EECS E6893 Big Data Analytics
Intro to Big Data Analytics on GCP

Qingcheng Yu, qy2281@columbia.edu

Agenda

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

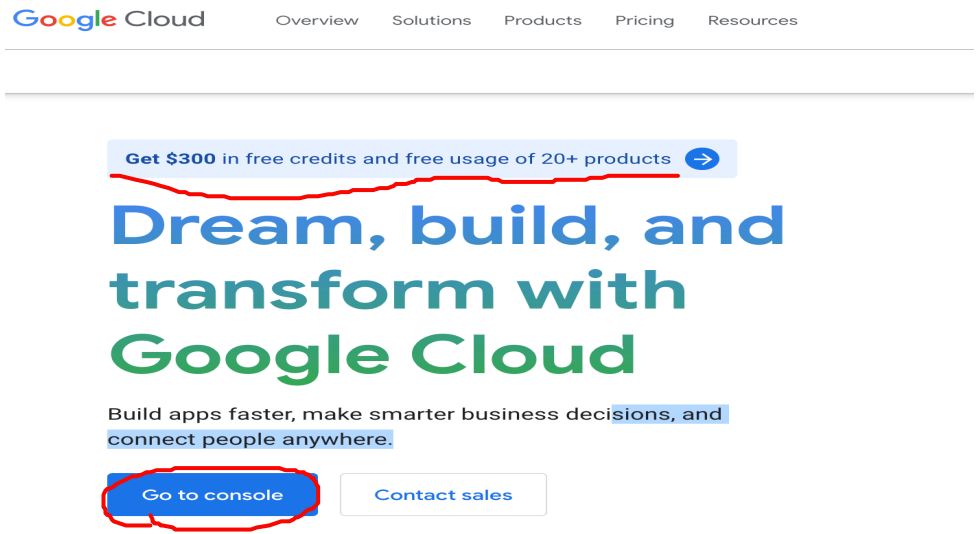
GCP

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

Google Cloud Platform (GCP)

GCP Setup

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



Google Cloud Overview Solutions Products Pricing Resources Contact Us Docs Support English Sign in Start free

Build what's next. Better software. Faster.

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

Get started for free Contact sales

Start running workloads for free

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

Built by developers, for developers

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

Estimate your costs

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

Try Google Cloud for free

Step 1 of 2 Account Information

Yu Jacky
jackyyu2021111@gmail.com

SWITCH ACCOUNT

Country

United States

What best describes your organization or needs?

Please select
Other

Terms of Service

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

Required to continue

CONTINUE

Access to all Google Cloud products

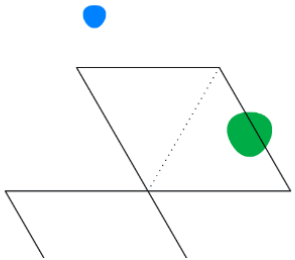
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

$300 credit for free

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

No autocharge after free trial ends

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



Try Google Cloud for free

Step 2 of 2 Payment Information Verification

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

Payments profile

Create new payments profile +

SUBMIT

Access to all Google Cloud products

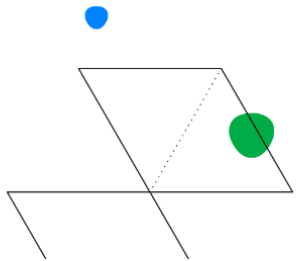
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

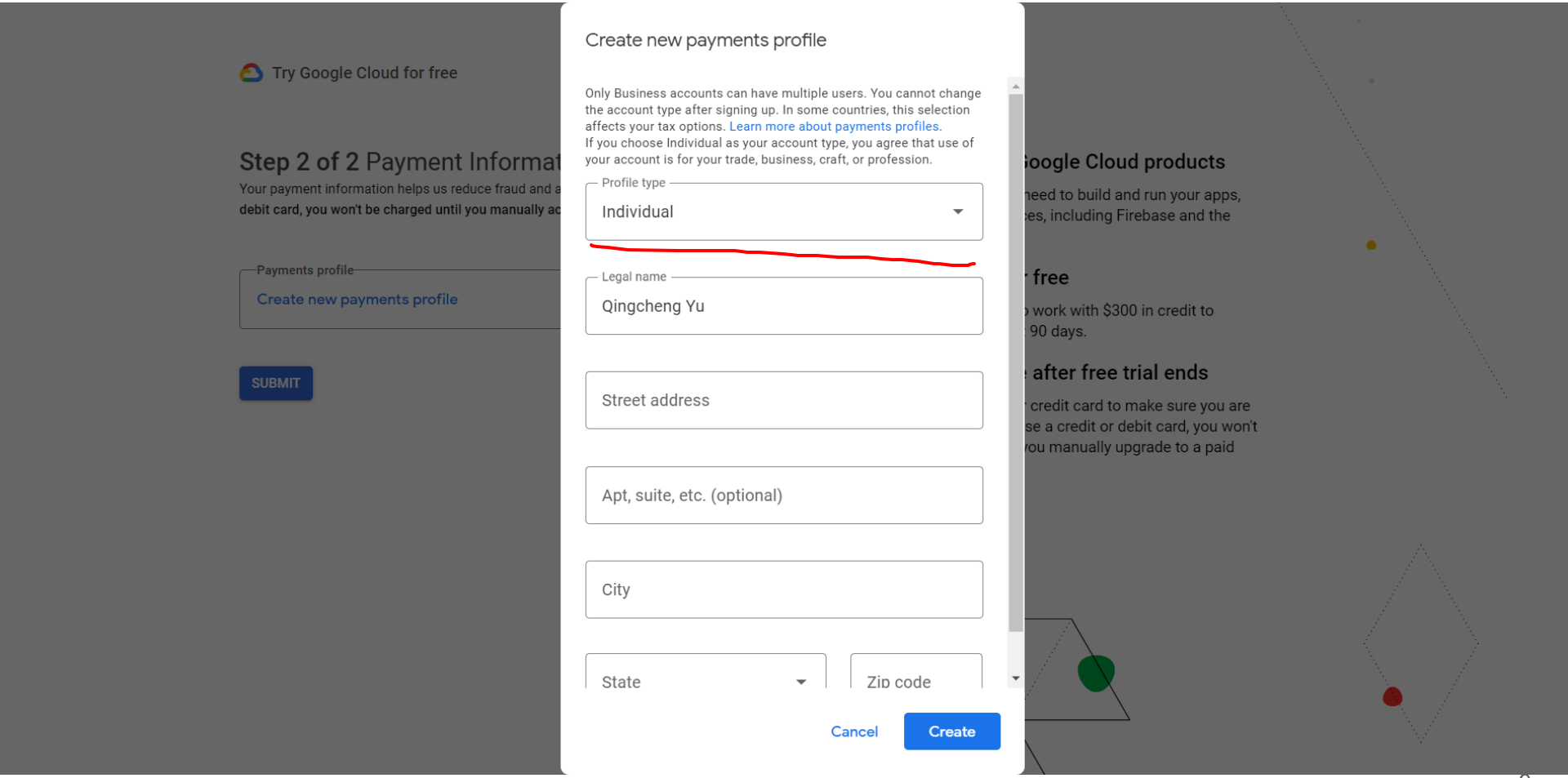
$300 credit for free

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

No autocharge after free trial ends

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




Try Google Cloud for free
Step 2 of 2 Payment Informat
Your payment information helps us reduce fraud and a
debit card, you won't be charged until you manually ac
Payments profile
Create new payments profile
SUBMIT
Create new payments profile
Only Business accounts can have multiple users. You cannot change the account type after signing up. In some countries, this selection affects your tax options. Learn more about payments profiles.
If you choose Individual as your account type, you agree that use of your account is for your trade, business, craft, or profession.
Profile type
Individual
Legal name
Qingcheng Yu
Street address
Apt, suite, etc. (optional)
City
State
Zip code
Cancel
Create
oogle Cloud products
eed to build and run your apps,
es, including Firebase and the
free
work with $300 in credit to
90 days.
after free trial ends
credit card to make sure you are
se a credit or debit card, you won't
ou manually upgrade to a paid

Try Google Cloud for free

Step 2 of 2 Payment Information Verification

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

Payments profile

Qingcheng Yu
Individual • United States • ID: 6169-5513-9110 Change

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

Payment method

Add payment method +

SUBMIT

Access to all Google Cloud products

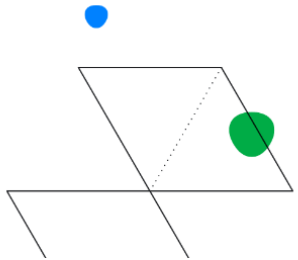
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

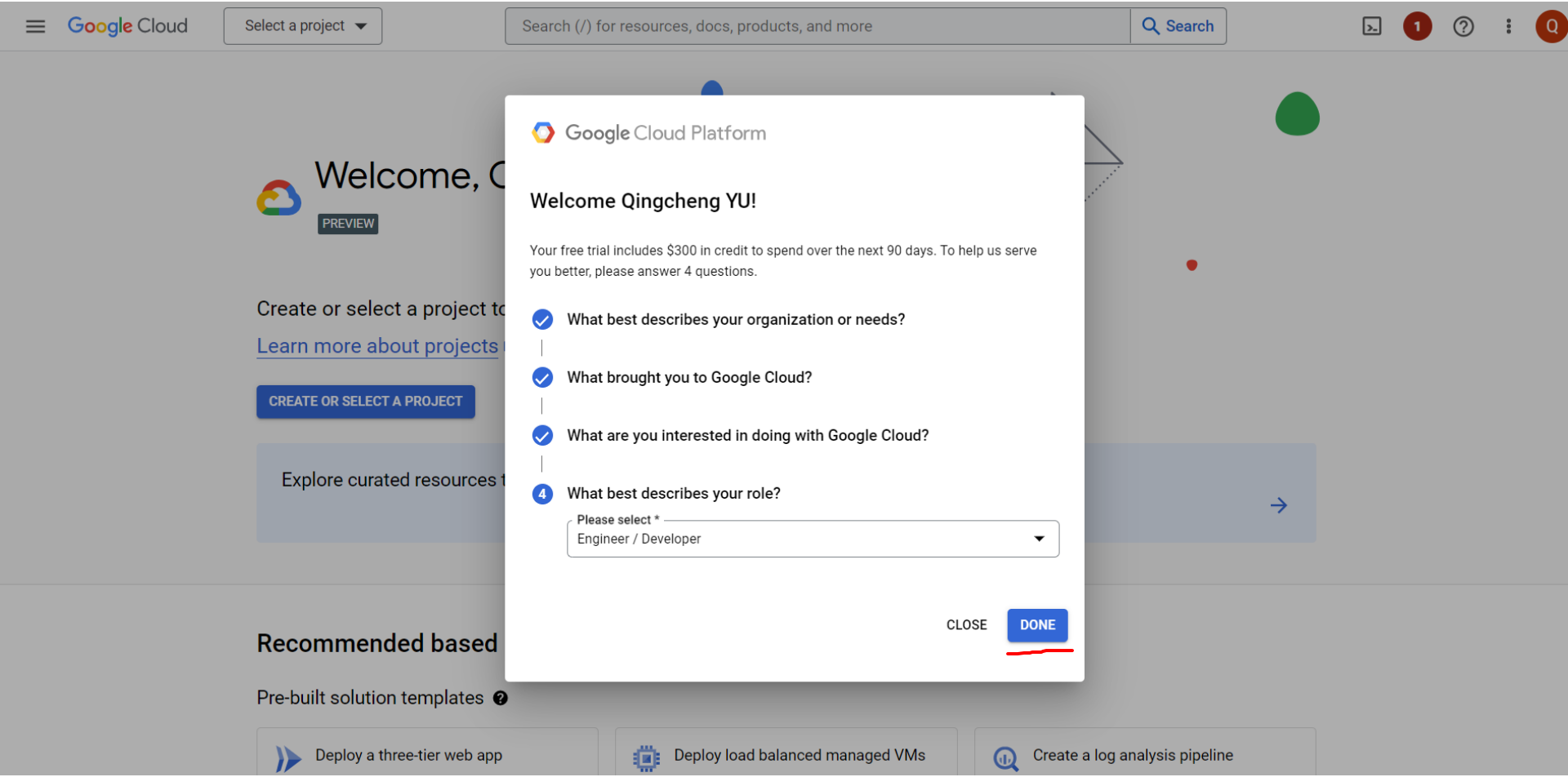
$300 credit for free

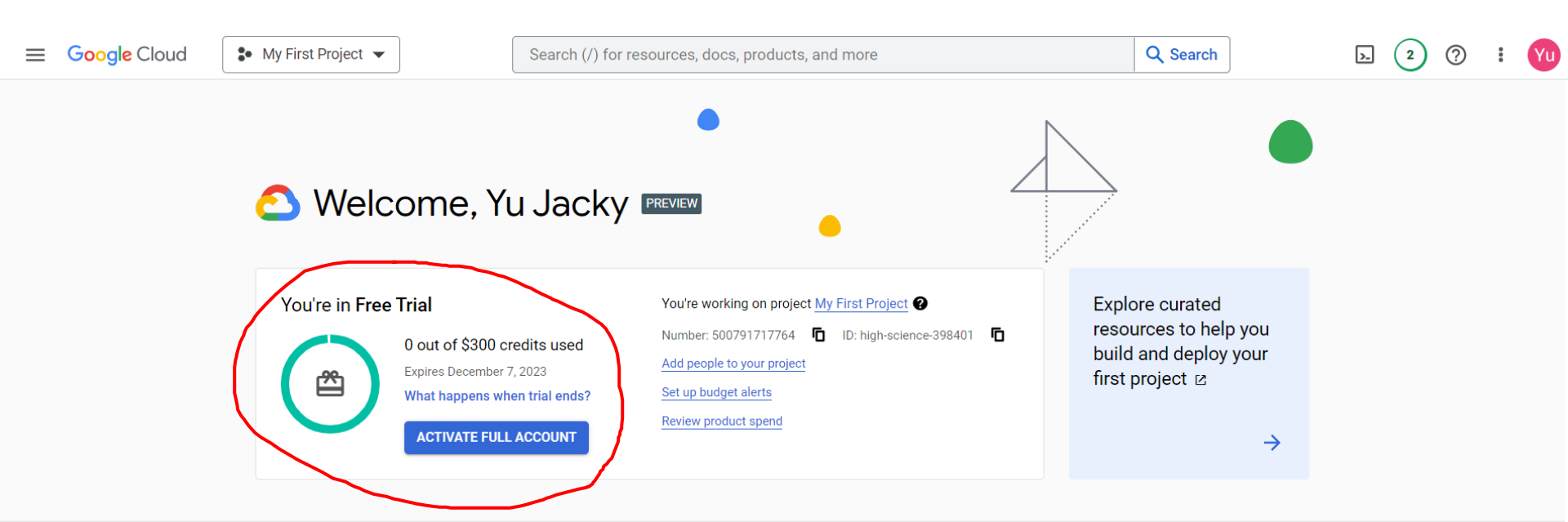
Put Google Cloud to work with $300 in credit to spend over the next 90 days.

