GCP Dataproc

Google GCP

Data Proc Cluster setup on Google Cloud Platform

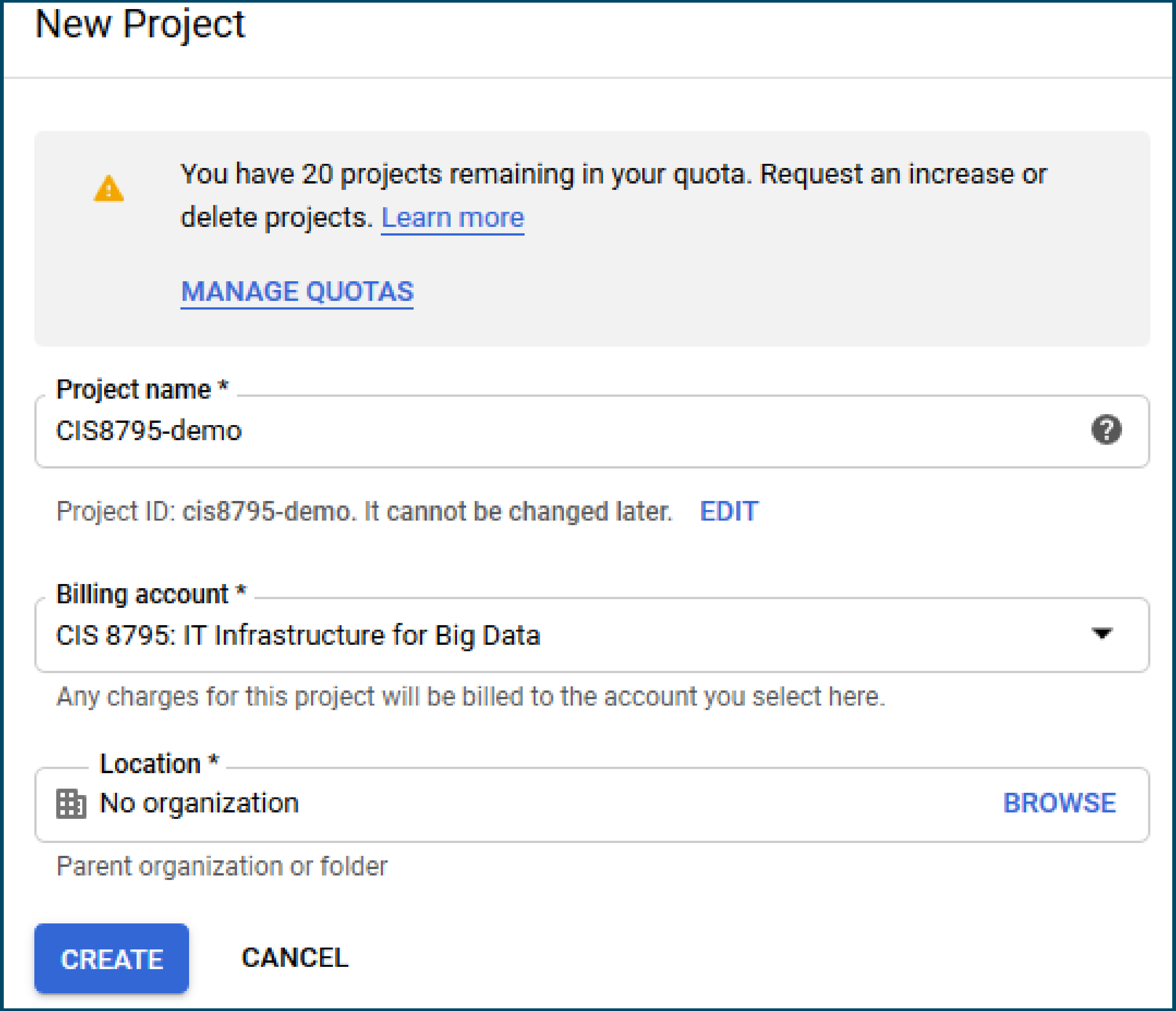
https://cloud.google.com/dataproc/docs/tutorials/jupyter-notebook

Signing up

- GCP offers $300 credit for new users.
- Navigate to this link to get started.
- Login with your Gmail ID to begin.
- Select “United States” as country and “Continue”.
- Select Account Type as “Individual” and fill in your personal details.
- You will be asked for credit card details to make sure you’re not a robot. **Google does not charge you even if you exhaust $300 credit.**
- You’re set if you reach “Getting Started” page.

Setting up the project

- Open the link given in the title slide and follow the instructions.
- For all the works you do in GCP, it is mandatory to select a project.
- Click the "Go to the project selector page" button to set up a project.
- Once you're in the dashboard, select "Create Project" on top right.
- Give a name to the project.
- Click "Create".



New Project

⚠ You have 20 projects remaining in your quota. Request an increase or delete projects. Learn more

MANAGE QUOTAS

Project name *
CIS8795-demo

Project ID: cis8795-demo. It cannot be changed later. EDIT

Billing account *
CIS 8795: IT Infrastructure for Big Data

Any charges for this project will be billed to the account you select here.

