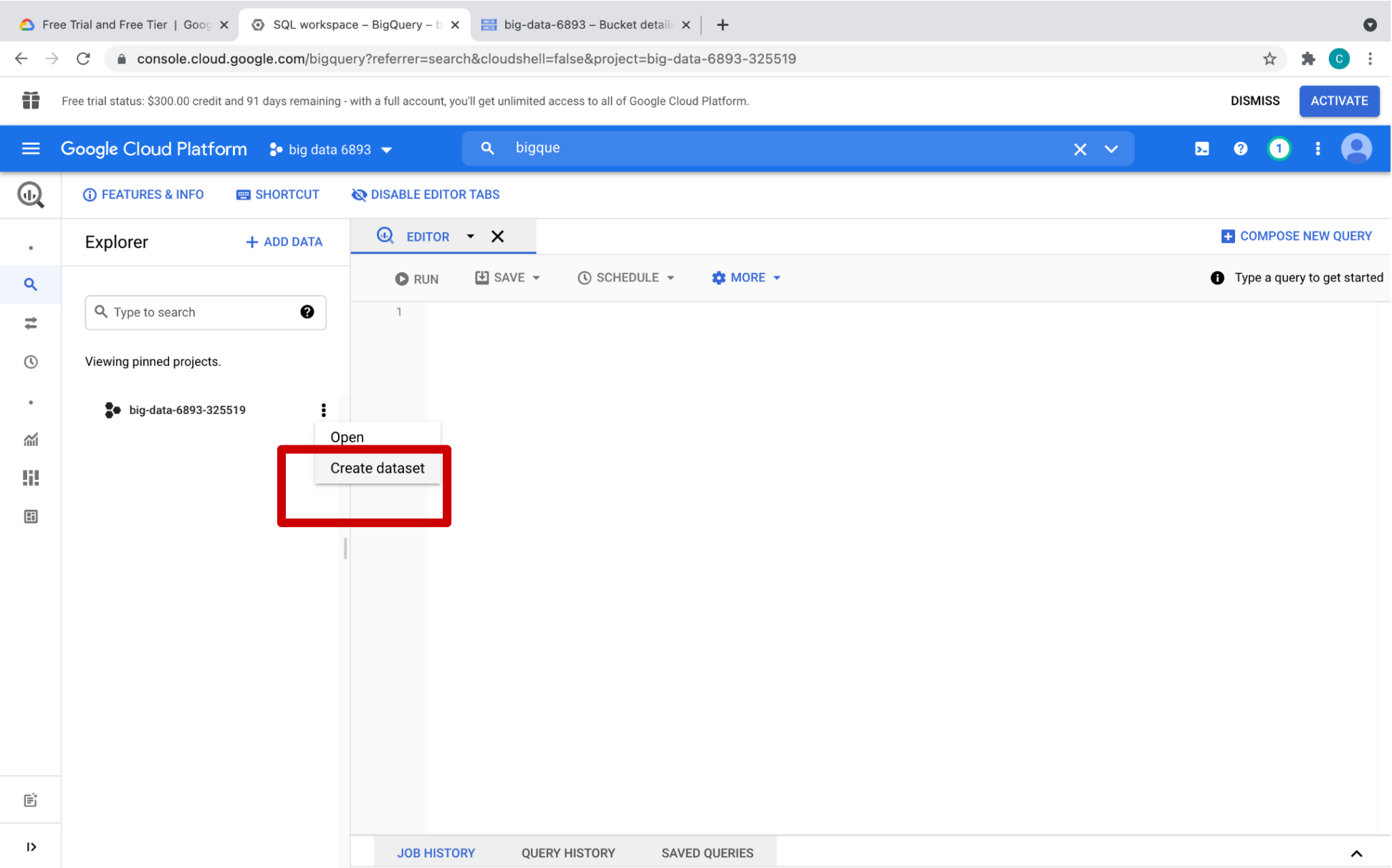
# BigQuery

- 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





# BigQuery

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 Cloudbig data 6893

Searchbigqu

Explorer

+ ADD DATA

<

Type to search

?