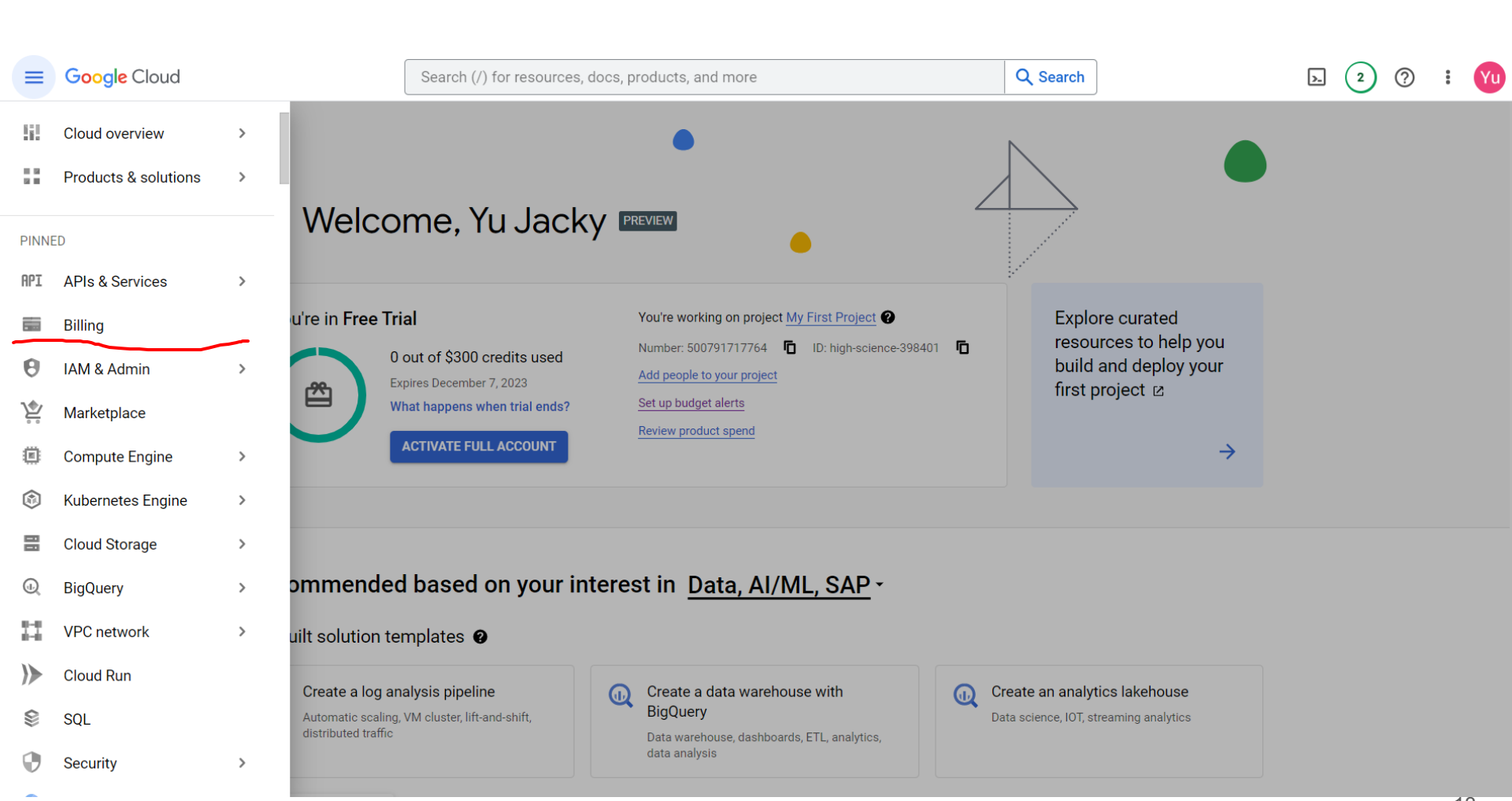
No autocharge after free trial ends

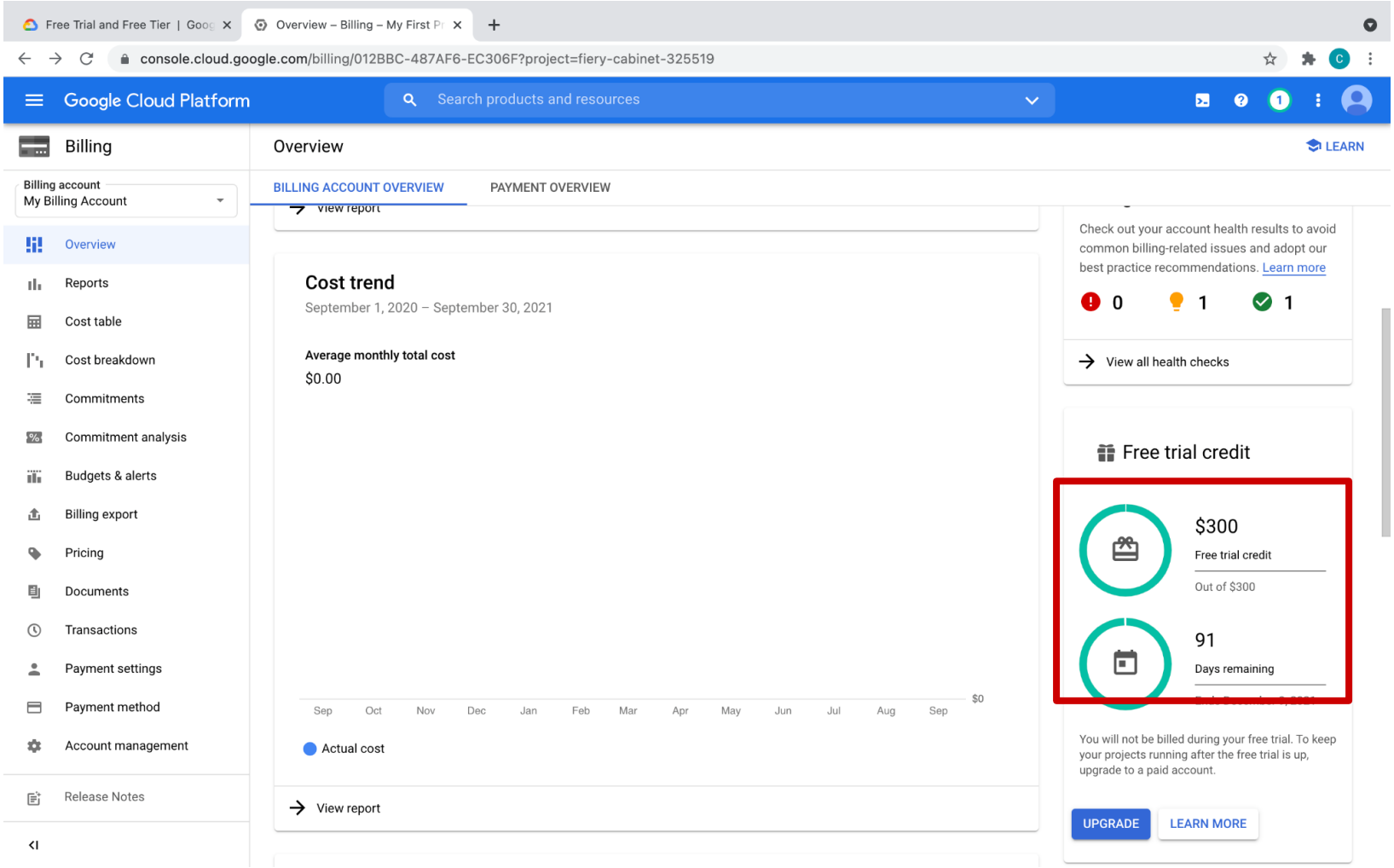
We ask you for your credit card to make sure you are not a robot. If you use a credit or debit card, you won't be charged unless you manually upgrade to a paid account.




Google Cloud
Select a project
Search (/) for resources, docs, products, and more
Search
Welcome, C
PREVIEW
Create or select a project to
Learn more about projects
CREATE OR SELECT A PROJECT
Explore curated resources
Google Cloud Platform
Welcome Qingcheng YU!
Your free trial includes $300 in credit to spend over the next 90 days. To help us serve you better, please answer 4 questions.
What best describes your organization or needs?
What brought you to Google Cloud?
What are you interested in doing with Google Cloud?
4
What best describes your role?
Please select *
Engineer / Developer
CLOSE
DONE
Recommended based
Pre-built solution templates
Deploy a three-tier web app
Deploy load balanced managed VMs
Create a log analysis pipeline

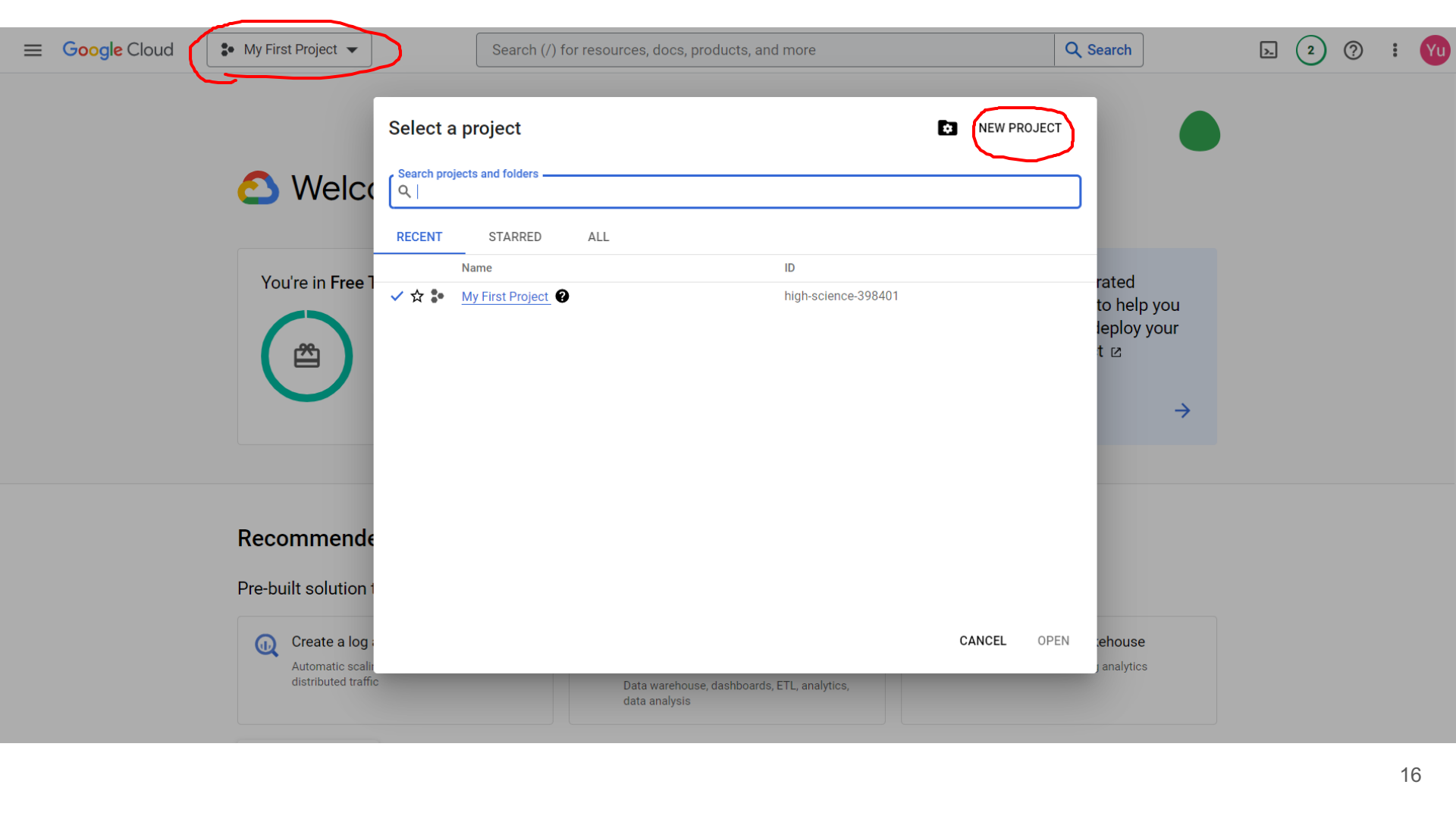

Google Cloud
My First Project
Search (/) for resources, docs, products, and more
Search
2
Yu
Welcome, Yu Jacky
PREVIEW
You're in Free Trial
0 out of $300 credits used
Expires December 7, 2023
What happens when trial ends?
ACTIVATE FULL ACCOUNT
You're working on project My First Project
Number: 500791717764
ID: high-science-398401
Add people to your project
Set up budget alerts
Review product spend
Explore curated resources to help you build and deploy your first project
Recommended based on your interest in Data, AI/ML, SAP
Pre-built solution templates
Create a log analysis pipeline
Automatic scaling, VM cluster, lift-and-shift, distributed traffic
Create a data warehouse with BigQuery
Data warehouse, dashboards, ETL, analytics, data analysis
Create an analytics lakehouse
Data science, IOT, streaming analytics


Google Cloud
Search (/) for resources, docs, products, and more
Search
Cloud overview
Products & solutions
PINNED
APIs & Services
Billing
IAM & Admin
Marketplace
Compute Engine
Kubernetes Engine
Cloud Storage
BigQuery
VPC network
Cloud Run
SQL
Security
Welcome, Yu Jacky
PREVIEW
u're in Free Trial
0 out of $300 credits used
Expires December 7, 2023
What happens when trial ends?
ACTIVATE FULL ACCOUNT
You're working on project My First Project
Number: 500791717764
ID: high-science-398401
Add people to your project
Set up budget alerts
Review product spend
Explore curated resources to help you build and deploy your first project
ommended based on your interest in Data, AI/ML, SAP
uilt solution templates
Create a log analysis pipeline
Automatic scaling, VM cluster, lift-and-shift, distributed traffic
Create a data warehouse with BigQuery
Data warehouse, dashboards, ETL, analytics, data analysis
Create an analytics lakehouse
Data science, IOT, streaming analytics


Free Trial and Free Tier | Goog
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
Overview
LEARN
Billing account
My Billing Account
BILLING ACCOUNT OVERVIEW
PAYMENT OVERVIEW
Overview
Reports
Cost table
Cost breakdown
Commitments
Commitment analysis
Budgets & alerts
Billing export
Pricing
Documents
Transactions
Payment settings
Payment method
Account management
Release Notes
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
You will not be billed during your free trial. To keep your projects running after the free trial is up, upgrade to a paid account.
UPGRADE
LEARN MORE

GCP: Create project

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


Google Cloud
My First Project
Search (/) for resources, docs, products, and more
Search
Select a project
NEW PROJECT
Search projects and folders
RECENT
STARRED
ALL
Name
ID
My First Project
high-science-398401
CANCEL
OPEN
You're in Free T
Recommende
Pre-built solution
Create a log
Automatic scali
distributed traffic
Data warehouse, dashboards, ETL, analytics, data analysis
ehouse
analytics

Google Cloud

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

Search

2

Yu

New Project

You have 10 projects remaining in your quota. Request an increase or delete projects. Learn more

MANAGE QUOTAS

Project name *

EECS6893

Project ID: eecs6893-398401. It cannot be changed later. EDIT

Location *

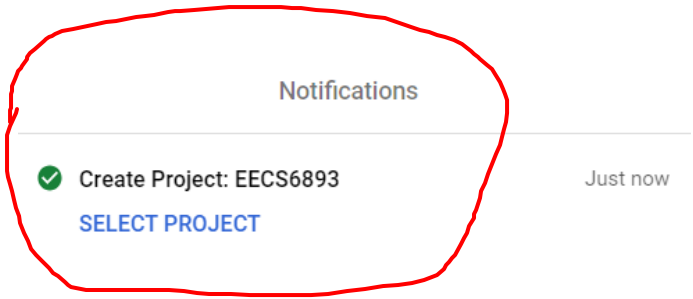
No organization

BROWSE

Parent organization or folder

CREATE

CANCEL



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

DASHBOARD | ACTIVITY | RECOMMENDATIONS | CUSTOMIZE

Project info

Project name
EECS6893

Project number
755979763552

Project ID
eecs6893-398401

ADD PEOPLE TO THIS PROJECT

Go to project settings

Resources

BigQuery
Data warehouse/analytics

SQL
Managed MySQL, PostgreSQL, SQL Server

Compute Engine
VMs, GPUs, TPUs, Disks

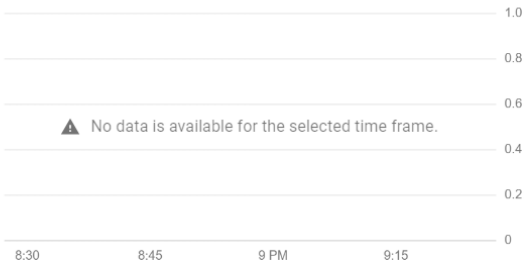
Storage
Multi-class multi-region object storage

Cloud Functions
Event-driven serverless functions

App Engine

APIs

Requests (requests/sec)