Location *
No organization BROWSE

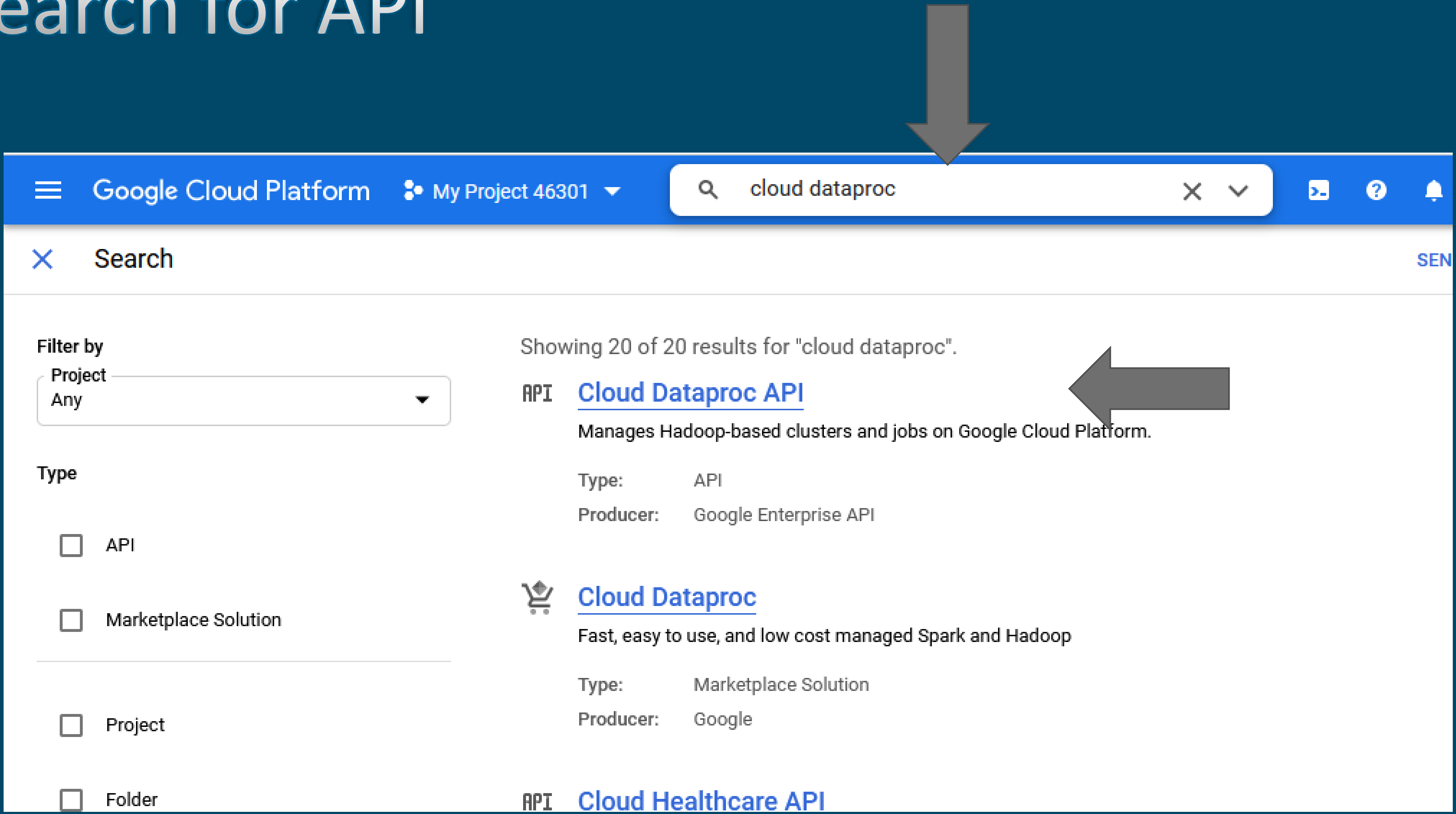
Parent organization or folder

CREATE CANCEL

Enable APIs

- Since we are working on Data Proc which uses the Compute Engine for underlying infrastructure, we need to enable the corresponding APIs before spinning up the cluster.
- Click on "Enable the APIs" button.
  - See the button on the https://cloud.google.com/dataproc/docs/tutorials/jupyter-notebook page
- Select the project you just created and hit "Continue".

Search for API



Filter by

Project: Any

Type

□ API

□ Marketplace Solution

□ Project

□ Folder

Showing 20 of 20 results for "cloud dataproc".

API **Cloud Dataproc API**

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

Type: API

Producer: Google Enterprise API

Cloud Dataproc

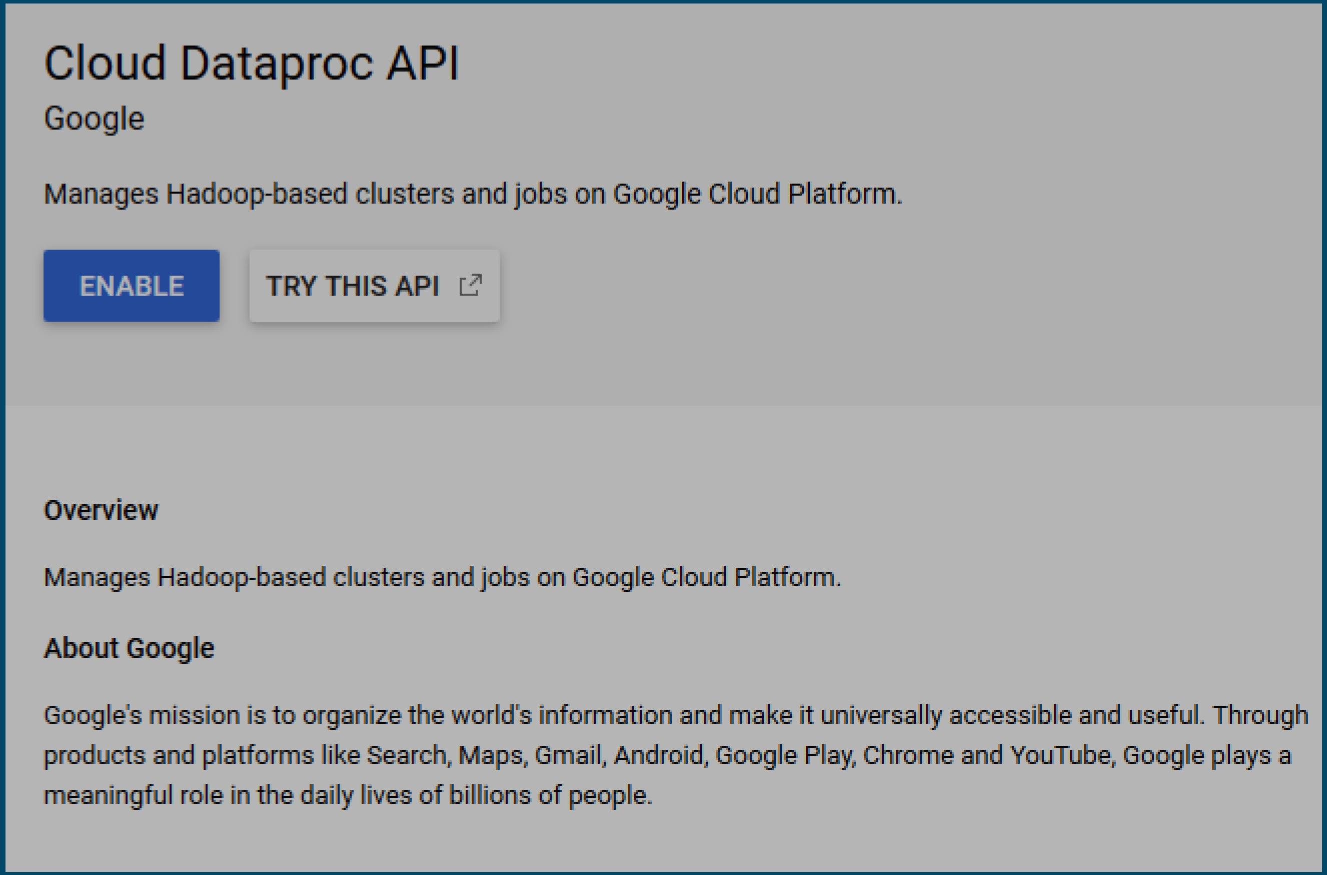
Fast, easy to use, and low cost managed Spark and Hadoop

Type: Marketplace Solution

Producer: Google

API **Cloud Healthcare API**

Search for: Cloud Dataproc API
select enable (ensure your project is selected)