Viewing pinned projects.

big-data-6893-362015

RUN

SAVE

SHARE

SCHEDULE

MORE

1

PERSONAL HISTORY

PROJECT HISTORY

Create dataset

Project ID

big-data-6893-362015

CHANGE

Dataset ID \*

dataset1

Letters, numbers, and underscores allowed

Data location

?

Default table expiration

Enable table expiration

?

Default maximum table age

Days

Advanced options

^

Encryption

?

Google-managed encryption key

No configuration required

Customer-managed encryption key (CMEK)

Manage via Google Cloud Key Management Service

Default Collation

Enable default collation

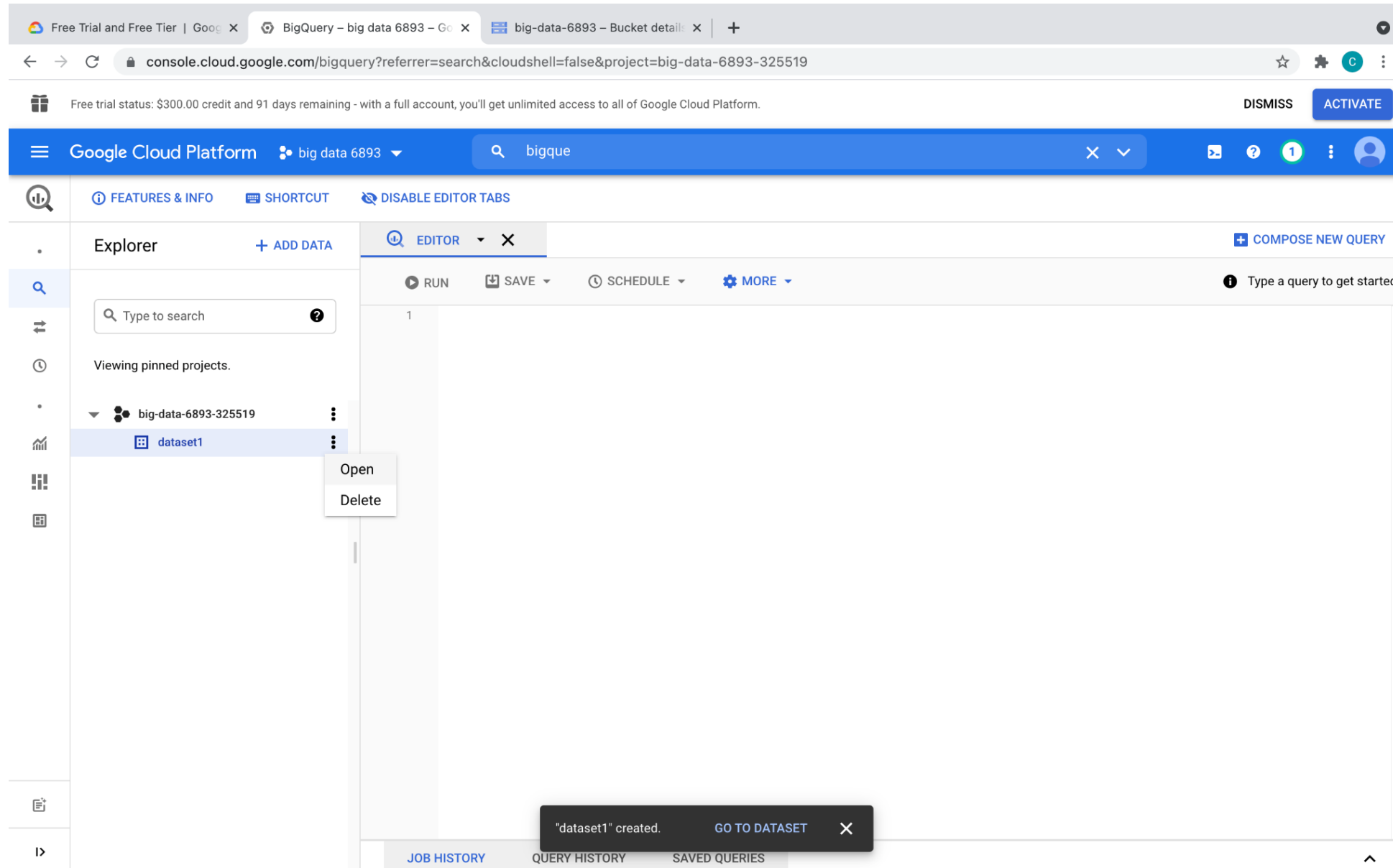
?