Go to APIs overview

Google Cloud Platform status

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

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

Go to Cloud status dashboard

Billing

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

Take a tour of billing

View detailed charges

Monitoring

Create my dashboard

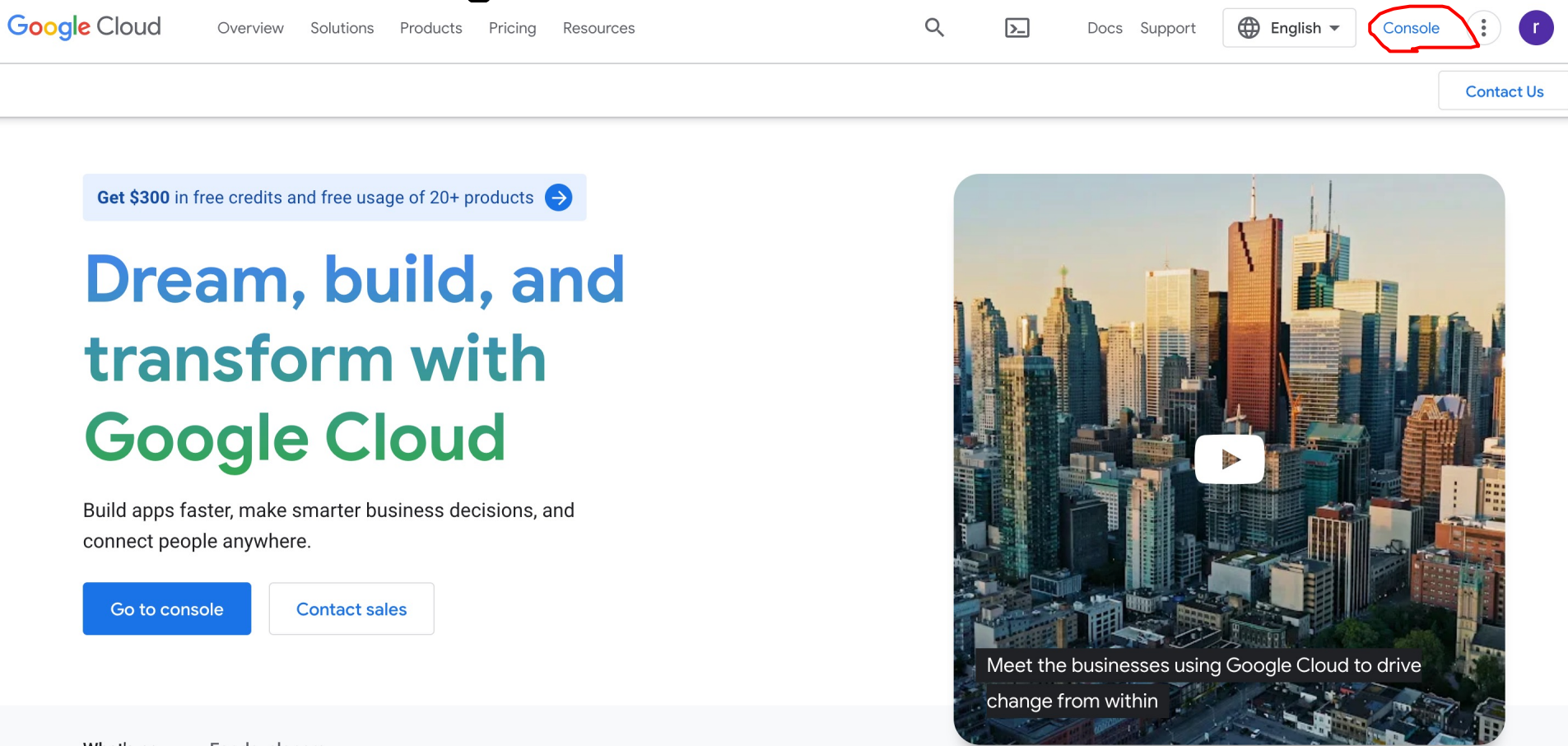
Set up alerting policies

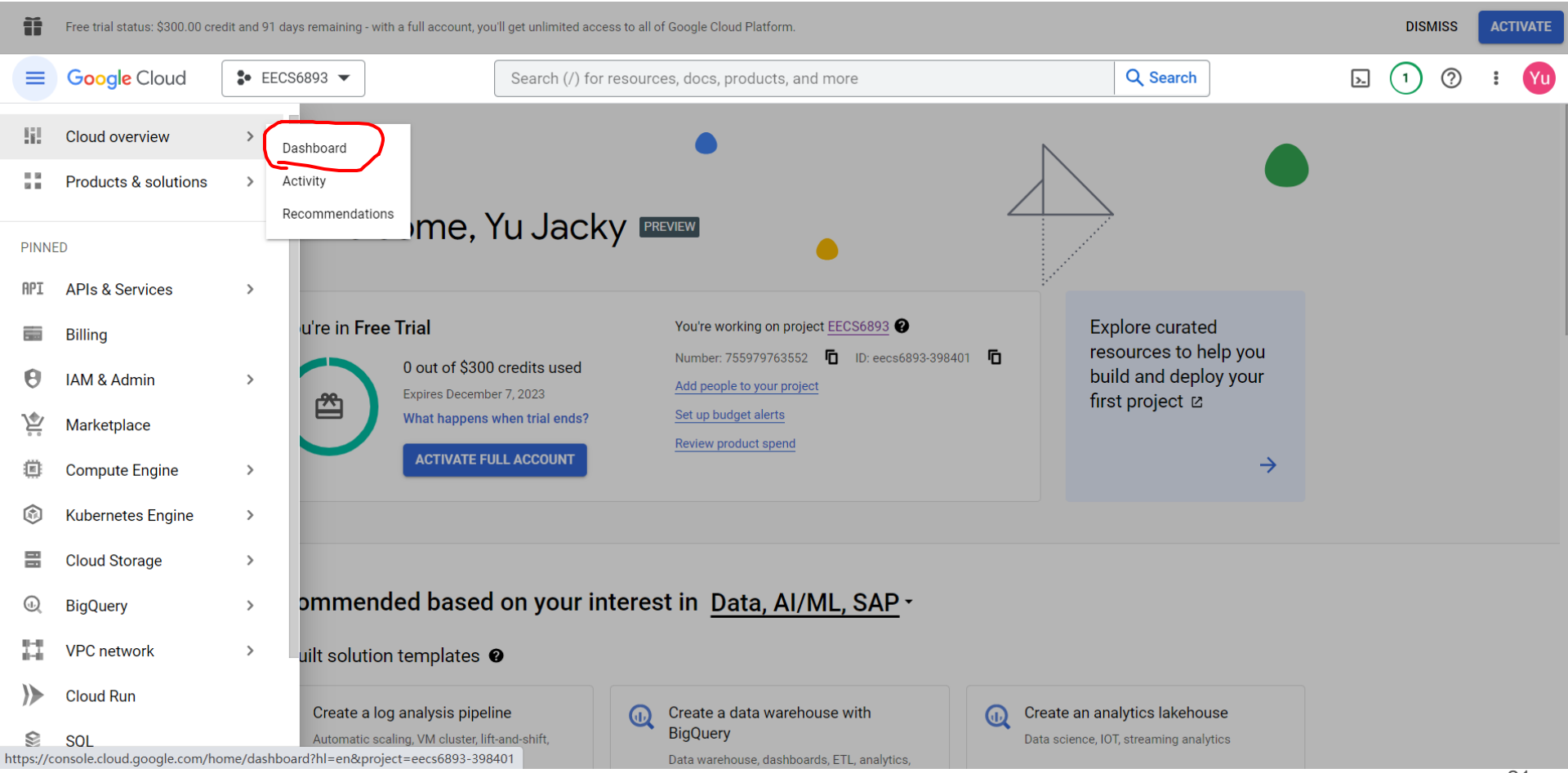
Create uptime checks

GCP: Interaction

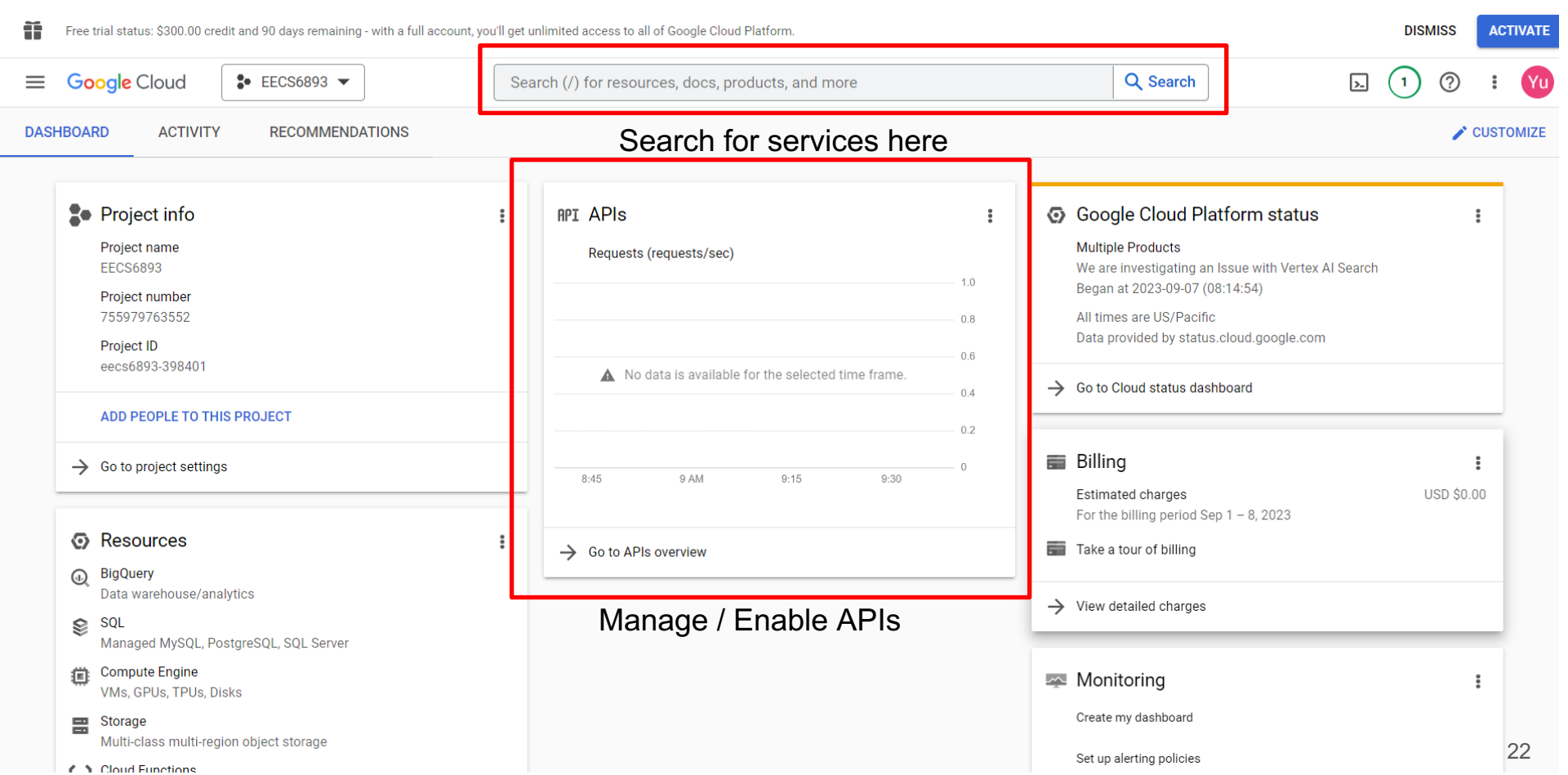
- Graphical UI / console: Useful to create VMs, set up clusters, provision resources, manage teams, etc
- Command line tools / Cloud SDK: Useful for interacting from local host and using the resources once provisioned. E.x. ssh into instances, submit jobs, copy files, etc
- Cloud Shell: Same as command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

Search in Google: GCP console




Free trial status: $300.00 credit and 91 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.
DISMISS
ACTIVATE
Google Cloud
EECS6893
Search (/) for resources, docs, products, and more
Search
Yu
Cloud overview
Dashboard
Activity
Recommendations
Products & solutions
PINNED
APIs & Services
Billing
IAM & Admin
Marketplace
Compute Engine
Kubernetes Engine
Cloud Storage
BigQuery
VPC network
Cloud Run
SQL
me, Yu Jacky
PREVIEW
u're in Free Trial
0 out of $300 credits used
Expires December 7, 2023
What happens when trial ends?
ACTIVATE FULL ACCOUNT
You're working on project EECS6893
Number: 755979763552
ID: eecs6893-398401
Add people to your project
Set up budget alerts
Review product spend
Explore curated resources to help you build and deploy your first project
ommended based on your interest in Data, AI/ML, SAP
uilt solution templates
Create a log analysis pipeline
Automatic scaling, VM cluster, lift-and-shift,
Create a data warehouse with BigQuery
Data warehouse, dashboards, ETL, analytics,
Create an analytics lakehouse
Data science, IOT, streaming analytics
https://console.cloud.google.com/home/dashboard?hl=en&project=eecs6893-398401

GCP: console


Free trial status: $300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.
DISMISS
ACTIVATE
Google Cloud
EECS6893
Search (/) for resources, docs, products, and more
Search
DASHBOARD
ACTIVITY
RECOMMENDATIONS
CUSTOMIZE
Search for services here
Project info
Project name
EECS6893
Project number
755979763552
Project ID
eecs6893-398401
ADD PEOPLE TO THIS PROJECT
Go to project settings
APIs
Requests (requests/sec)
1.0
0.8
0.6
0.4
0.2
0
No data is available for the selected time frame.
8:45
9 AM
9:15
9:30
Go to APIs overview
Manage / Enable APIs
Google Cloud Platform status
Multiple Products
We are investigating an Issue with Vertex AI Search
Began at 2023-09-07 (08:14:54)
All times are US/Pacific
Data provided by status.cloud.google.com
Go to Cloud status dashboard
Billing
Estimated charges
USD $0.00
For the billing period Sep 1 – 8, 2023
Take a tour of billing
View detailed charges
Resources
BigQuery
Data warehouse/analytics
SQL
Managed MySQL, PostgreSQL, SQL Server
Compute Engine
VMs, GPUs, TPUs, Disks
Storage
Multi-class multi-region object storage
Cloud Functions
Monitoring
Create my dashboard
Set up alerting policies

GCP: Cloud SDK

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

Filter

gcloud CLI
Product overview
gcloud CLI overview
gcloud CLI cheat sheet

Quickstart
Install the Google Cloud CLI

How-to guides
All how-to guides
Installing the gcloud CLI
Setting up the gcloud CLI
Managing gcloud CLI components
Scripting gcloud CLI commands
Enabling accessibility features
Using gcloud interactive shell
Uninstalling the gcloud CLI

Installing the latest gcloud CLI version (445.0.0)

Note: If you are behind a proxy/firewall, see the proxy settings page for more information on installation.

Linux | Debian/Ubuntu | Red Hat/Fedora/CentOS | macOS | Windows | Chromebook

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

1. Download the Google Cloud CLI installer.

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

```
(New-Object Net.WebClient).DownloadFile("https://dl.google.com/dl/cloudsdk/channels/rapid

& $env:Temp\GoogleCloudSDKInstaller.exe
```

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

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

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

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

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

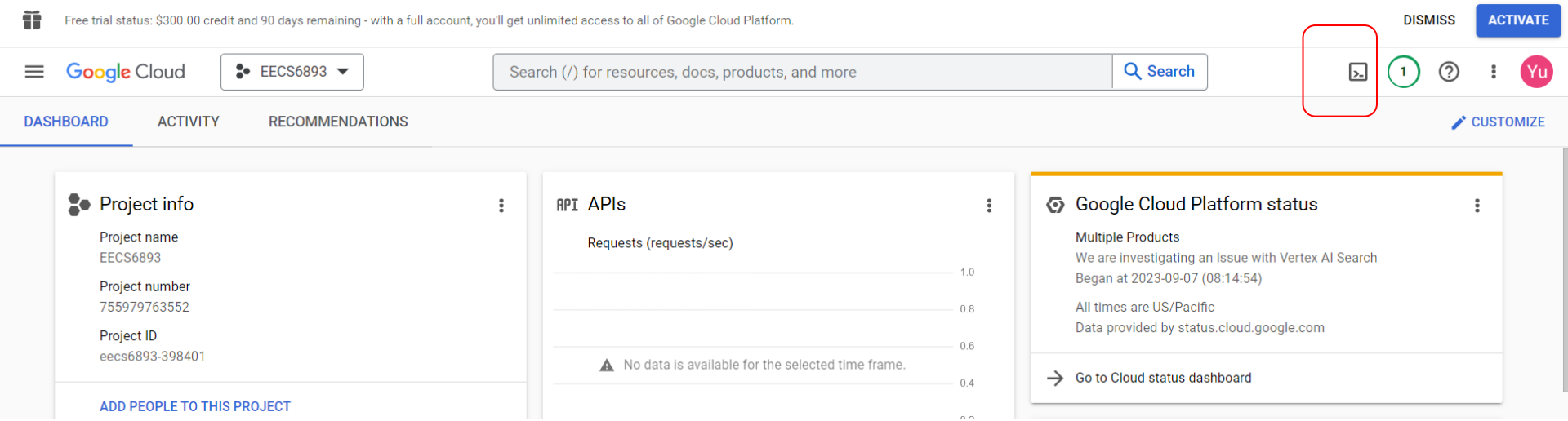
```
Pick cloud project to use:
 [1] eecs6893-398401
 [2] high-science-398401
 [3] resonant-time-398400
 [4] Enter a project ID
 [5] Create a new project
Please enter numeric choice or text value (must exactly match list item):
```

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

```
C:\Users\11518\AppData\Local\Google\Cloud SDK>gcloud info
Google Cloud SDK [402.0.0]

Platform: [Windows, x86_64] uname_result(system='Windows', node='LAPTOP-LJO7H8BA', release='10', version='10.0.19044', machine='AMD64')
Locale: ('zh_CN', 'cp1252')
Python Version: [3.9.12 (tags/v3.9.12:b28265d, Mar 23 2022, 23:52:46) [MSC v.1929 64 bit (AMD64)]]
Python Location: [C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\platform\bundledpython\python.exe]
OpenSSL: [OpenSSL 1.1.1n  15 Mar 2022]
Requests Version: [2.25.1]
urllib3 Version: [1.26.9]
Site Packages: [Disabled]

Installation Root: [C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk]
Installed Components:
  beta: [2022.09.12]
  bq: [2.0.75]
  core: [2022.09.12]
  gcloud-crc32c: [1.0.0]
  gsutil: [5.13]
System PATH: [C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\bin\..\bin\sdk;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\bin;C:\Program Files\Common Files\Oracle\Java\javapath;C
:\Program Files (x86)\Common Files\Oracle\Java\javapath;E:\manager wizard\ChemScript\Lib;C:\Program Files (x86)\Intel\iCLS Client\;C:\Program Files\Intel\iCLS Client\;C:\Windows\system32;C:\Windows;C:\Windows\Sy
stem32\Wbem;C:\Windows\System32\WindowsPowerShell\v1.0\;C:\Program Files (x86)\NVIDIA Corporation\PhysX\Common;C:\Program Files (x86)\Intel\Intel(R) Management Engine Components\DAL;C:\Program Files\Intel\Intel(
R) Management Engine Components\DAL;C:\Program Files (x86)\Intel\Intel(R) Management Engine Components\IPT;C:\Program Files\Intel\Intel(R) Management Engine Components\IPT;C:\WINDOWS\system32;C:\WINDOWS;C:\WINDO
WS\System32\Wbem;C:\WINDOWS\System32\WindowsPowerShell\v1.0\;C:\WINDOWS\System32\OpenSSH\;C:\Program Files (x86)\Wolfram Research\WolframScript\;E:\MATLAB_R2020a\bin;C:\Program Files\Java\jdk1.8.0_202\bin;E:\Had
oop\hadoop-3.2.4\hadoop-3.2.4\bin;E:\Hadoop\hadoop-3.2.4\hadoop-3.2.4\sbin;E:\Git\cmd;E:\Git\Git LFS;E:\Node.js\;C:\ProgramData\chocolatey\bin;E:\Program Files (x86)\Eclipse\Sumo\bin\;E:\Program Files (x86)\Ecli
pse\Sumo\tools\;C:\Program Files\MySQL\MySQL Shell 8.0\bin\;E:\python\python3.9.1\Scripts\;E:\python\python3.9.1\;C:\Users\11518\AppData\Local\Microsoft\WindowsApps;E:\??????\MiKTeX\miktex\bin\x64\;C:\Program Fi
les\MPICH2\bin;E:\LAMMPS 64-bit 14May2021\bin;E:\Pycharm\PyCharm Community Edition 2022.1.4\bin;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\bin;C:\Users\11518\AppData\Roaming\TinyTeX\bin\win32
;E:\Microsoft VS Code\bin;C:\Users\11518\AppData\Local\GitHubDesktop\bin;E:\Fiddler;C:\Users\11518\AppData\Local\Pandoc\;C:\Users\11518\AppData\Roaming\npm;]
Python PATH: [C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\lib\third_party;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\lib;E:\LAMMPS 64-bit 14May2021\Python;C:\Users\11518\Ap
pData\Local\Google\Cloud SDK\google-cloud-sdk\platform\bundledpython\python39.zip;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\platform\bundledpython\DLLs;C:\Users\11518\AppData\Local\Google\Cl
oud SDK\google-cloud-sdk\platform\bundledpython\lib;C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\platform\bundledpython]
Cloud SDK on PATH: [True]
Kubectl on PATH: [False]

Installation Properties: [C:\Users\11518\AppData\Local\Google\Cloud SDK\google-cloud-sdk\properties]
User Config Directory: [C:\Users\11518\AppData\Roaming\gcloud]
Active Configuration Name: [default]
Active Configuration Path: [C:\Users\11518\AppData\Roaming\gcloud\configurations\config_default]

Account: [jackyyu2021111@gmail.com]
Project: [eecs6893-398401]

Current Properties:
  [accessibility]
    screen_reader: [False] (property file)
  [core]
    account: [jackyyu2021111@gmail.com] (property file)
```