Cloud Dataproc API

Google

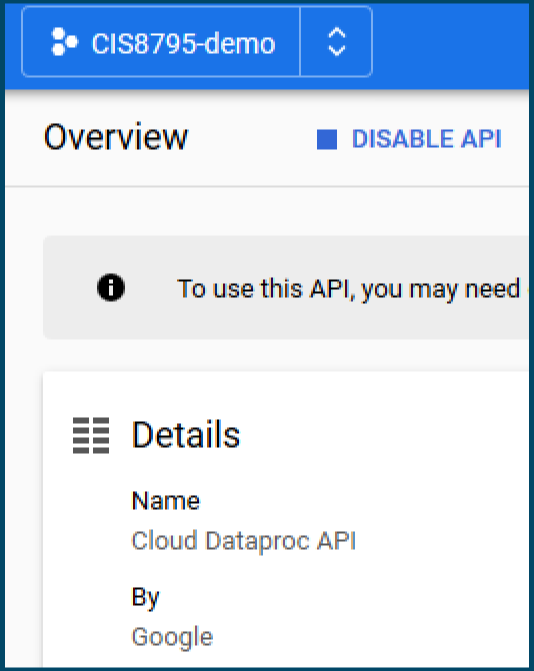
Manages Hadoop-based clusters and jobs on Google Cloud Platform.

Overview

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

About Google

Google's mission is to organize the world's information and make it universally accessible and useful. Through products and platforms like Search, Maps, Gmail, Android, Google Play, Chrome and YouTube, Google plays a meaningful role in the daily lives of billions of people.