Default Collation

CREATE DATASET

CANCEL

# BigQuery



# BigQuery

Free Trial and Free Tier | Google

SQL workspace - BigQuery - b

big-data-6893 - Bucket detail

+

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

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

FEATURES & INFO

SHORTCUT

DISABLE EDITOR TABS

Explorer + ADD DATA

Viewing pinned projects.

big-data-6893-325519

dataset1

EDITOR DATASET1 + COMPOSE NEW QUERY

big-data-6893-325519:dataset1 + Create table SHARE DATASET AUTHORIZE ROUTINES COPY DATASET DELETE DATASET

Description

None

Labels

None

Dataset info

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

"dataset1" created. GO TO DATASET

JOB HISTORY QUERY HISTORY SAVED QUERIES

# BigQuery

Free Trial and Free Tier | Google Cloud

SQL workspace - BigQuery - 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...

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

Google Cloud Platform | big data 6893 | bigque

FEATURES & INFO | SHORTCUT | DISABLE EDITOR TABS

Explorer | + ADD DATA

big-data-6893-325519:dataset1

Description | Dataset info

Dataset ID: big-data-6893-325519:dataset1

Created: Sep 9, 2023

Default table expiration: Never

Last modified: Sep 9, 2023

Data location: US

Create table

Source

Create table from

Empty table

Google Cloud Storage

Destination

Search for a

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)

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

"dataset1" created.

GO TO DATASET

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

Google Cloud Platform | big data 6893 | bigque

FEATURES & INFO | SHORTCUT | DISABLE EDITOR TABS

Explorer | + ADD DATA

big-data-6893-325519:dataset1

Description | Dataset info

Dataset ID: big-data-6893-325519:dataset1

Created: Sep 9, 2023

Default table expiration: Never

Last modified: Sep 9, 2023

Data location: US

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 \*

big-data-6893-362015

Dataset \*

dataset1

Table \*

bike\_data

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

Table type

Native table

Schema

Source file defines the schema.

Partition and cluster settings

Partitioning

No partitioning

CREATE TABLE

CANCEL

Choose a file

6893\_data\_22ta

data\_citibike\_stations.csv

Filename

data\_citibike\_stations.csv

SELECT

CANCEL

# BigQuery

Free Trial and Free Tier | Google

SQL workspace – BigQuery – b

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...

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

Google Cloud Platformbig data 6893bigque

Explorer+ ADD DATA

big-data-6893-325519:dataset1

Dataset info

Dataset ID	big-data-6893-325519
Created	Sep 9, 2023
Default table expiration	Never
Last modified	Sep 9, 2023
Data location	US

JOB HISTORYQUERY

Create table

Source

Create table from:Google Cloud Storage

Select file from GCS bucket:big-data-6893/data/data\_citibike\_stations.csv

File format:CSV

Source Data Partitioning

Destination

Search for a projectEnter a project name

Project namebig data 6893

Dataset namedataset1

Table typeNative table

Table namebike\_data

Schema

Auto detect

Schema and input parameters

Schema will be automatically generated.