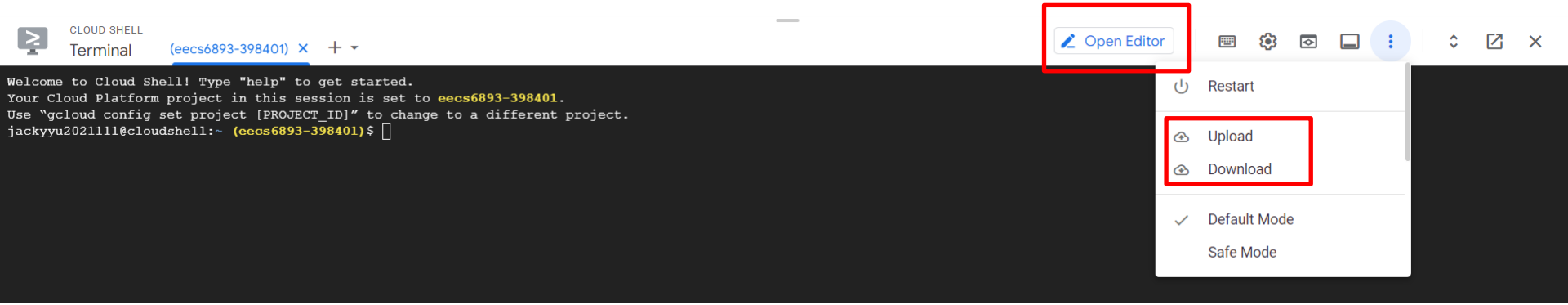
```
C:\Users\11518\AppData\Local\Google\Cloud SDK>gcloud auth list
     Credentialed Accounts
ACTIVE  ACCOUNT
*       jackyyu2021111@gmail.com
        qy2281@columbia.edu

To set the active account, run:
    $ gcloud config set account `ACCOUNT`


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

Your current Google Cloud CLI version is: 402.0.0
The latest available version is: 445.0.0

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

GCP: Cloud Shell



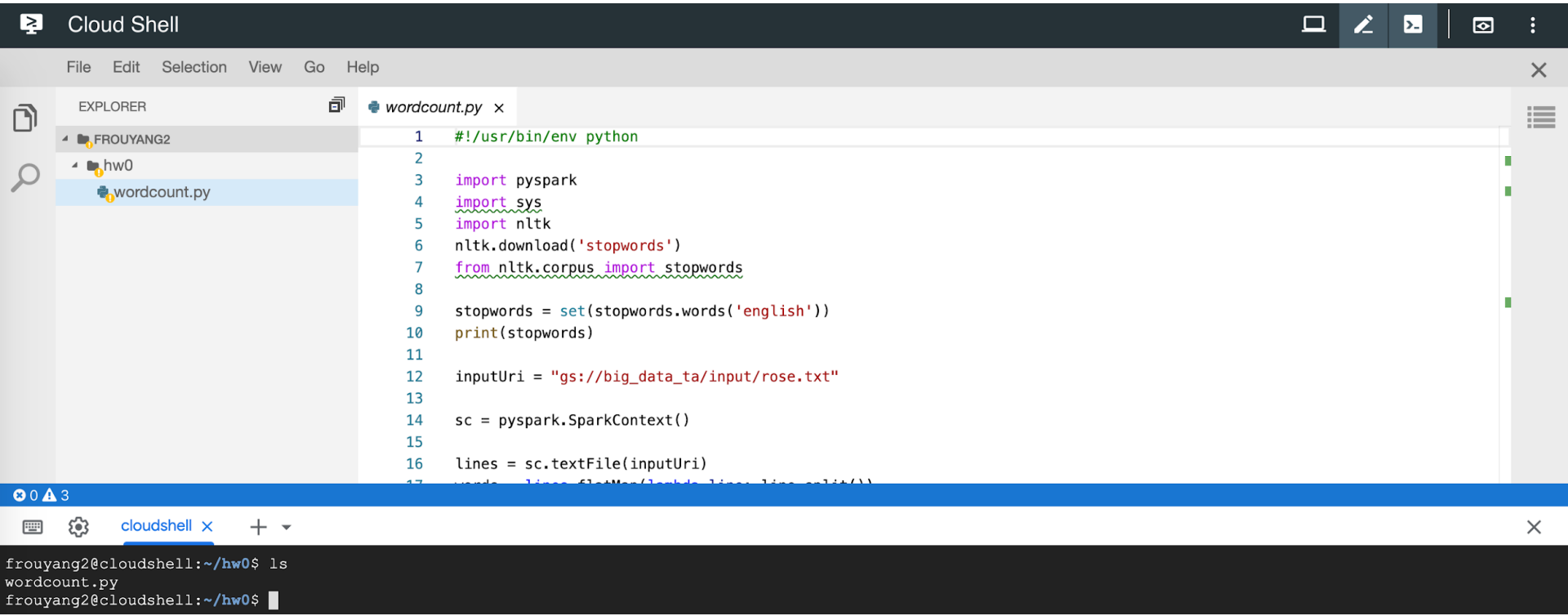
persistent home directory :). The most useful way to complete the HW0

GCP: Cloud Shell



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

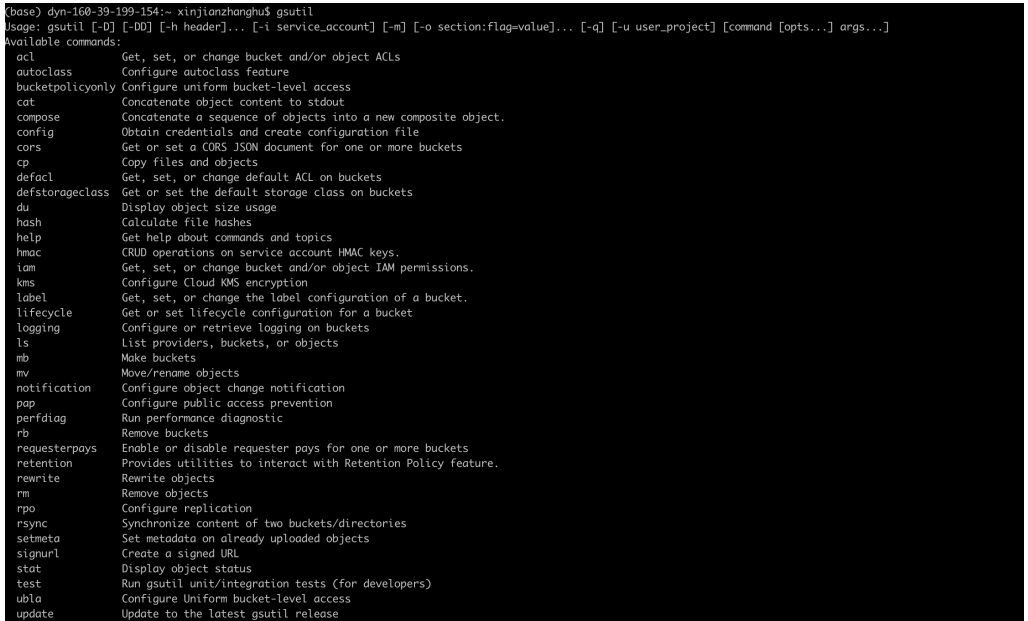
GCP: Cloud Shell Code Editor



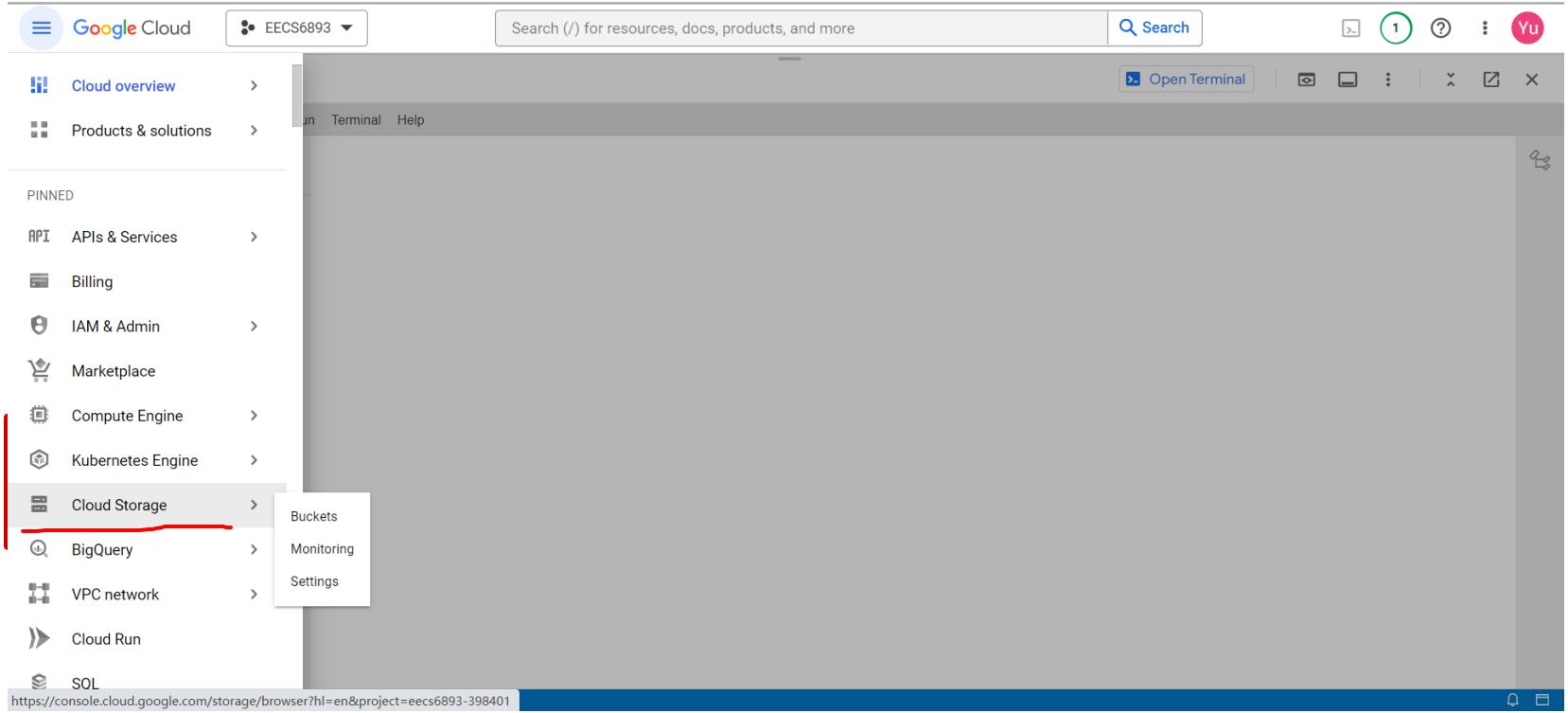
Cloud Storage

Cloud Storage

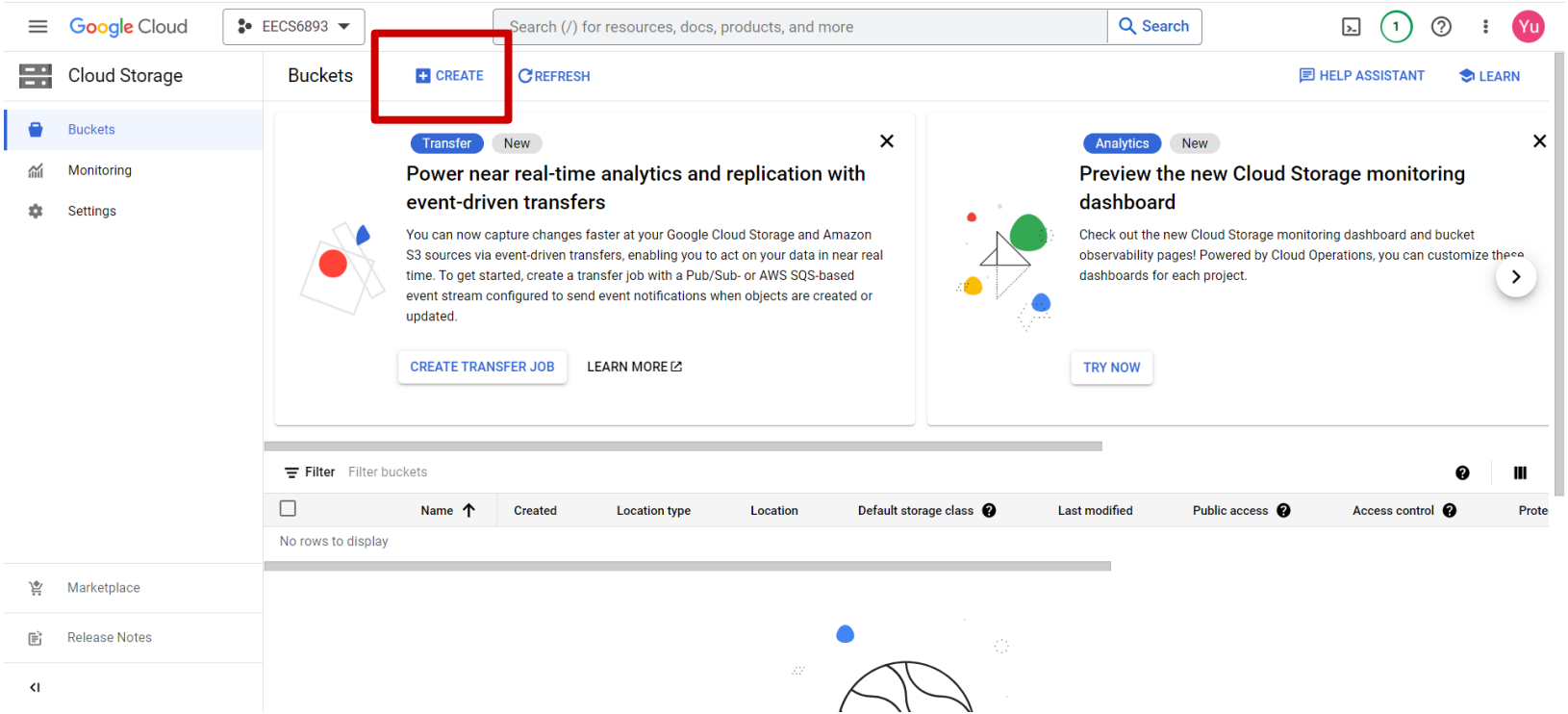
- Online file storage system
- Graphical UI through console
- Command line tool: `gsutil`



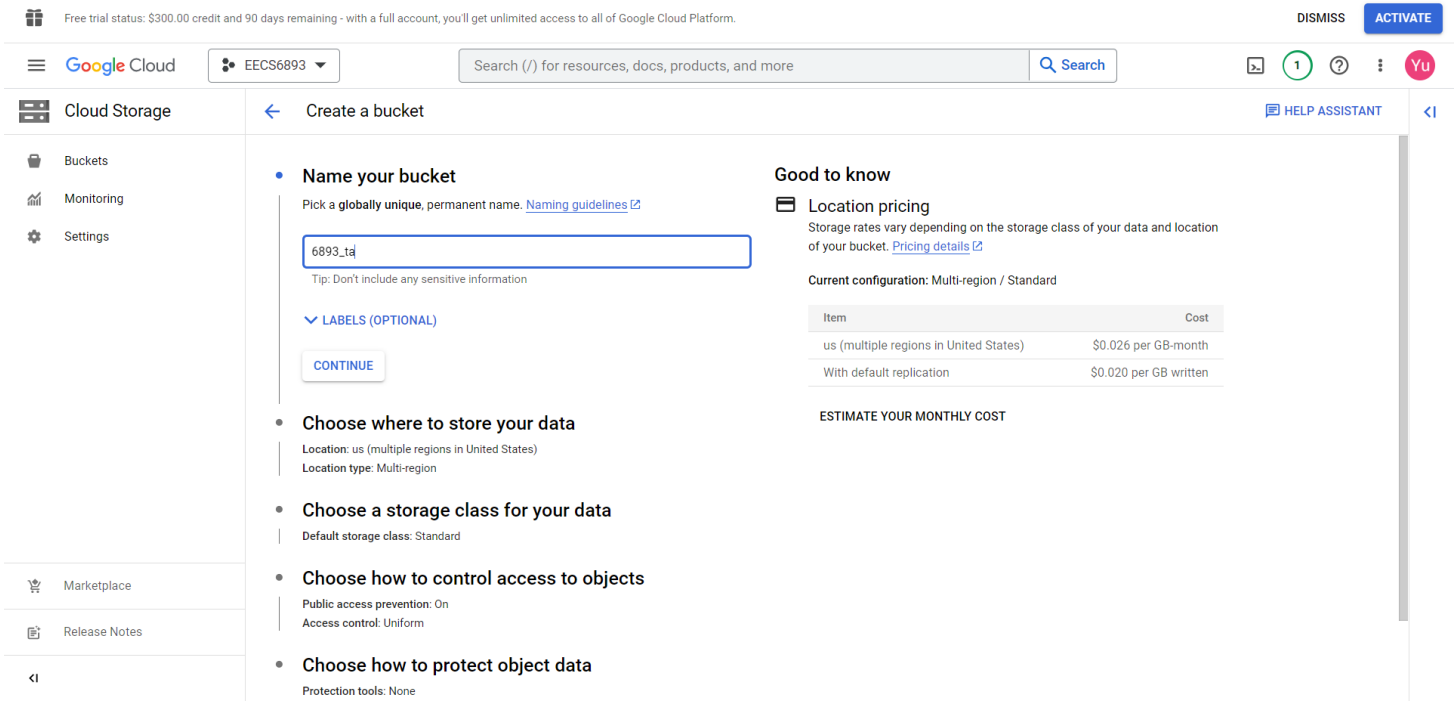
Cloud Storage



Cloud Storage

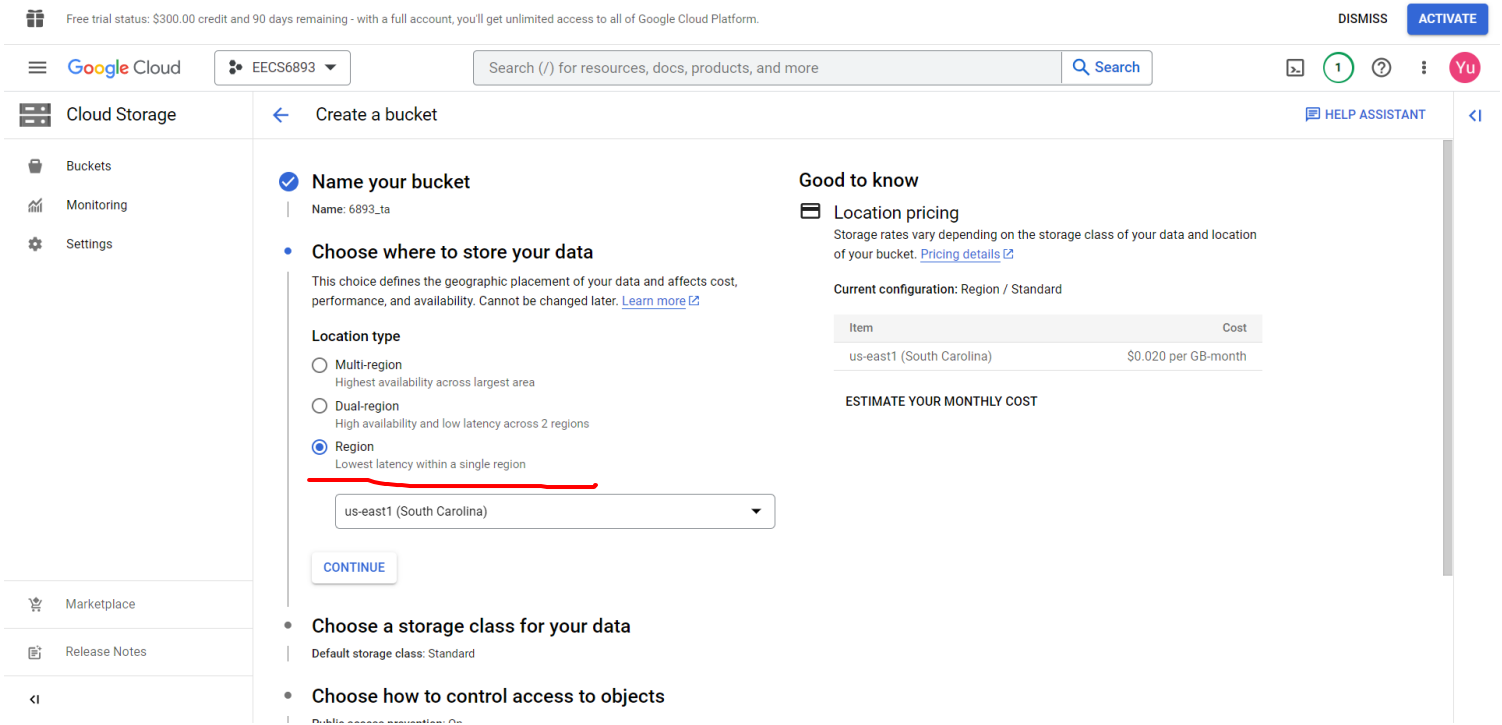


Cloud Storage

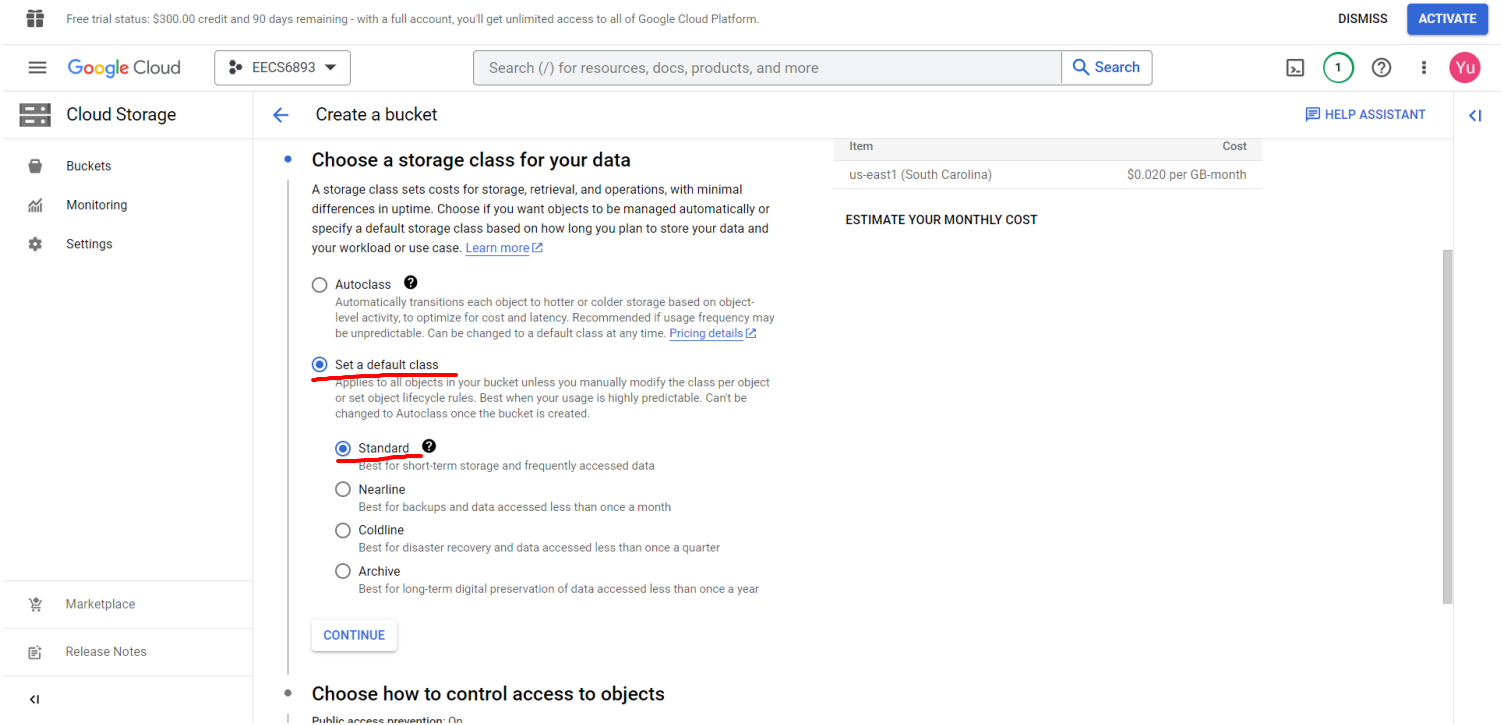


Name your own bucket

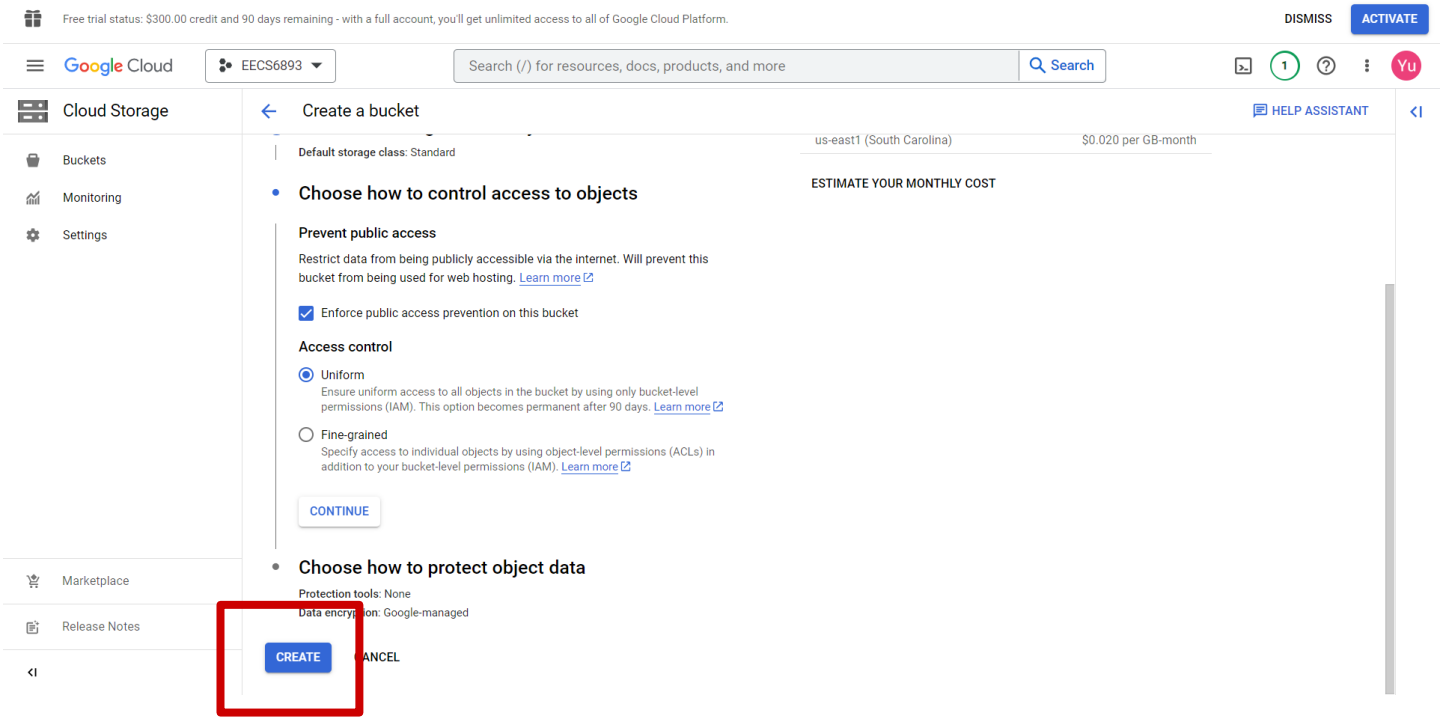
Cloud Storage



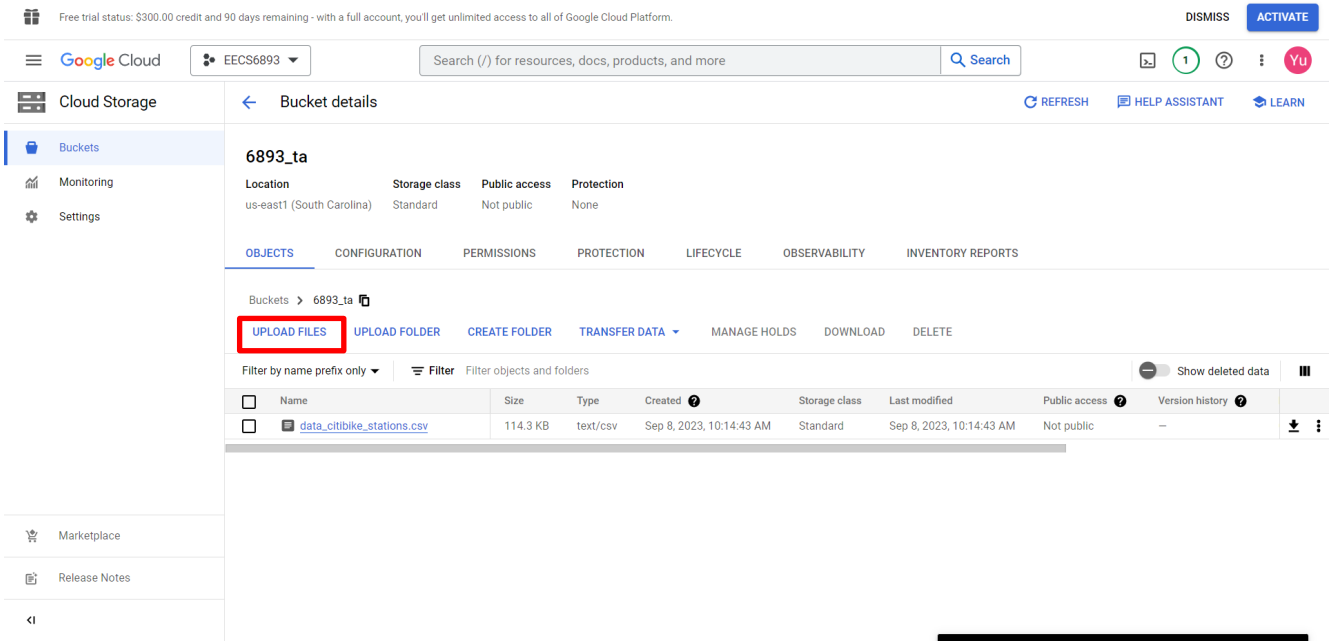
Cloud Storage



Cloud Storage



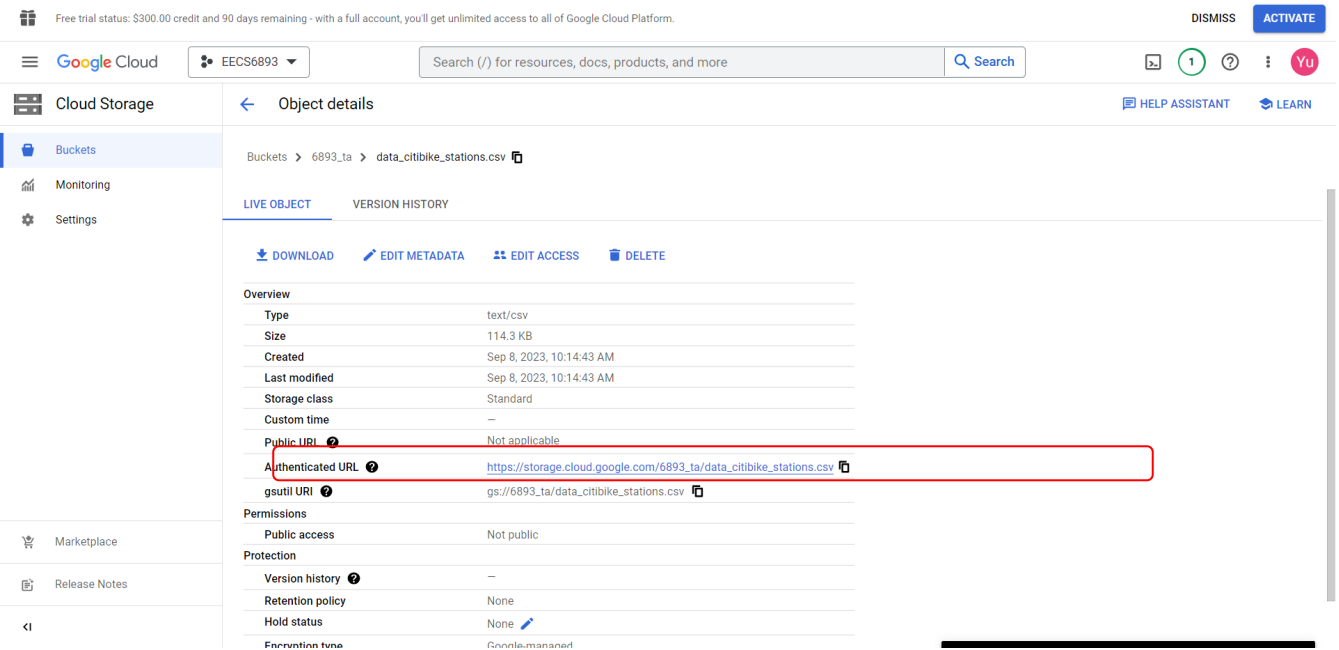
Cloud Storage



dataset provided in HW0 details

Cllick on the uploaded dataset file

Cloud Storage



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

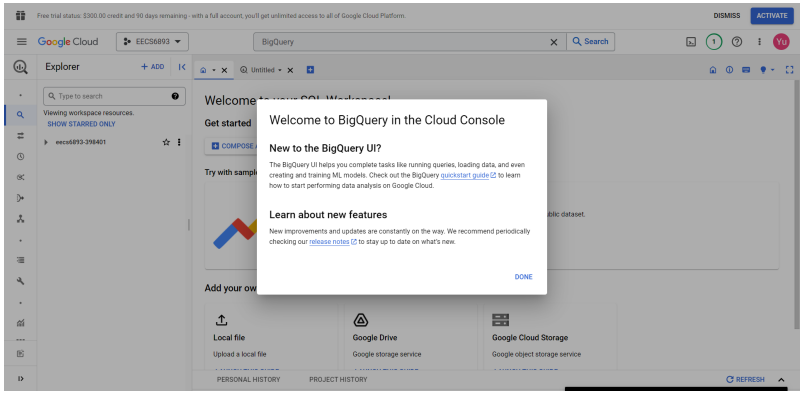
Cloud Storage - `gsutil`

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

BigQuery

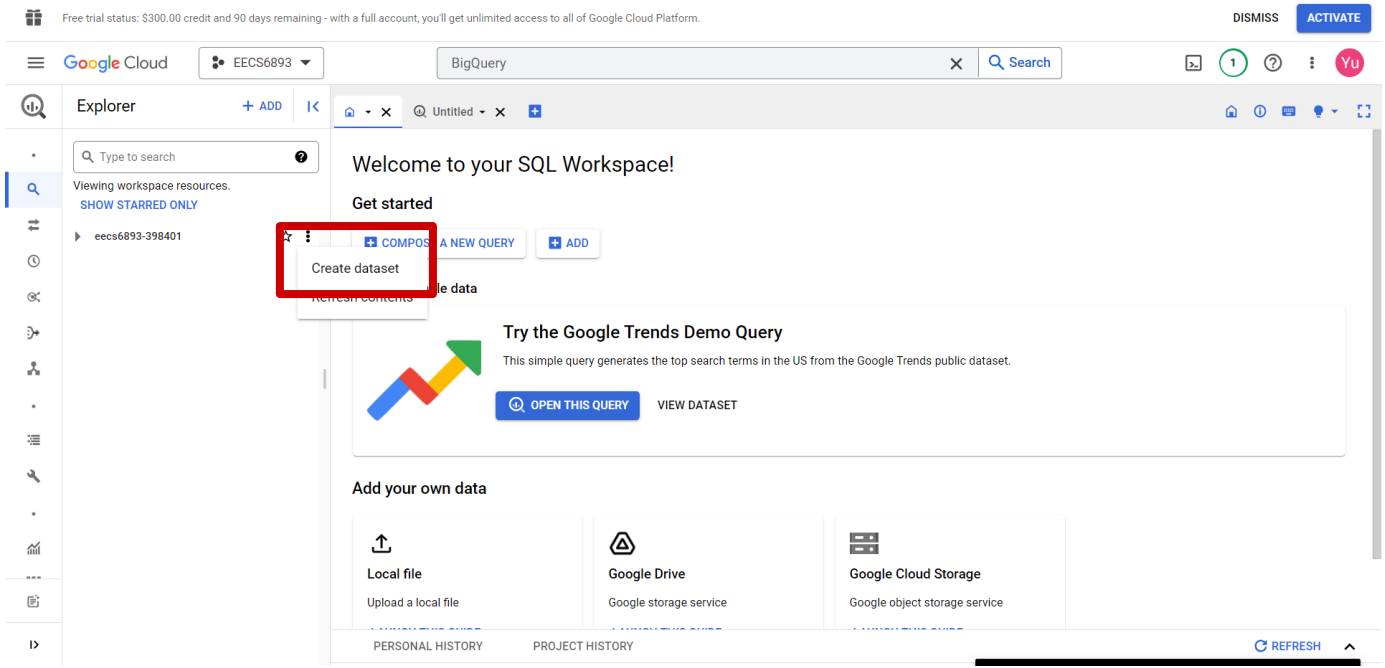
BigQuery

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

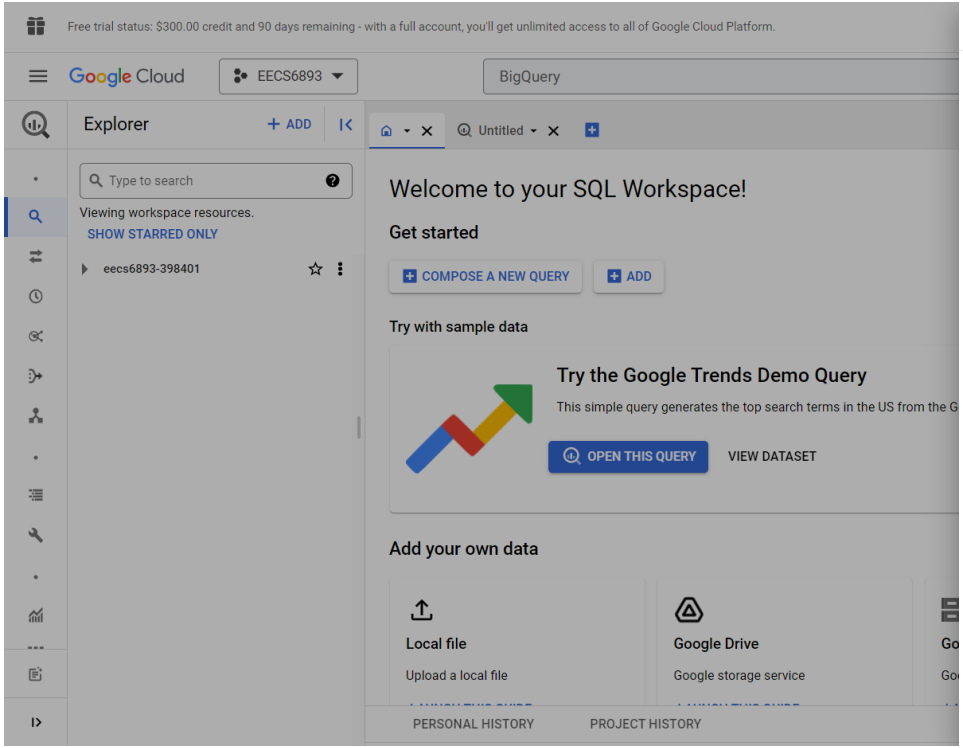


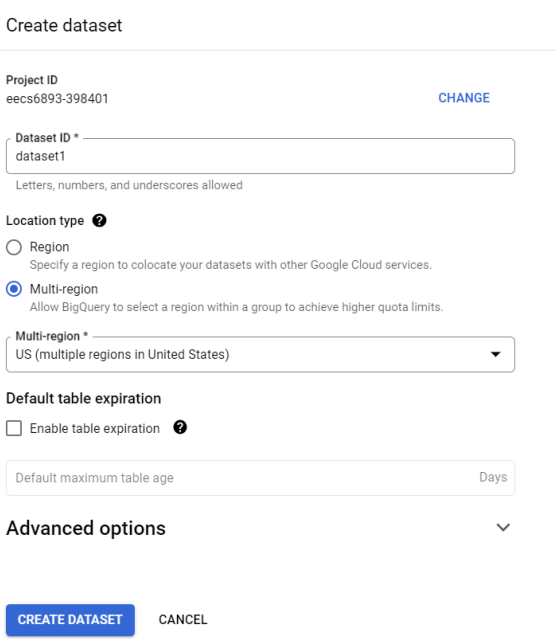
search for BigQuery and go for it

BigQuery

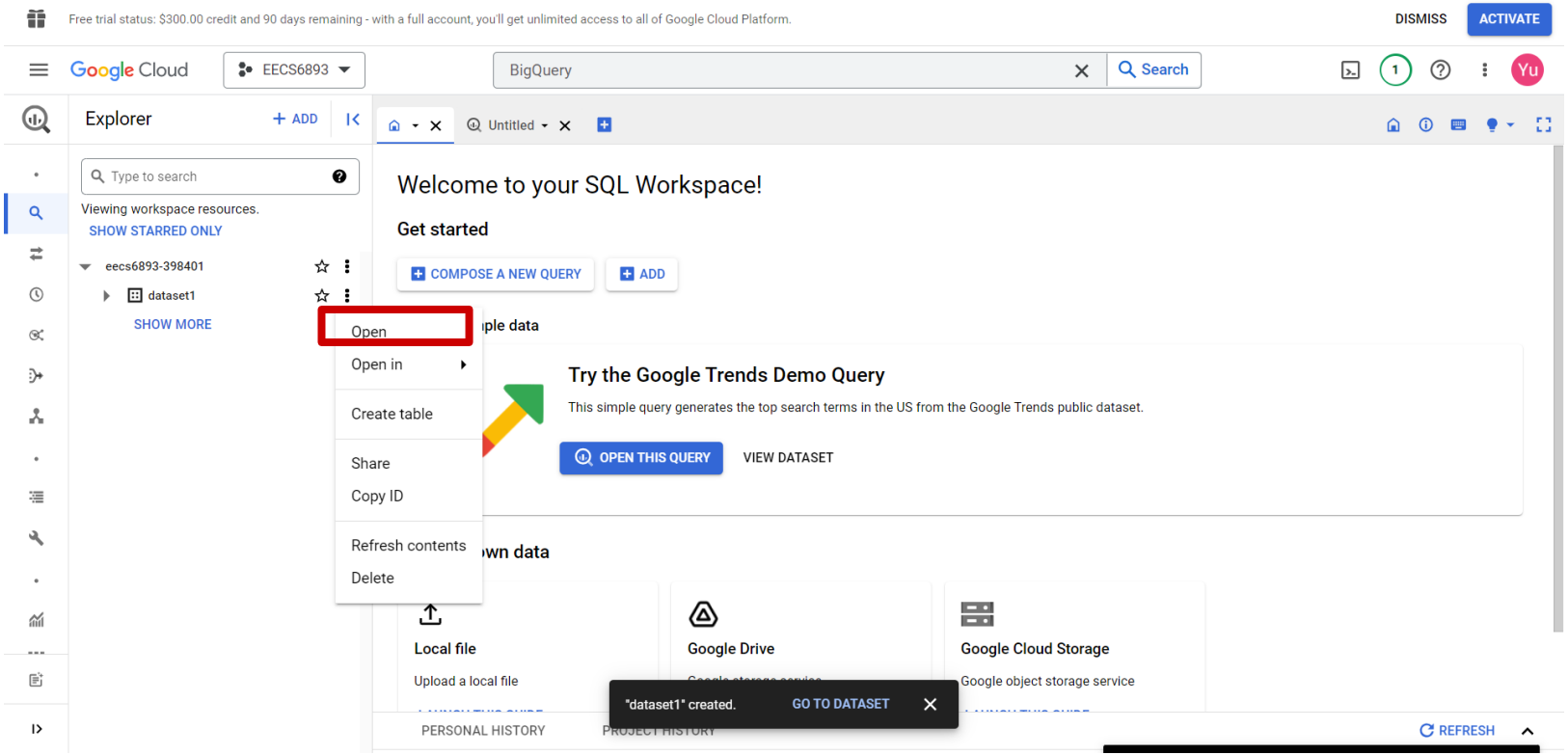

Free trial status: $300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.
DISMISS
ACTIVATE
Google Cloud
EECS6893
BigQuery
Search
Explorer