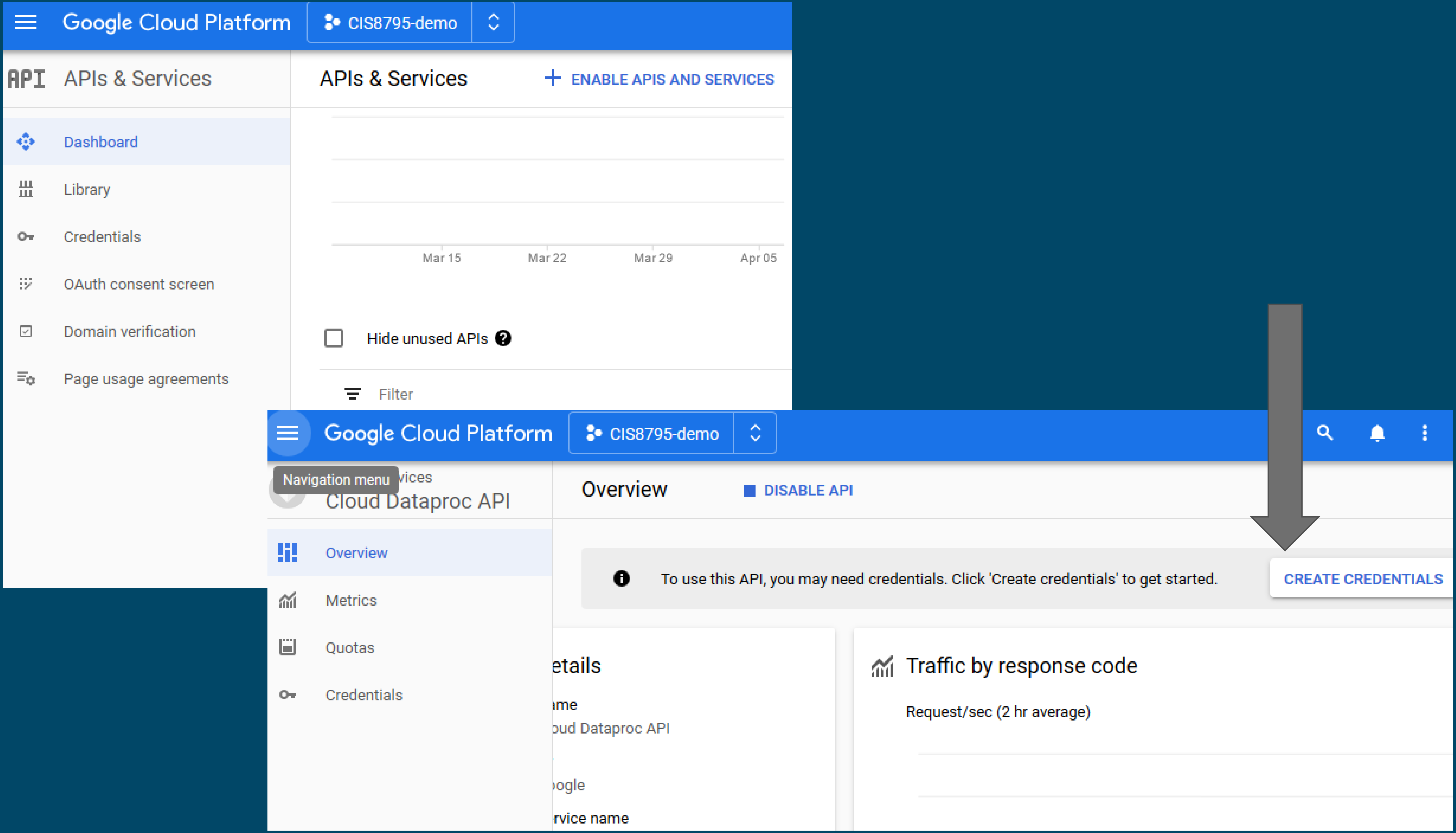


To use this API, you may need

Details

Name
Cloud Dataproc API

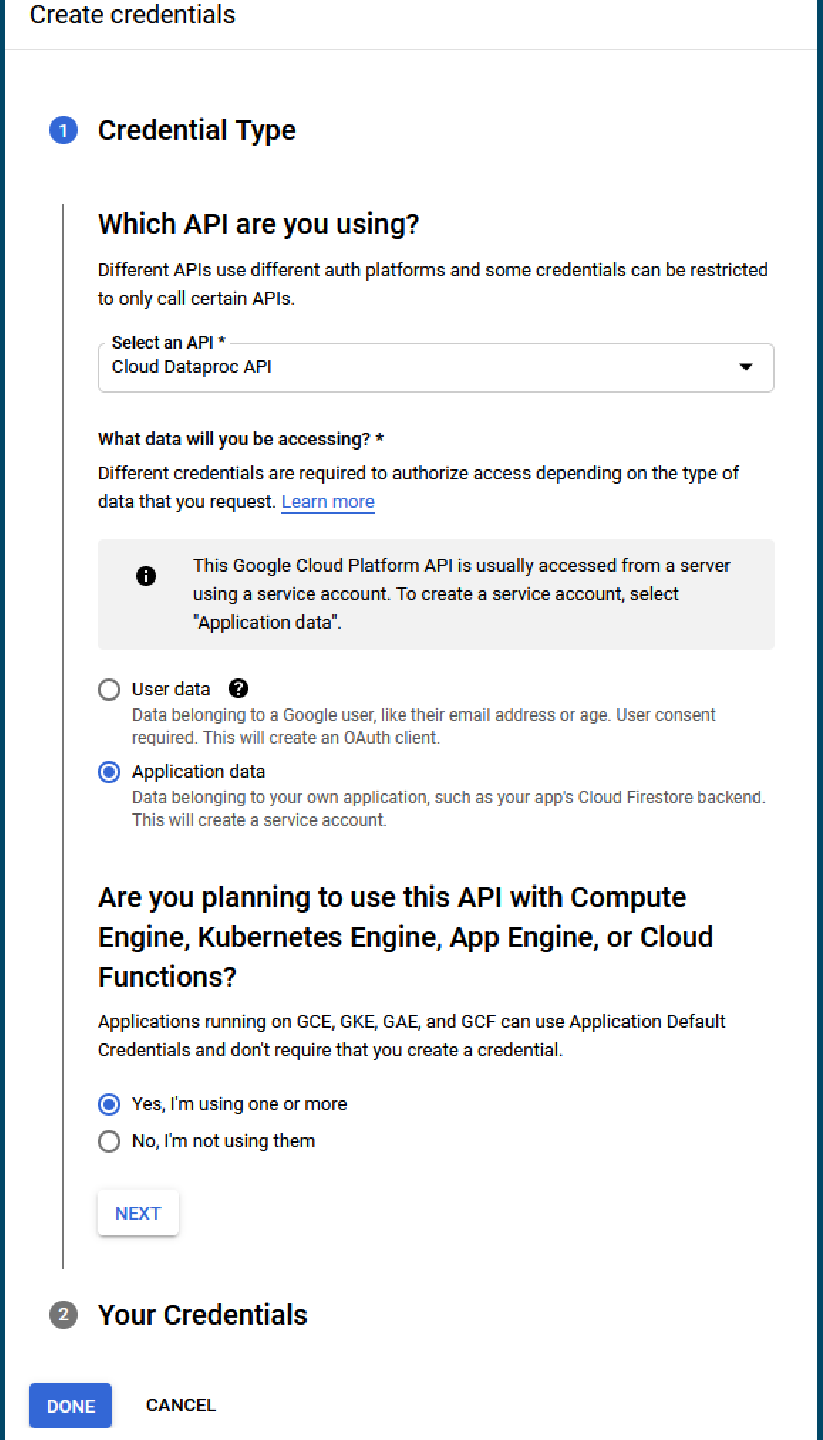
By
Google



To use this API, you may need credentials. Click 'Create credentials' to get started.

Establishing the Credentials

- Once the APIs are enabled, you need to set your credentials by clicking "Create Credentials".
    - You'll need to wait for the API's to be enabled
    - API & Services -> Credentials
- Select "Application Data" and "Yes, I'm using one or more"
- And click "Done"
- Once credentials are established, Click on "CONFIGURE CONSENT SCREEN"
    - External
- Give a name for the application, and click on "Save"



Create credentials

1 Credential Type

Which API are you using?

Different APIs use different auth platforms and some credentials can be restricted to only call certain APIs.

Select an API *
Cloud Dataproc API

What data will you be accessing? *

Different credentials are required to authorize access depending on the type of data that you request. Learn more

This Google Cloud Platform API is usually accessed from a server using a service account. To create a service account, select "Application data".

○ User data ?
Data belonging to a Google user, like their email address or age. User consent required. This will create an OAuth client.

● Application data
Data belonging to your own application, such as your app's Cloud Firestore backend. This will create a service account.

Are you planning to use this API with Compute Engine, Kubernetes Engine, App Engine, or Cloud Functions?

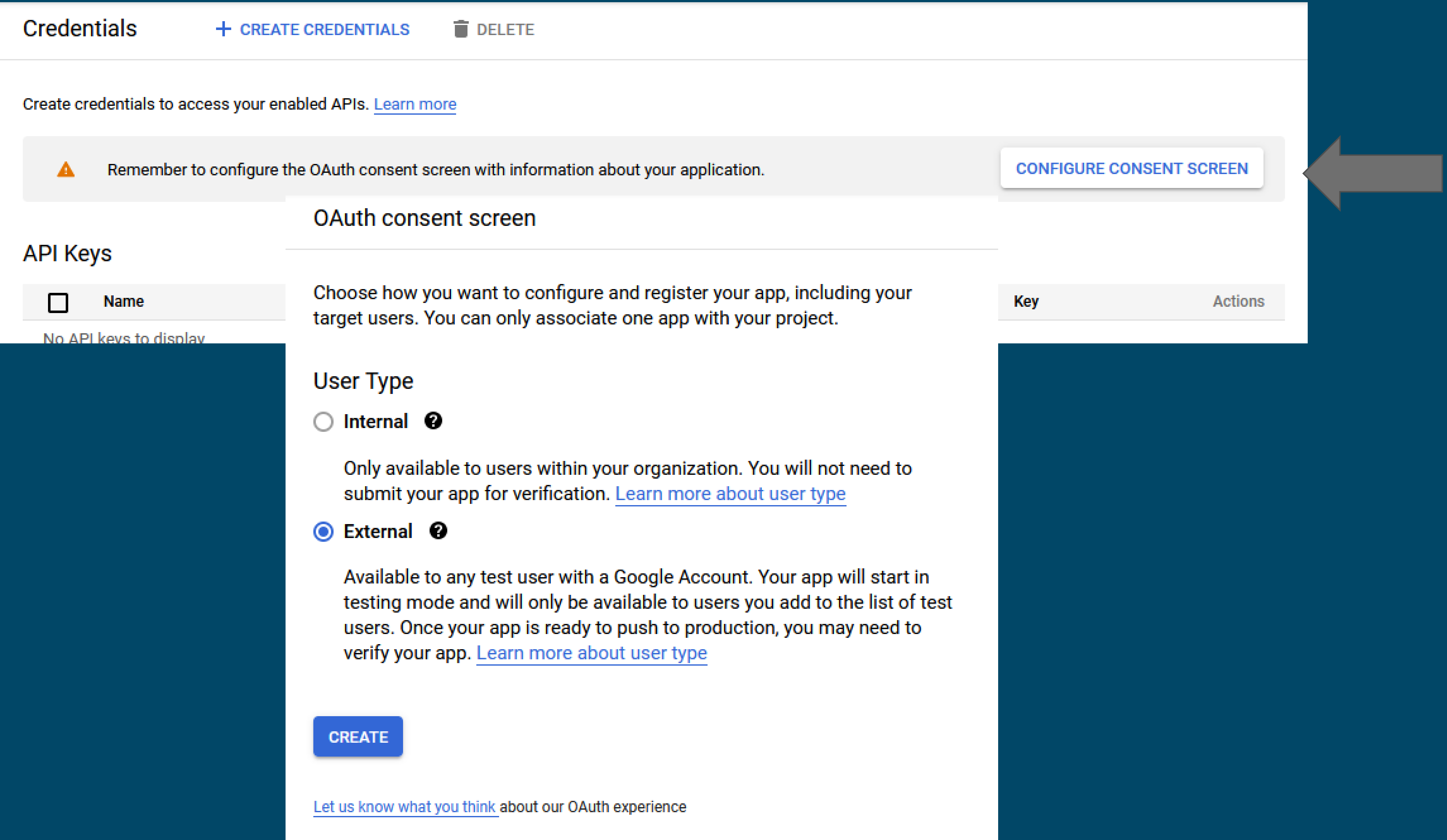
Applications running on GCE, GKE, GAE, and GCF can use Application Default Credentials and don't require that you create a credential.