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)

Advanced options

Create tableCancel

# 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=!1m5!1m4!1m3!1sbig-data-6893-325519!2sbquxj

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

Google Cloud Platform big data 6893 bigque

FEATURES & INFO SHORTCUT DISABLE EDITOR TABS

Explorer + ADD DATA

Type to search

Viewing pinned projects.

big-data-6893-325519 dataset1 bike\_data

UNSAVE... 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	eightd_has_key_dispenser	num_bikes_availab
1	3206	Hilltop	JC019	40.7311689	-74.0575736	70	KEY,CREDITCARD	26	false	
2	3195	Sip Ave	JC056	40.73089709786179	-74.06391263008118	70	KEY,CREDITCARD	34	false	
3	3640	Journal Square	JC103	40.73367	-74.0625	70	KEY,CREDITCARD	18	false	
4	3481	York St	JC096	40.71649	-74.04105	70	KEY,CREDITCARD	22	false	

JOB HISTORY QUERY HISTORY SAVED QUERIES

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

Google Cloud big data 6893 Search bigqu

Explorer + ADD DATA

Type to search

Viewing pinned projects.

big-data-6893-362015 dataset1 bike\_data

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				
eightd_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				
eightd_has_available_keys	BOOLEAN	NULLABLE				
last_reported	TIMESTAMP	NULLABLE				