ADD
Type to search
Viewing workspace resources.
SHOW STARRED ONLY
eecs6893-398401
Create dataset
Untitled
Welcome to your SQL Workspace!
Get started
COMPOSE A NEW QUERY
ADD
Try the Google Trends Demo Query
This simple query generates the top search terms in the US from the Google Trends public dataset.
OPEN THIS QUERY
VIEW DATASET
Add your own data
Local file
Upload a local file
Google Drive
Google storage service
Google Cloud Storage
Google object storage service
PERSONAL HISTORY
PROJECT HISTORY
REFRESH

BigQuery

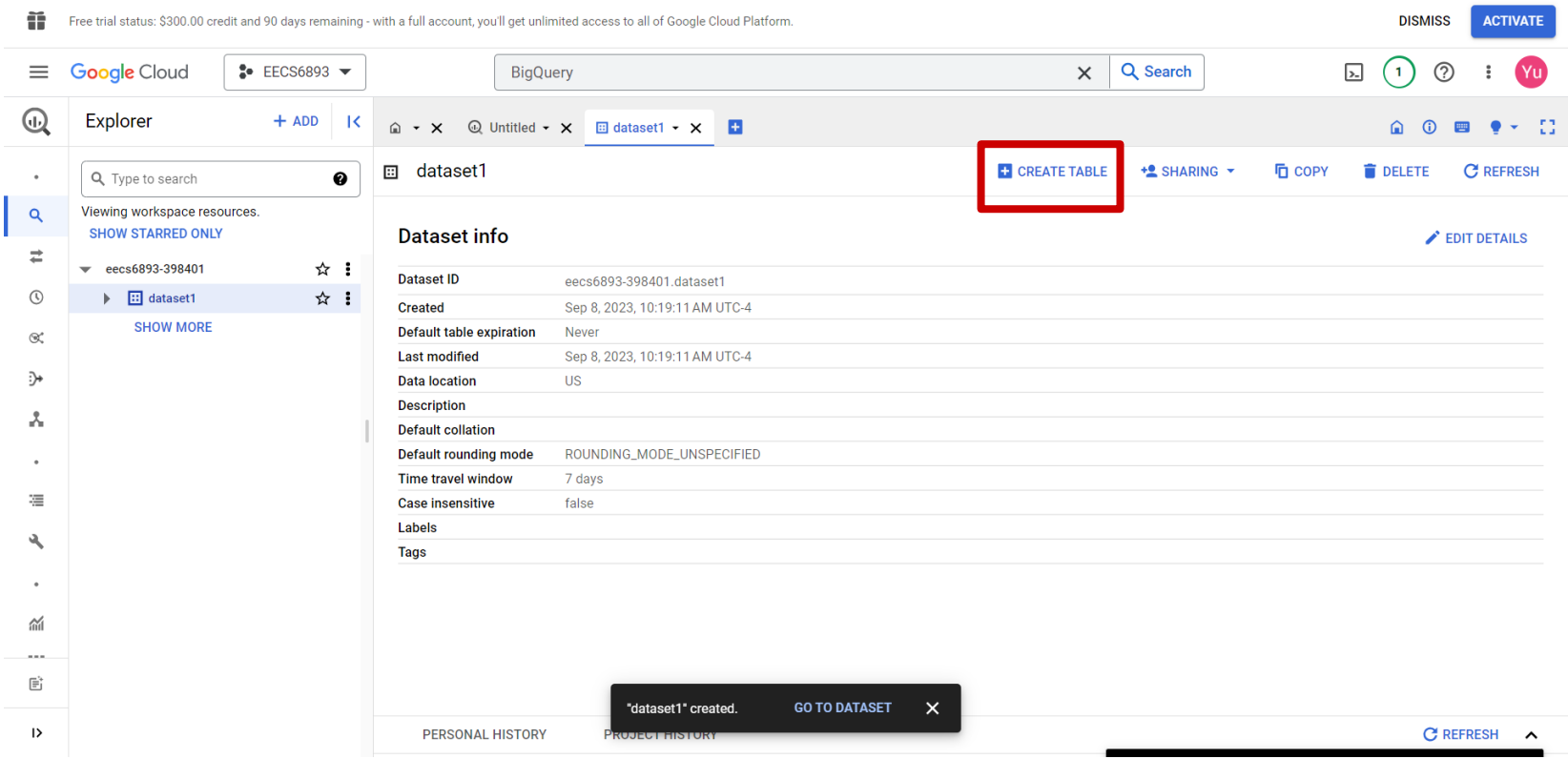

Free trial status: $300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.
Google Cloud
EECS6893
BigQuery
Explorer
+ ADD
Type to search
Viewing workspace resources.
SHOW STARRED ONLY
eecs6893-398401
Untitled
Welcome to your SQL Workspace!
Get started
COMPOSE A NEW QUERY
ADD
Try with sample data
Try the Google Trends Demo Query
This simple query generates the top search terms in the US from the G
OPEN THIS QUERY
VIEW DATASET
Add your own data
Local file
Upload a local file
Google Drive
Google storage service
PERSONAL HISTORY
PROJECT HISTORY


Create dataset
Project ID
eecs6893-398401
CHANGE
Dataset ID *
dataset1
Letters, numbers, and underscores allowed
Location type
Region
Specify a region to colocate your datasets with other Google Cloud services.
Multi-region
Allow BigQuery to select a region within a group to achieve higher quota limits.
Multi-region *
US (multiple regions in United States)
Default table expiration
Enable table expiration
Default maximum table age
Days
Advanced options
CREATE DATASET
CANCEL

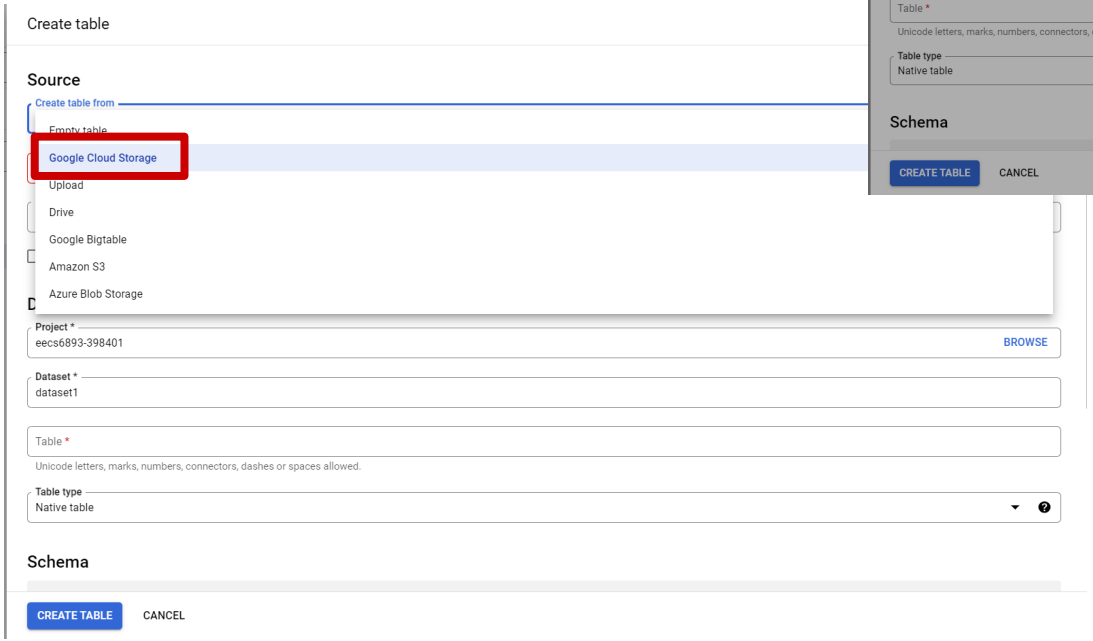
BigQuery

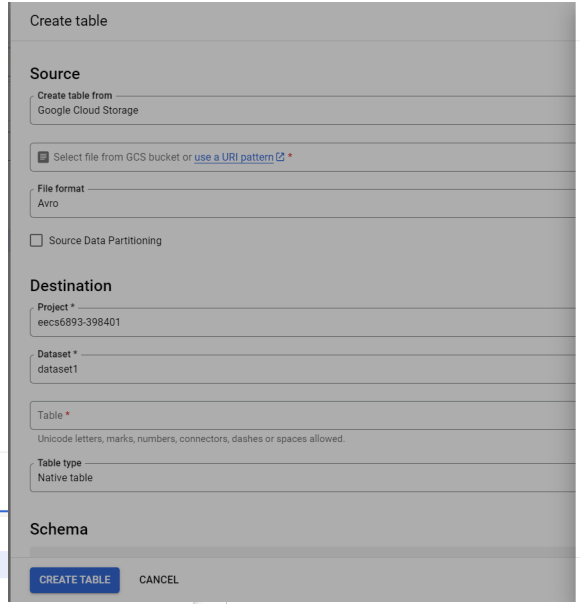

Free trial status: $300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.
DISMISS
ACTIVATE
Google Cloud
EECS6893
BigQuery
Search
Explorer
+ ADD
Type to search
Viewing workspace resources.
SHOW STARRED ONLY
eecs6893-398401
dataset1
SHOW MORE
Open
Open in
Create table
Share
Copy ID
Refresh contents
Delete
Untitled
Welcome to your SQL Workspace!
Get started
COMPOSE A NEW QUERY