● Yes, I'm using one or more

○ No, I'm not using them

2 Your Credentials

2021 style credentials



Credentials

Create credentials to access your enabled APIs. Learn more

⚠ Remember to configure the OAuth consent screen with information about your application.

API Keys

□	Name	Key	Actions
No API keys to display			

OAuth consent screen

Choose how you want to configure and register your app, including your target users. You can only associate one app with your project.

User Type

○ Internal

Only available to users within your organization. You will not need to submit your app for verification. Learn more about user type

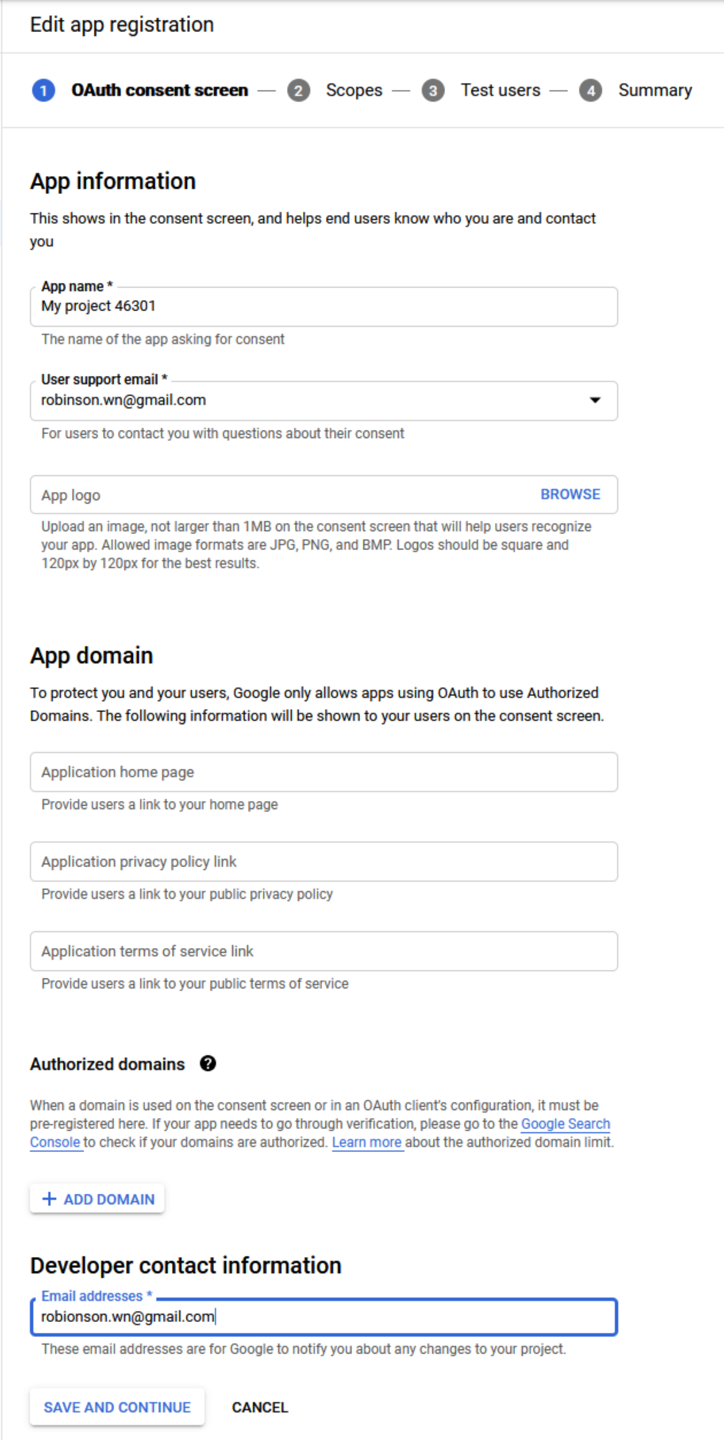
● External

Available to any test user with a Google Account. Your app will start in testing mode and will only be available to users you add to the list of test users. Once your app is ready to push to production, you may need to verify your app. Learn more about user type

Let us know what you think about our OAuth experience

Application registration

- Give app name
- Email for
  - User support
  - Developer contact
- Save
- Go to Home Dashboard



Edit app registration
1 OAuth consent screen — 2 Scopes — 3 Test users — 4 Summary

App information

This shows in the consent screen, and helps end users know who you are and contact you

App name *
My project 46301
The name of the app asking for consent

User support email *
robinson.wn@gmail.com
For users to contact you with questions about their consent

App logo BROWSE
Upload an image, not larger than 1MB on the consent screen that will help users recognize your app. Allowed image formats are JPG, PNG, and BMP. Logos should be square and 120px by 120px for the best results.

App domain

To protect you and your users, Google only allows apps using OAuth to use Authorized Domains. The following information will be shown to your users on the consent screen.

Application home page
Provide users a link to your home page

Application privacy policy link
Provide users a link to your public privacy policy