EDIT SCHEMA VIEW ROW ACCESS POLICIES

"bike\_data" created. GO TO TABLE

PERSONAL HISTORY PROJECT HISTORY

REFRESH

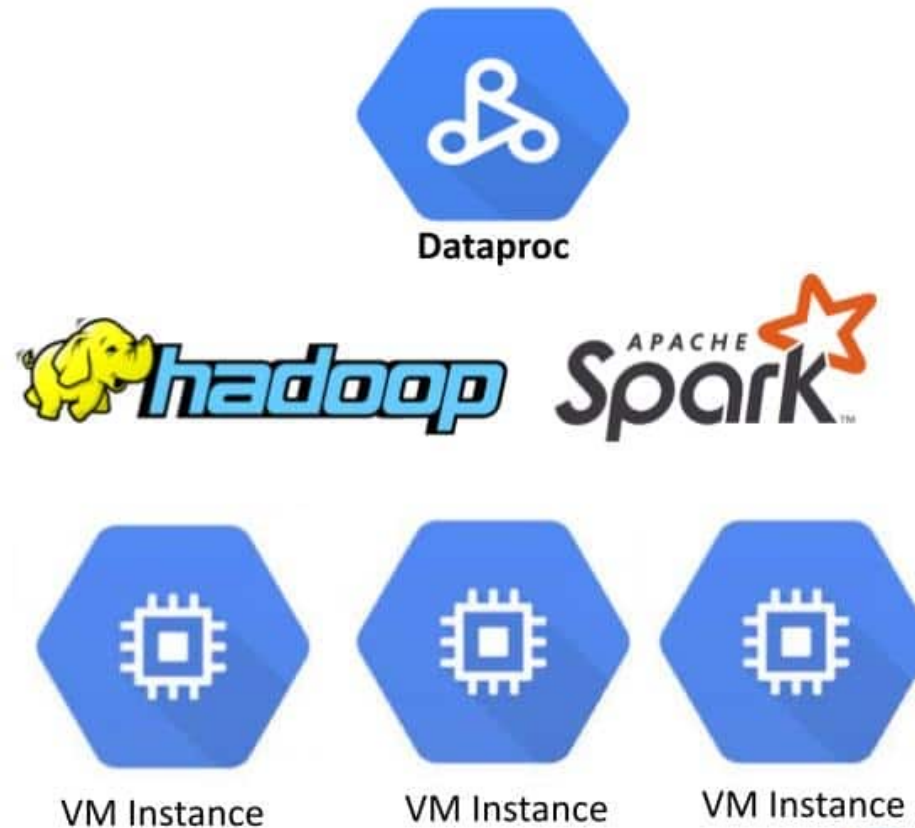


# Dataproc

# Dataproc

## What is dataproc?

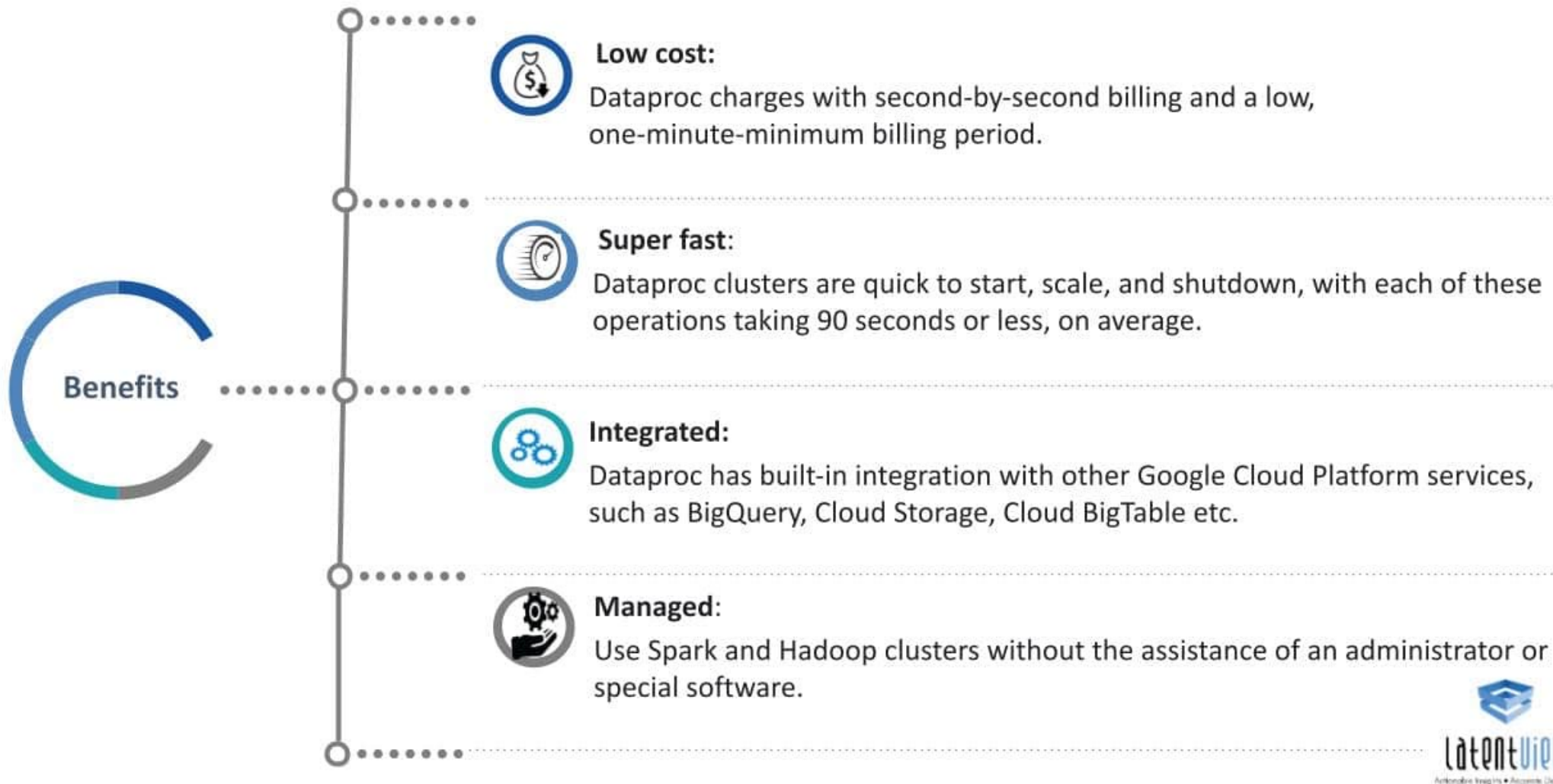
- Google Cloud Dataproc is a managed service for running **Apache Hadoop and Spark jobs**.
- Dataproc uses **Compute Engine instances** under the hood, but it takes care of the management details.
- Includes **Hadoop, Spark, Hive and Pig**.
- **Ideal for moving** existing code to GCP