ADD
Try the Google Trends Demo Query
This simple query generates the top search terms in the US from the Google Trends public dataset.
OPEN THIS QUERY
VIEW DATASET
Local file
Upload a local file
Google Drive
Google Cloud Storage
Google object storage service
"dataset1" created.
GO TO DATASET
PERSONAL HISTORY
REFRESH

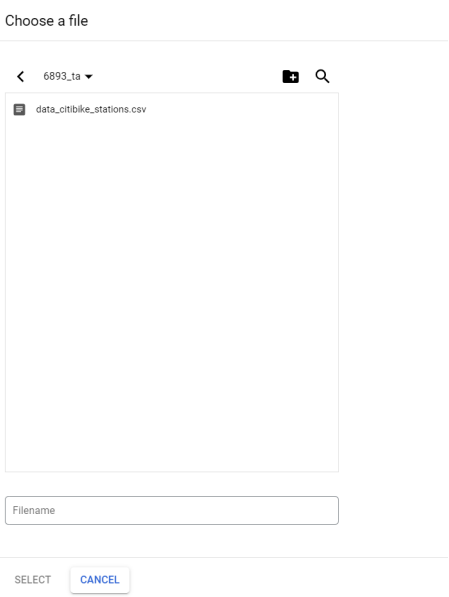
BigQuery


Free trial status: $300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.
DISMISS
ACTIVATE
Google Cloud
EECS6893
BigQuery
Search
Explorer
+ ADD
Type to search
Viewing workspace resources.
SHOW STARRED ONLY
eecs6893-398401
dataset1
SHOW MORE
Untitled
dataset1
dataset1
CREATE TABLE
SHARING
COPY
DELETE
REFRESH
Dataset info
EDIT DETAILS
Dataset ID eecs6893-398401.dataset1
Created Sep 8, 2023, 10:19:11 AM UTC-4
Default table expiration Never
Last modified Sep 8, 2023, 10:19:11 AM UTC-4
Data location US
Description
Default collation
Default rounding mode ROUNDING_MODE_UNSPECIFIED
Time travel window 7 days
Case insensitive false
Labels
Tags
"dataset1" created.
GO TO DATASET
PERSONAL HISTORY
PROJECT HISTORY
REFRESH

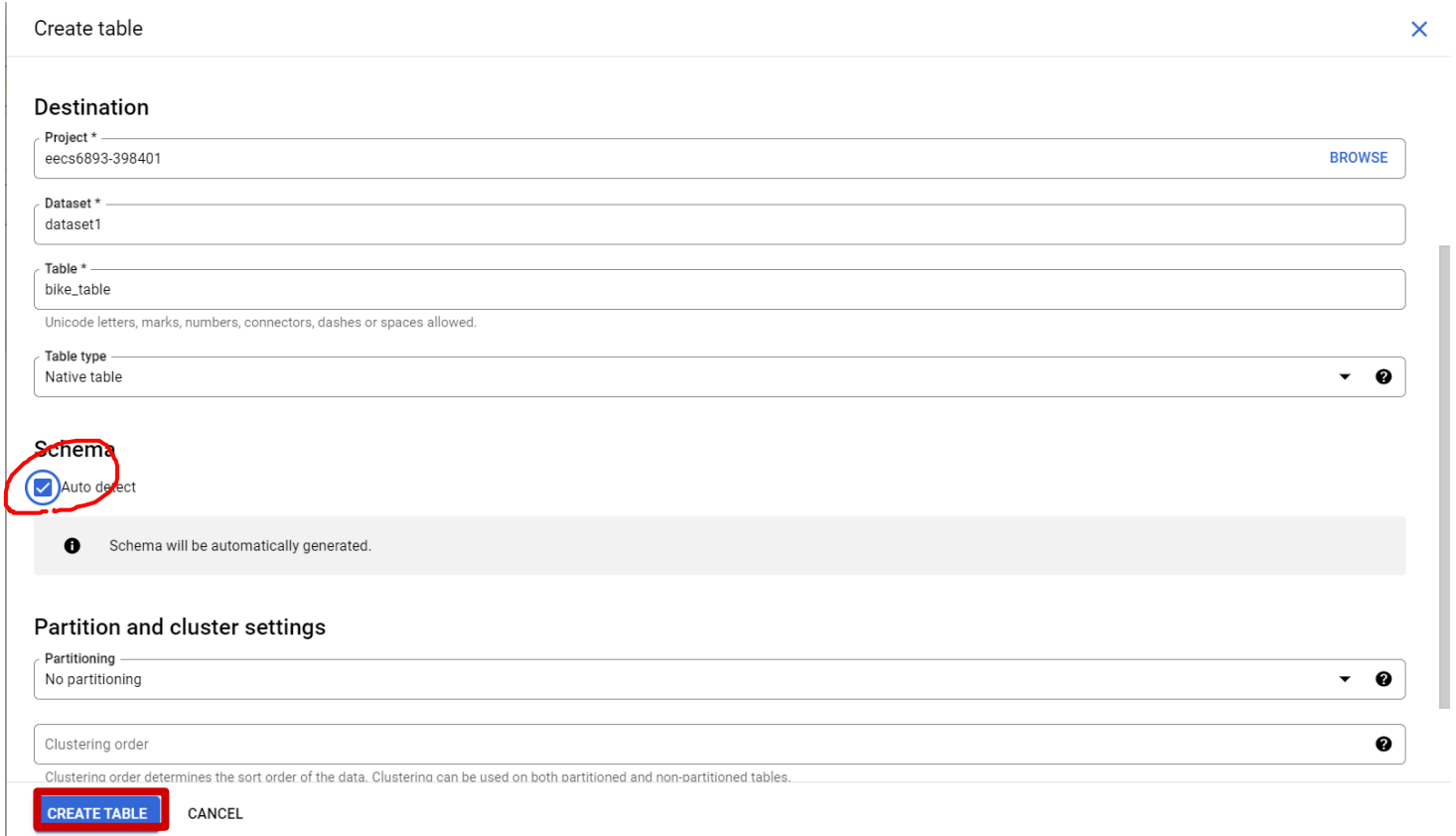
BigQuery


Create table
Source
Create table from
Empty table
Google Cloud Storage
Upload
Drive
Google Bigtable
Amazon S3
Azure Blob Storage
Project *
eecs6893-398401
BROWSE
Dataset *
dataset1
Table *
Unicode letters, marks, numbers, connectors, dashes or spaces allowed.
Table type
Native table
Schema
CREATE TABLE
CANCEL

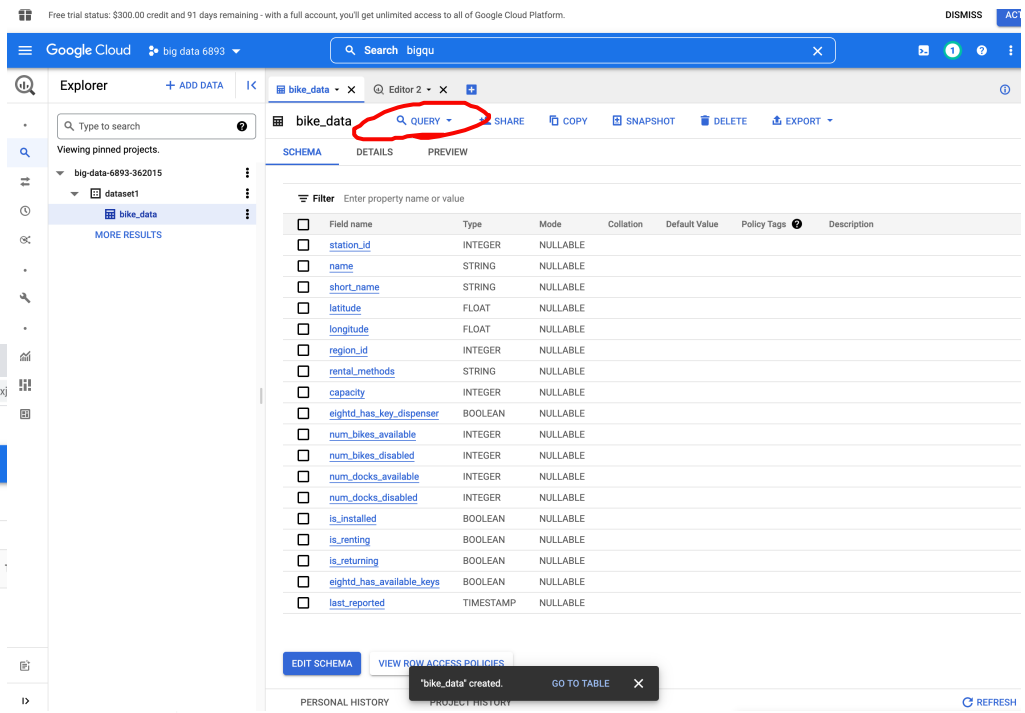

Create table
Source
Create table from
Google Cloud Storage
Select file from GCS bucket or use a URI pattern *
File format
Avro
Source Data Partitioning
Destination
Project *
eecs6893-398401
Dataset *
dataset1
Table *
Unicode letters, marks, numbers, connectors, dashes or spaces allowed.
Table type
Native table
Schema
CREATE TABLE
CANCEL

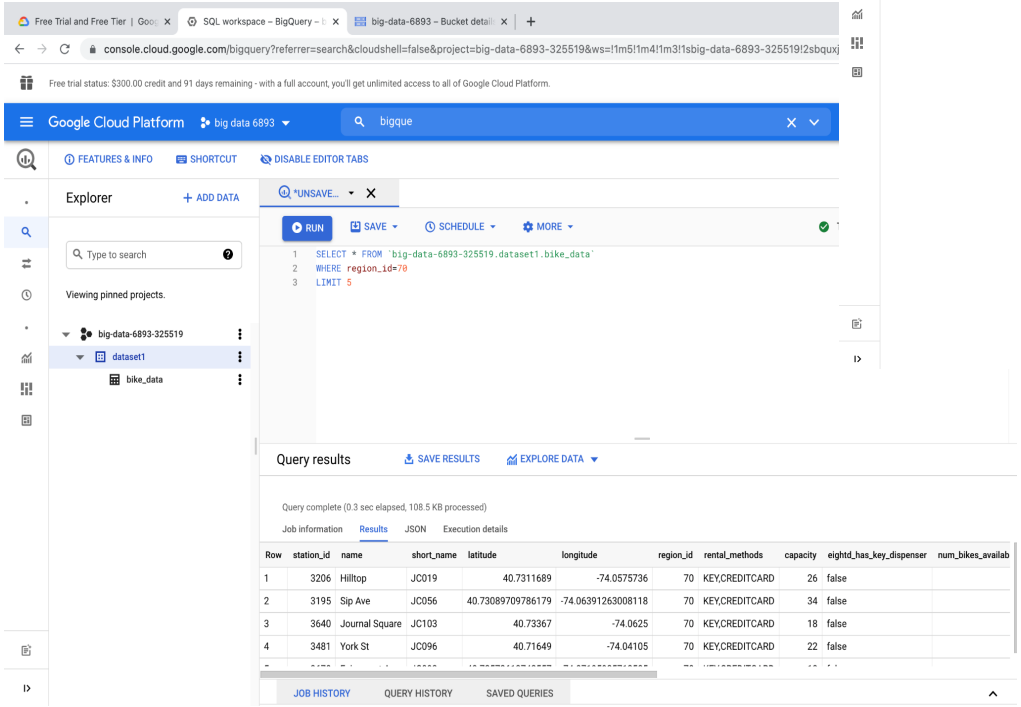

Choose a file
6893_ta
data_citibike_stations.csv
Filename
SELECT
CANCEL

BigQuery


Create table
Destination
Project *
eecs6893-398401
BROWSE
Dataset *
dataset1
Table *