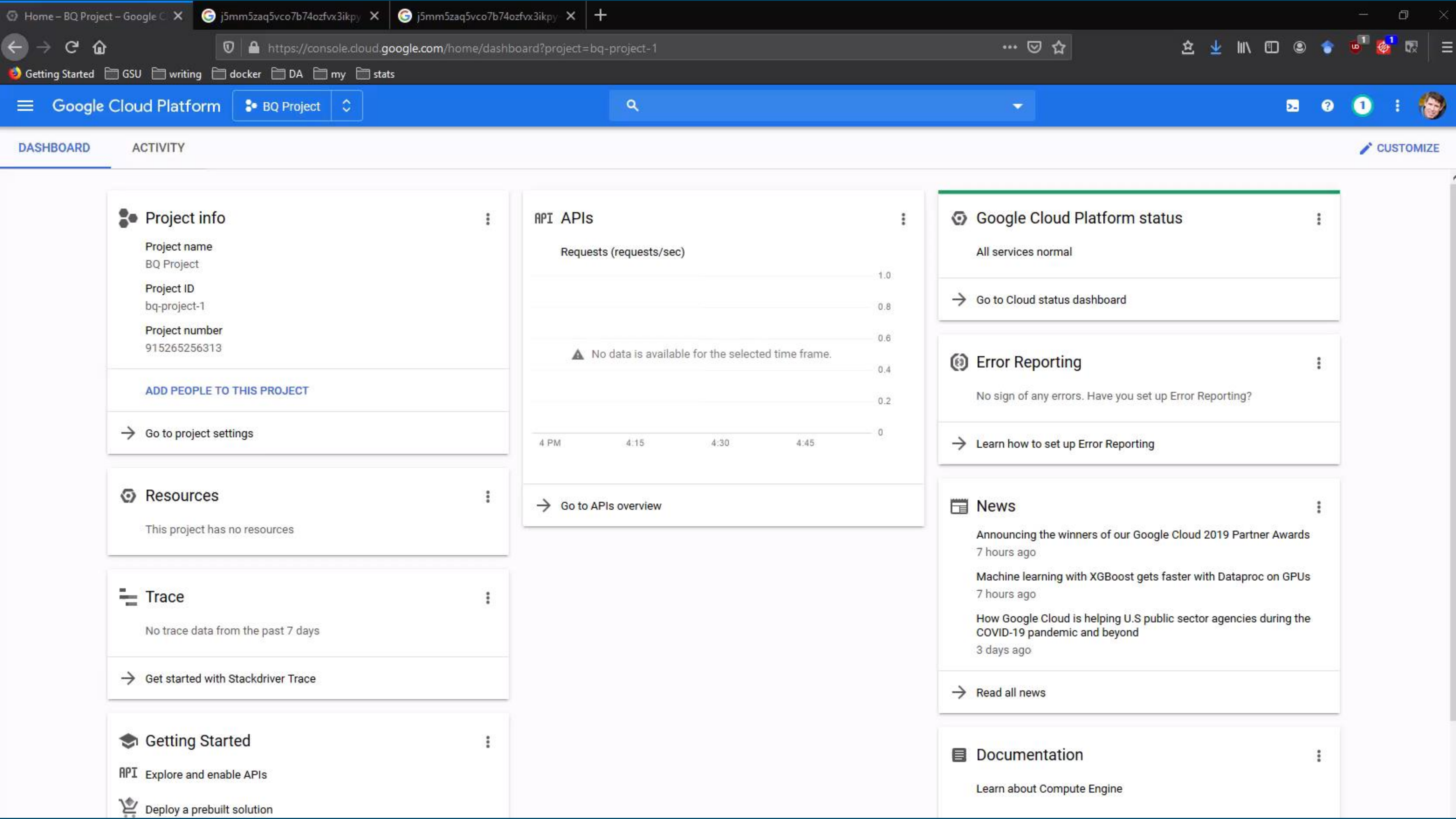
Application terms of service link
Provide users a link to your public terms of service

Authorized domains

When a domain is used on the consent screen or in an OAuth client's configuration, it must be pre-registered here. If your app needs to go through verification, please go to the Google Search Console to check if your domains are authorized. Learn more about the authorized domain limit.

Developer contact information

Email addresses *
robionson.wn@gmail.com
These email addresses are for Google to notify you about any changes to your project.


Home – BQ Project – Google C
j5mm5zaq5vco7b74ozfvx3ikpy
j5mm5zaq5vco7b74ozfvx3ikpy
https://console.cloud.google.com/home/dashboard?project=bq-project-1
Getting Started
GSU
writing
docker
DA
my
stats
Google Cloud Platform
BQ Project
DASHBOARD
ACTIVITY
CUSTOMIZE

Project info

Project name
BQ Project

Project ID
bq-project-1

Project number
915265256313

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 Stackdriver Trace

Getting Started

API Explore and enable APIs

Deploy a prebuilt solution

API APIs

Requests (requests/sec)

1.0
0.8
0.6
0.4
0.2
0
4 PM
4:15
4:30
4:45

No data is available for the selected time frame.

→ Go to APIs overview

Google Cloud Platform status

All services normal

→ Go to Cloud status dashboard

Error Reporting

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

→ Learn how to set up Error Reporting

News

Announcing the winners of our Google Cloud 2019 Partner Awards
7 hours ago

Machine learning with XGBoost gets faster with Dataproc on GPUs
7 hours ago

How Google Cloud is helping U.S public sector agencies during the COVID-19 pandemic and beyond
3 days ago