# Dataproc

## Why dataproc?



# Dataproc

search cloud data proc  
click on the API link

Free Trial and Free Tier | Google | SQL workspace – BigQuery – b | big-data-6893 – Bucket detail: | Cloud Dataproc API – Marketpl | +

← → ↺ console.cloud.google.com/marketplace/product/google/dataproc.googleapis.com?returnUrl=%2Fdataproc%2Fclusters%3Fcloudshell%3Dfalse%26project%3Dbig-data-689... ☆ ⚙️ 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 🔍 📧 ? 1 ⋮ 👤

←

**Cloud Dataproc API**  
Google Enterprise API  
Manages Hadoop-based clusters and jobs on Google Cloud Platform.  
**ENABLE** TRY THIS API 🔗

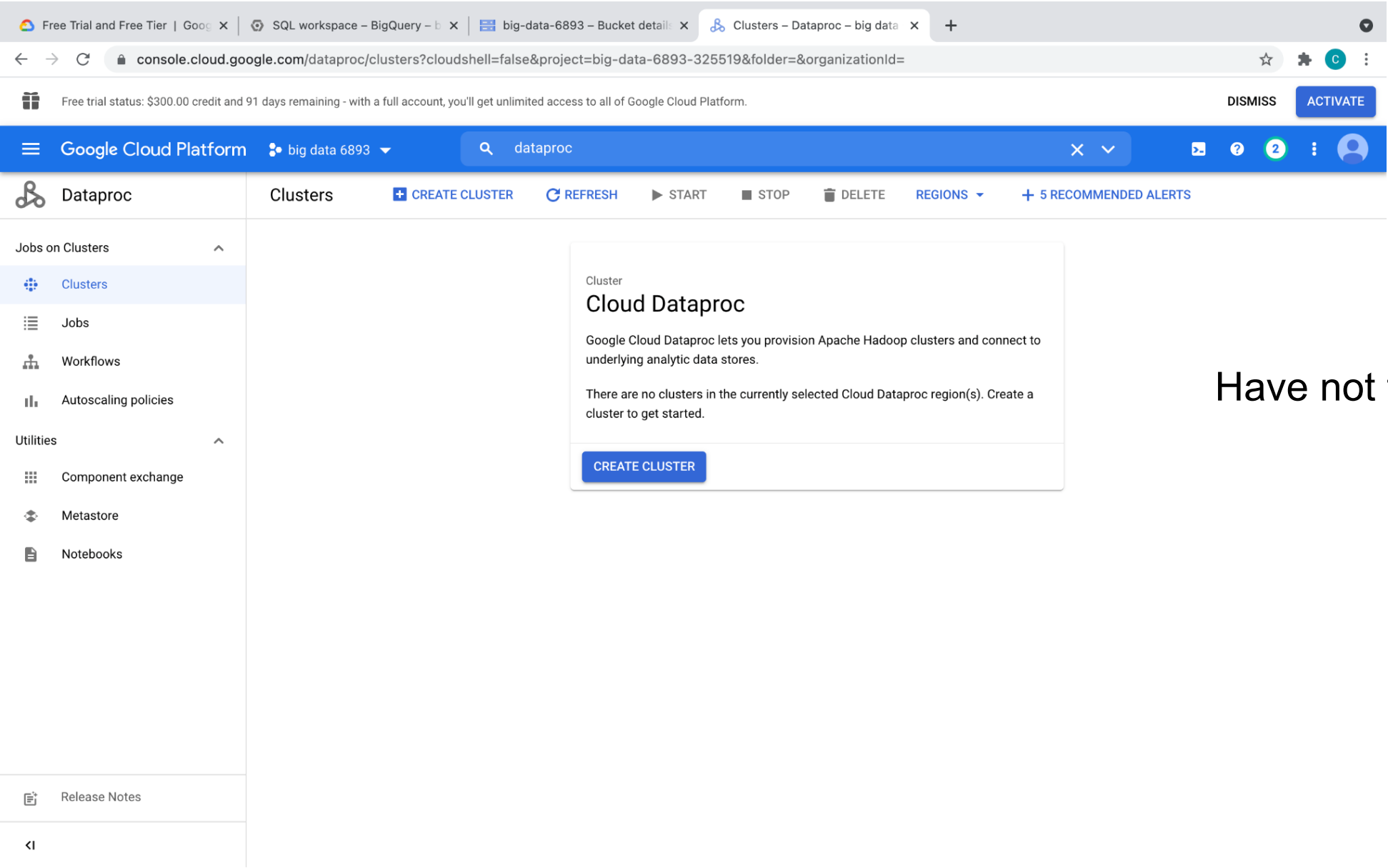
OVERVIEW DOCUMENTATION

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



Go to Cloud Dataproc

Have not tested GKE, should also be OK

# 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
- Customize cluster (optional)

Add cluster properties, features, and actions.

Manage security (optional)

Change access, encryption, and security settings.
- CREATE

CANCEL
- EQUIVALENT COMMAND LINE

▼

Release Notes

create cluster with Jupyter

## 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
```