bike_table
Unicode letters, marks, numbers, connectors, dashes or spaces allowed.
Table type
Native table
Schema
Auto detect
Schema will be automatically generated.
Partition and cluster settings
Partitioning
No partitioning
Clustering order
Clustering order determines the sort order of the data. Clustering can be used on both partitioned and non-partitioned tables.
CREATE TABLE
CANCEL

BigQuery


Free trial status: $300.00 credit and 91 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.
DISMISS
Google Cloud
big data 6893
Search bigqu
Explorer
ADD DATA
Type to search
Viewing pinned projects.
big-data-6893-362015
dataset1
bike_data
MORE RESULTS
bike_data
Editor 2
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


Free Trial and Free Tier | Goo
SQL workspace – BigQuery –
big-data-6893 – Bucket detail
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 40.730897097861 79 -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

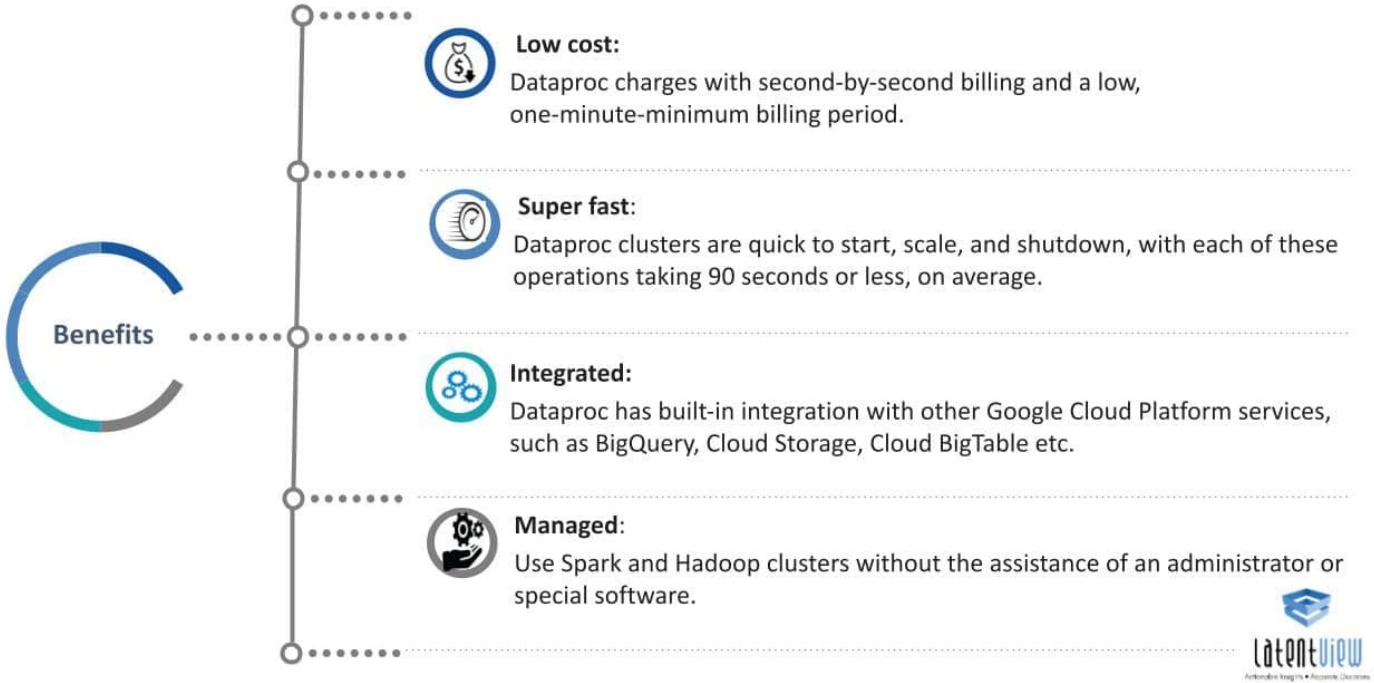
Dataproc

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

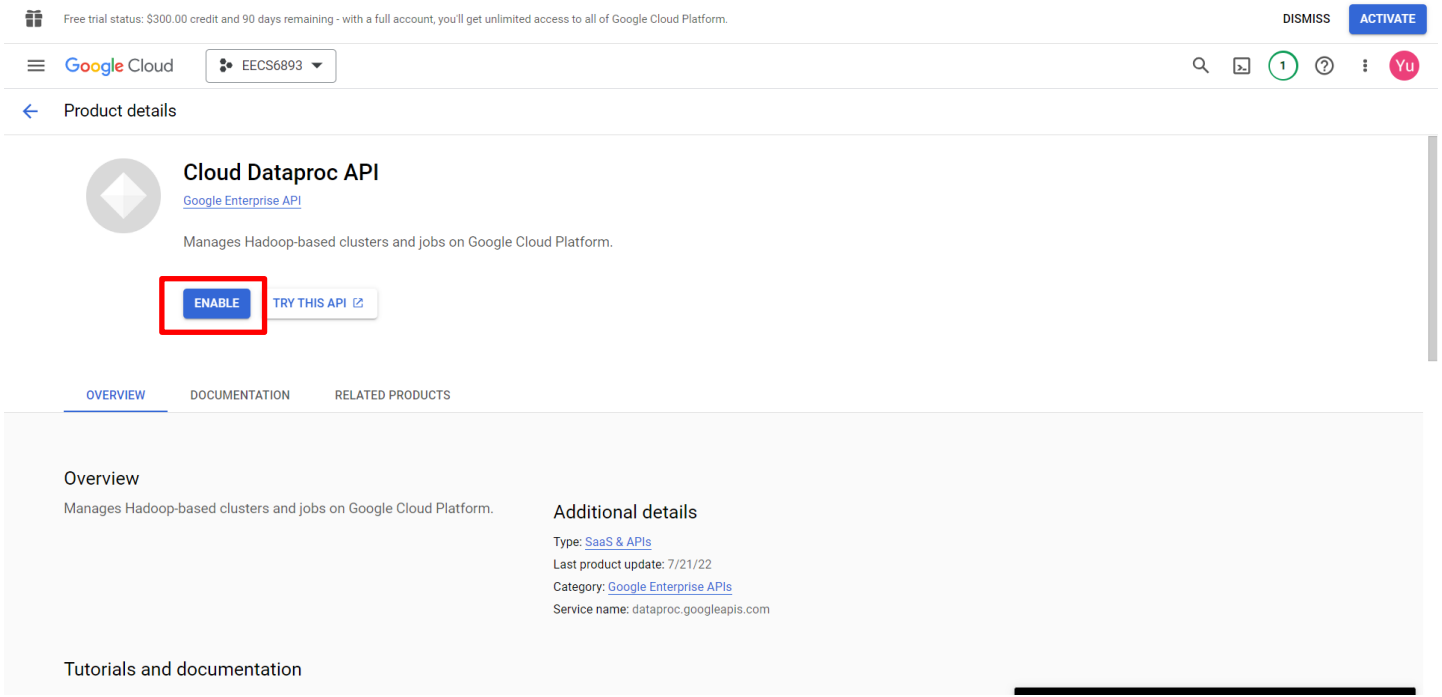
Dataproc

Why dataproc?

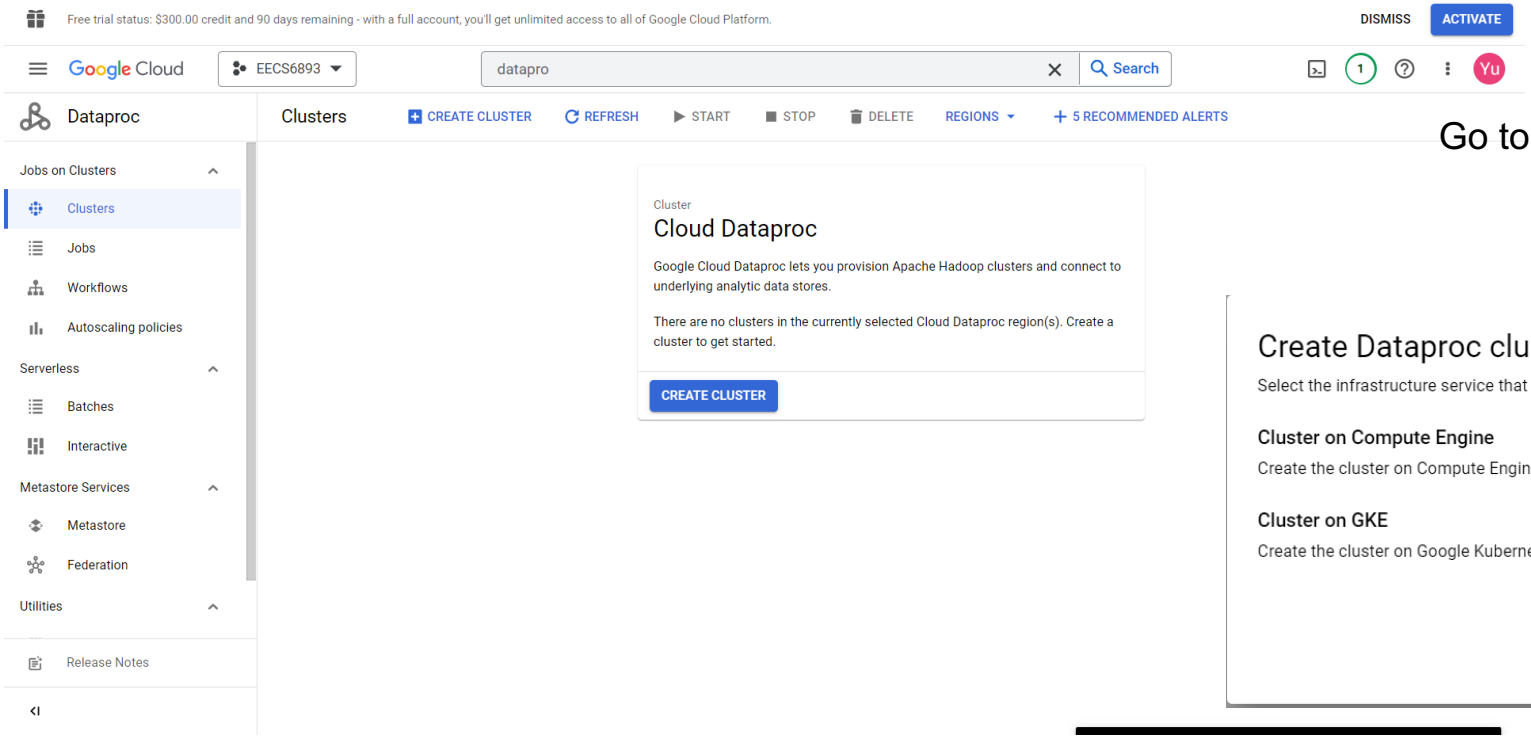


Dataproc

search cloud data proc
click on the API link



Dataproc - graphical UI



Go to Cloud Dataproc

Search cloud dataproc

Create a Dataproc cluster on Compute Engine

Set up cluster
Begin by providing basic information.

Configure nodes (optional)
Change node compute and storage capabilities.

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

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

ATE CANCEL

VALENT 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

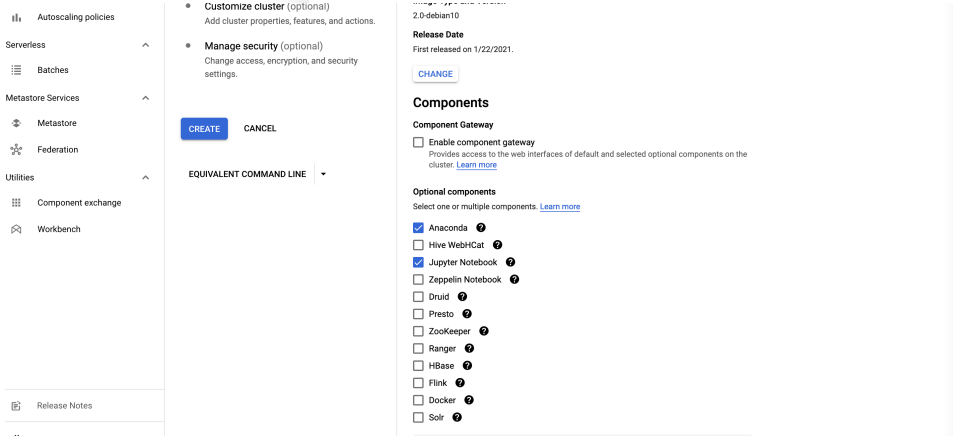
Autoscaling

Automates cluster resource management based on an autoscaling policy.

Policy
None

Enhanced Flexibility Mode

create cluster with Jupyter



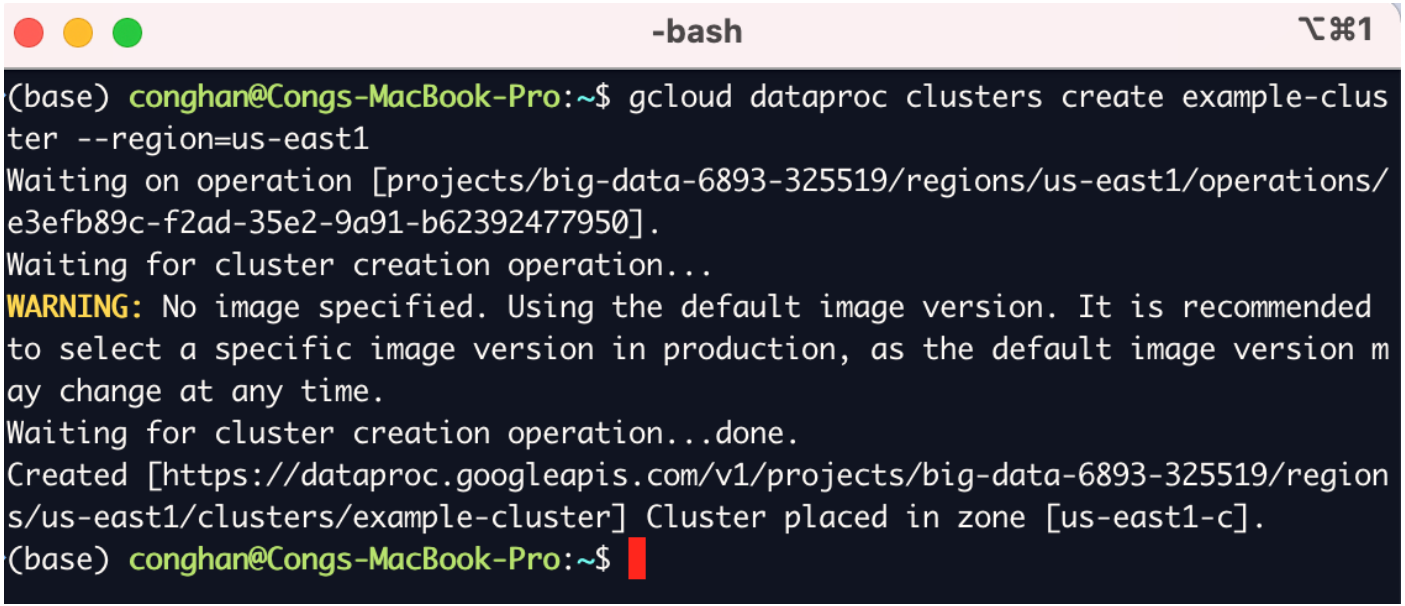
Dataproc - Cloud SDK

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@Congs-MacBook-Pro:~$ gcloud dataproc clusters create example-clus
ter --region=us-east1
```