→ Read all news

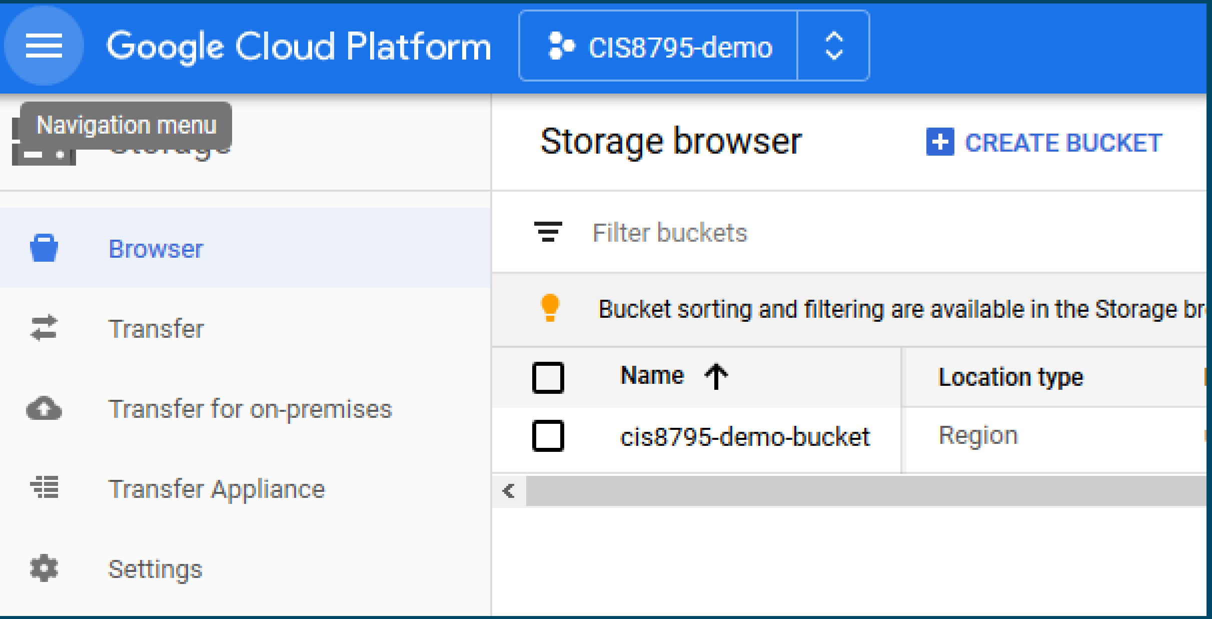
Documentation

Learn about Compute Engine

Create a Storage Bucket

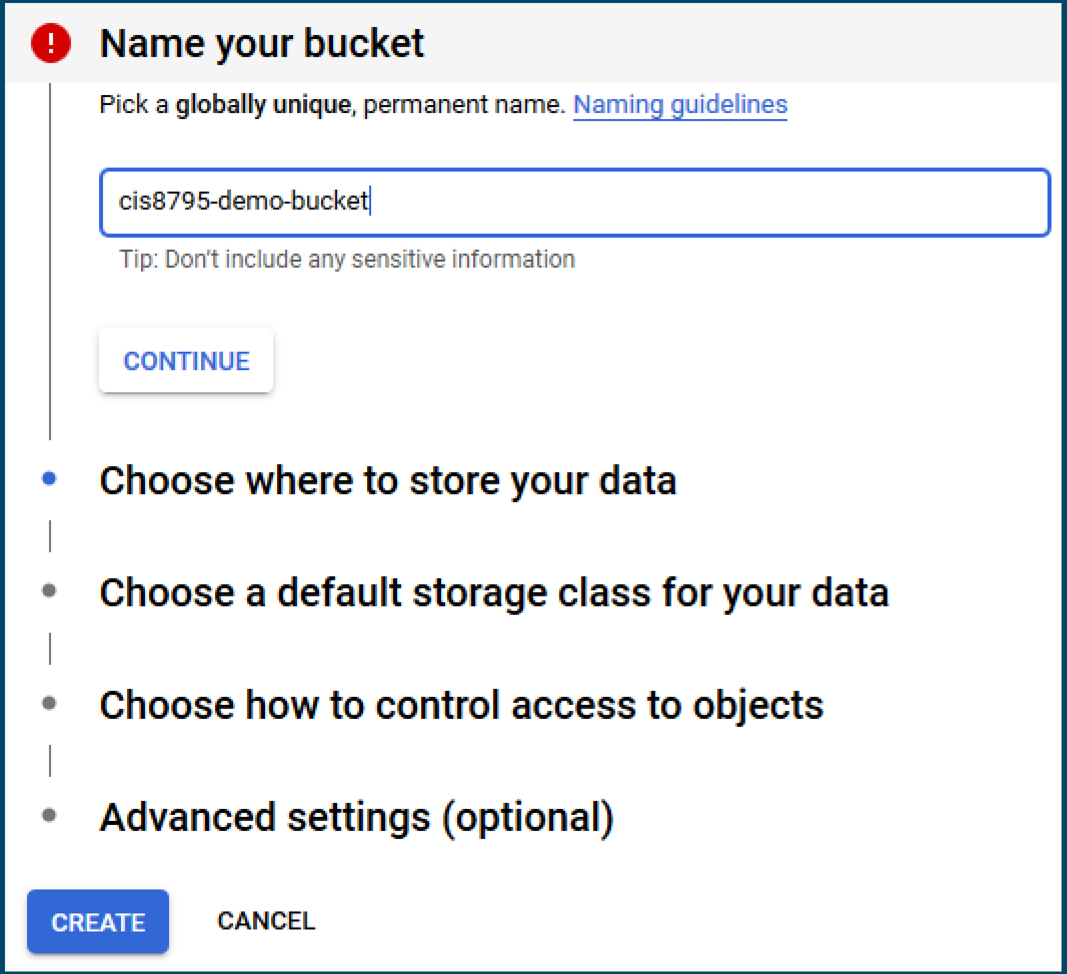
- Click on Go to the Cloud Storage Browser page
- Click on CREATE BUCKET
- Give your bucket a name.
- Select Region and hit Continue
- Select Standard as default storage class and hit Continue.
- Select Fine-grained access control and hit Continue.
- Select Google-managed key and hit Continue.
- Navigate to the bucket and click on "Upload files" and upload the .csv and .ipynb files.