# Dataprocc - Cloud SDK

Cluster creation (using Cloud SDK):

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

# Dataprocc - 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
SparkId702abe1n1r/1.1, (http/1.1)j10.0.0.0.0
Job [3f9861f7e3744a5580068001cdf48bf9] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-east1-881004012112-ixdi0md0/goog
le-cloud-dataproc-metainfo/7ff01079-3cec-47b3-b2f4-ba88665d16e1/jobs/3f9861f7e37
44a5580068001cdf48bf9/
driverOutputResourceUri: gs://dataproc-staging-us-east1-881004012112-ixdi0md0/go
ogle-cloud-dataproc-metainfo/7ff01079-3cec-47b3-b2f4-ba88665d16e1/jobs/3f9861f7e
3744a5580068001cdf48bf9/driveroutput
jobUuid: e5839c28-799f-3591-8dd8-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

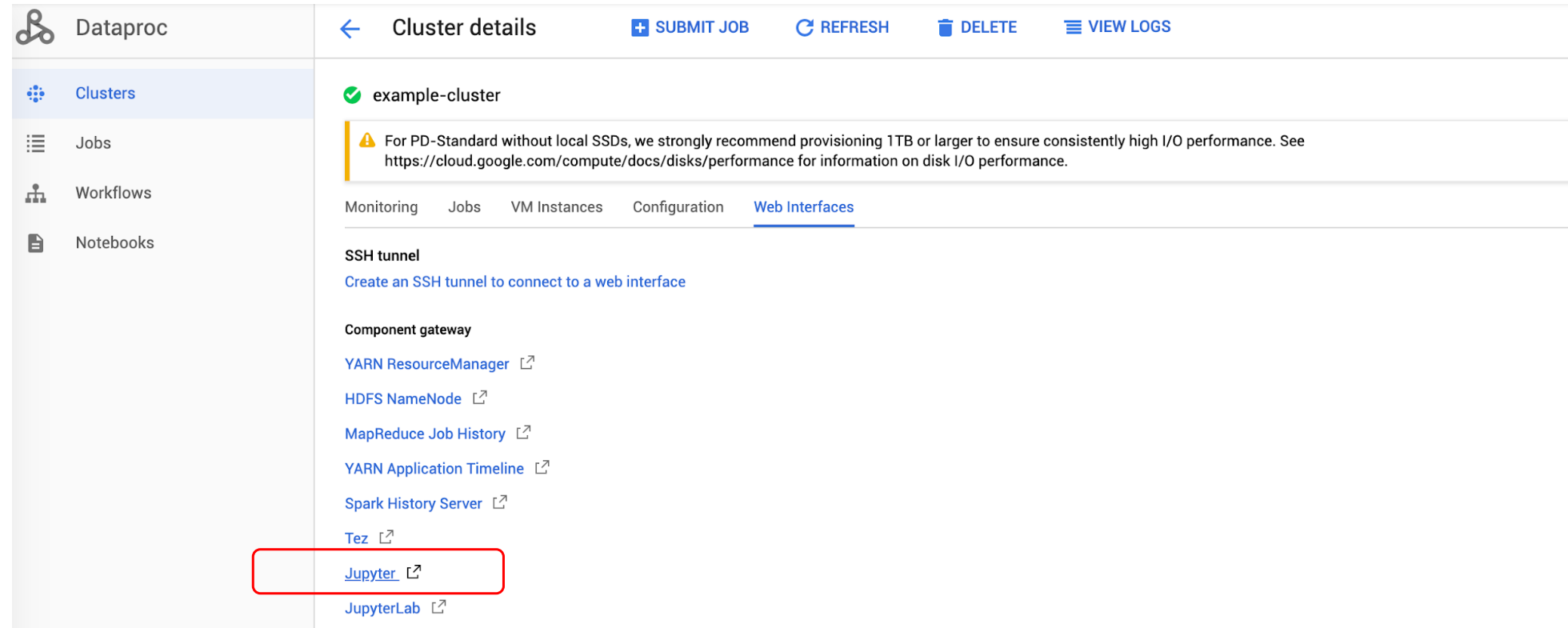
```
(base) conghan@Cong's-MacBook-Pro:~$ gcloud beta dataproc clusters create example-cluster --region=us-east1 --optional-components=ANACONDA,JUPYTER --image-version=1.3 --enable-component-gateway --bucket big-data-6893 --project big-data-6893-325519 --single-node --metadata 'PIP_PACKAGES=graphframes==0.6'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