Dataproc - Cloud SDK

Cluster creation (using Cloud SDK):



Dataproc - Cloud SDK

Submit a job - Pi calculation

```
(base) conghan@Congs-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.AHSProxy: 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.RMProxy: 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
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-ebe4f198110e
```

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

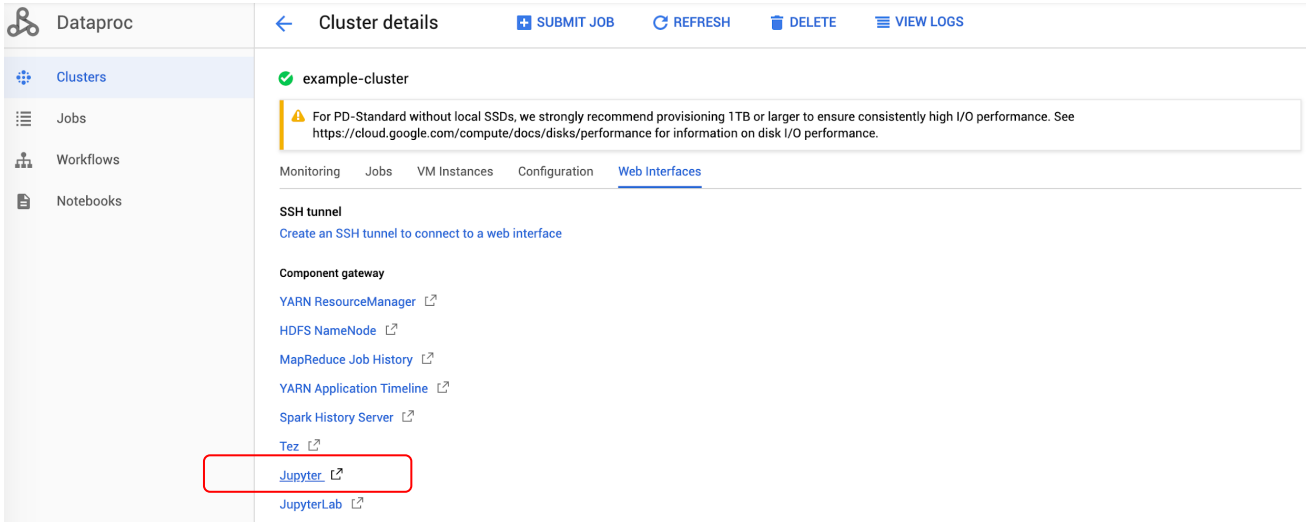
```
(base) conghan@Congs-MacBook-Pro:~$ gcloud beta dataproc clusters create example
-cluster --region=us-east1 --optional-components=ANACONDA,JUPYTER --image-versio
n=1.3 --enable-component-gateway --bucket big-data-6893 --project big-data-6893-
325519 --single-node --metadata 'PIP_PACKAGES=graphframes==0.6'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

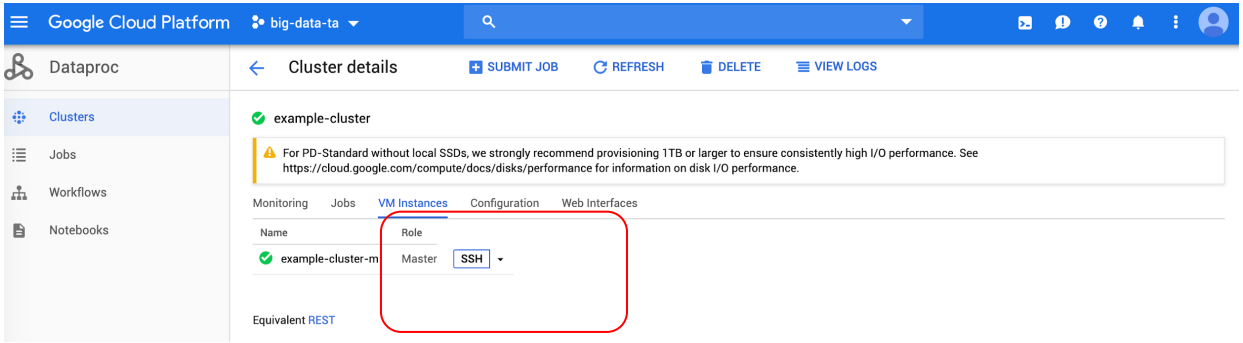


- Cloud SDK:
  - `gcloud dataproc jobs submit pyspark <your_program.py> --cluster=<cluster-name>`
  - View your jobs in console

- 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 fou
nd 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
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   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.