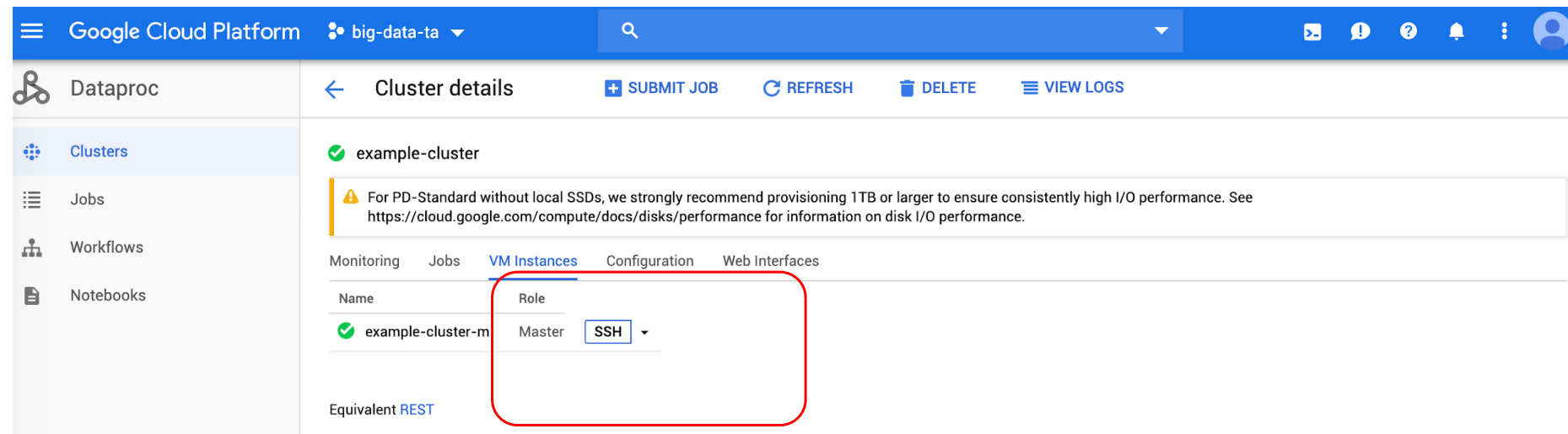
- Cloud SDK:

- `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
and so jobs will be scheduled in FIFO order. To use fair scheduling, configure pools in fairscheduler.xml or set spa
rk.scheduler.allocation.file to a file that contains the configuration.
Welcome to

  _ _ _ _ _
 _/ _ _ _ \_ Spark version 2.3.3
/_/_/_/_/_

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

# HW0

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

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