Create a bucket for your project



Bucket sorting and filtering are available in the Storage br

□	Name ↑	Location type
□	cis8795-demo-bucket	Region



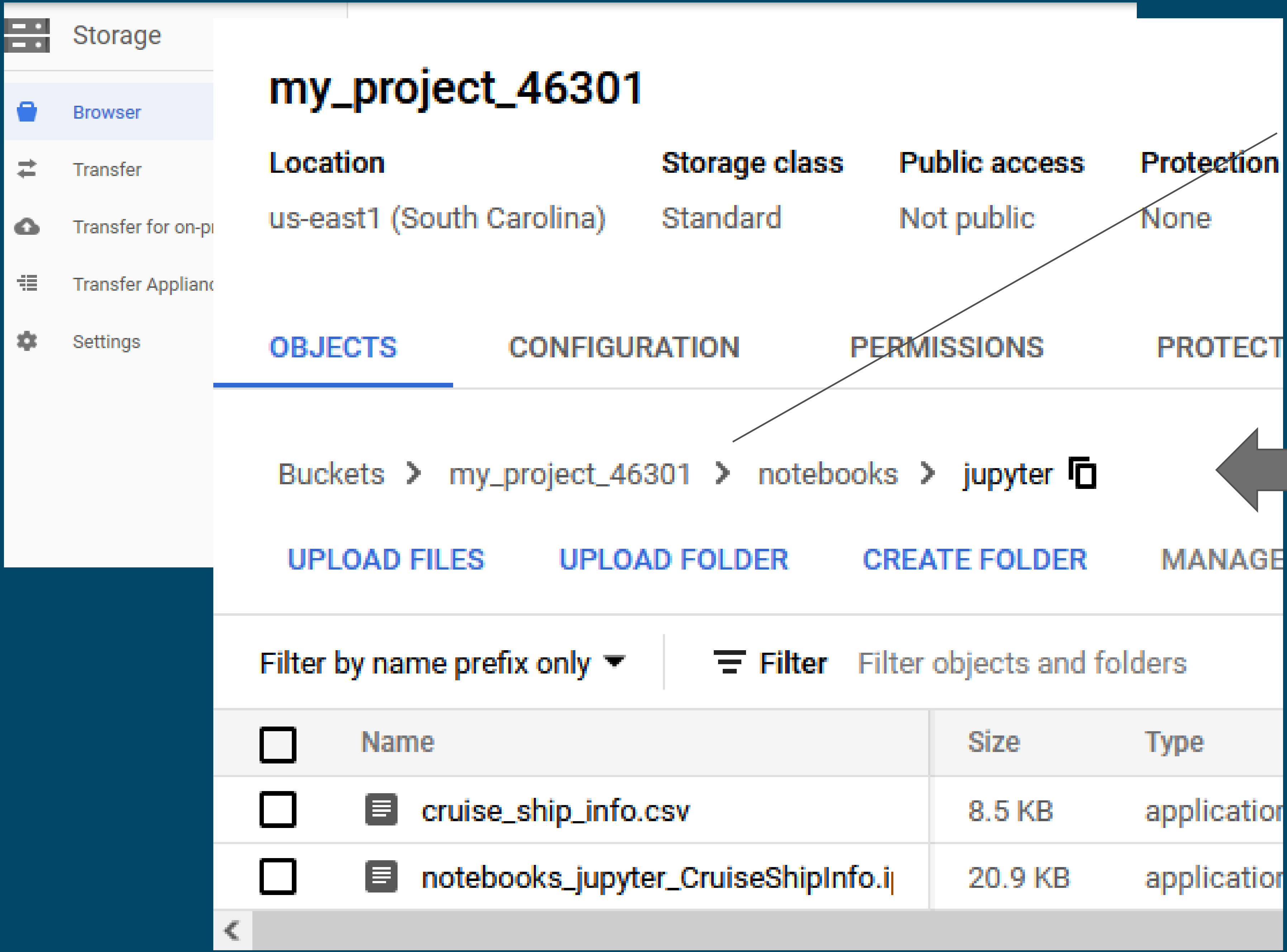
Name your bucket

Pick a **globally unique**, permanent name. Naming guidelines

cis8795-demo-bucket

Tip: Don't include any sensitive information

- Choose where to store your data
- Choose a default storage class for your data
- Choose how to control access to objects
- Advanced settings (optional)



my_project_46301

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

Buckets > my_project_46301 > notebooks > jupyter

□	Name	Size	Type
□	cruise_ship_info.csv	8.5 KB	applicatio
□	notebooks_jupyter_CruiseShipInfo.i	20.9 KB	applicatio

Upload notebooks & data after cluster creation

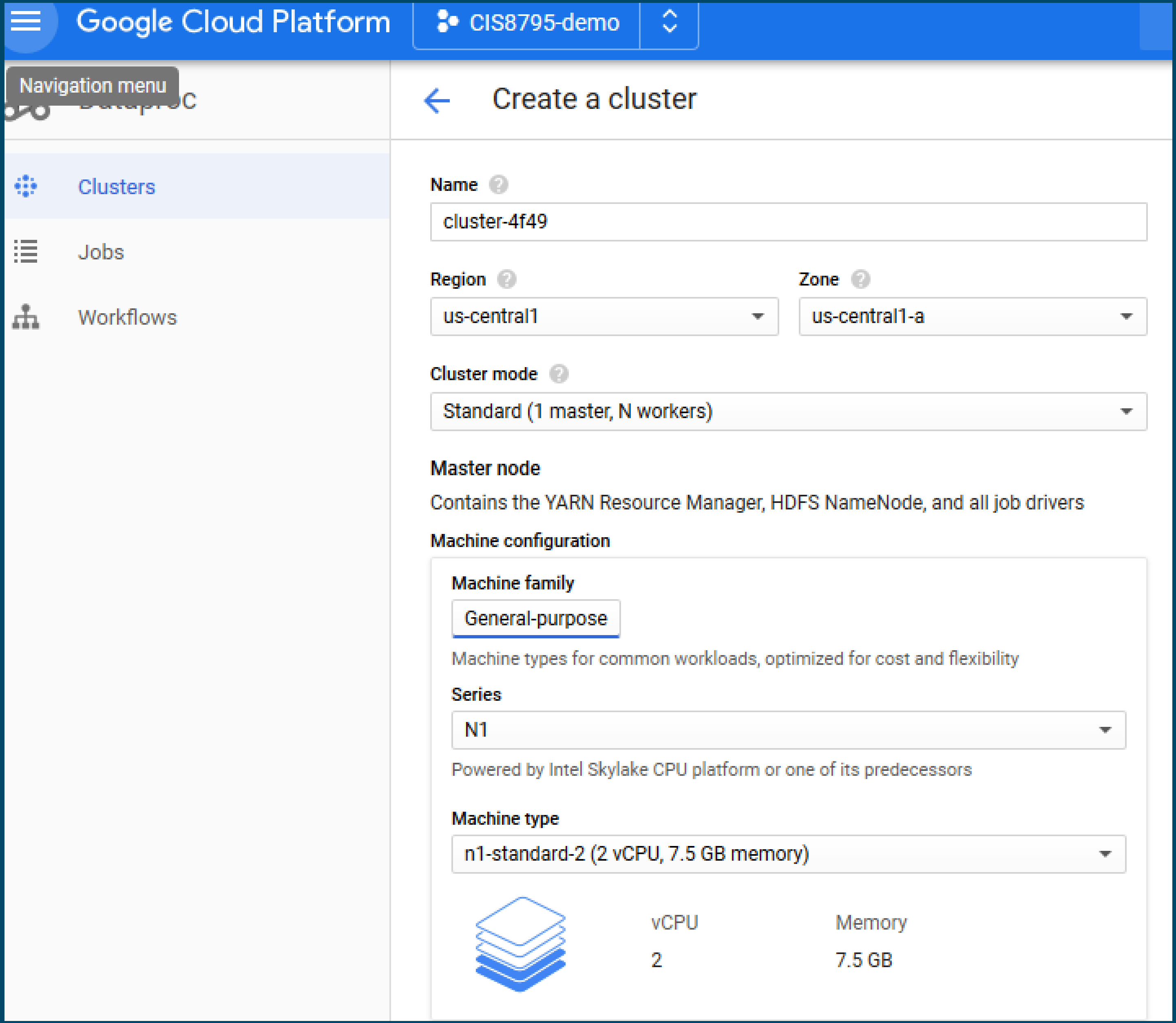
Path
project/notebooks/jupyter

Optional:
Install and Initialize the Cloud SDK

- Click on "Install and initialize the Cloud SDK".
- Choose your operating system and make sure Python is installed.
- Download the SDK file.
- And run the "gcloud init" command on your terminal.
- Select option 1 - Re-initialize this configuration [default] with new settings
- Choose your email ID.
- Choose your project.
- Choose No for default region.

Creating Cluster

- Click on "Cloud Console" and choose "Dataproc"
- Select "Clusters" and "Create cluster"
- Set up Cluster
  - Under "Component gateway", Enable (for web access)
  - Under Optional components, Select "Jupyter notebook"
- Configure nodes
  - Master Node to have 2 CPUs
  - Worker Node to have 2 CPUs
- Customize cluster
  - Under "Cloud Storage staging bucket" configure the bucket as your default storage option.
- Click "Create"



Create a cluster

Name
cluster-4f49

Region
us-central1

Zone
us-central1-a

Cluster mode
Standard (1 master, N workers)

Master node

Contains the YARN Resource Manager, HDFS NameNode, and all job drivers

Machine configuration

Machine family
General-purpose

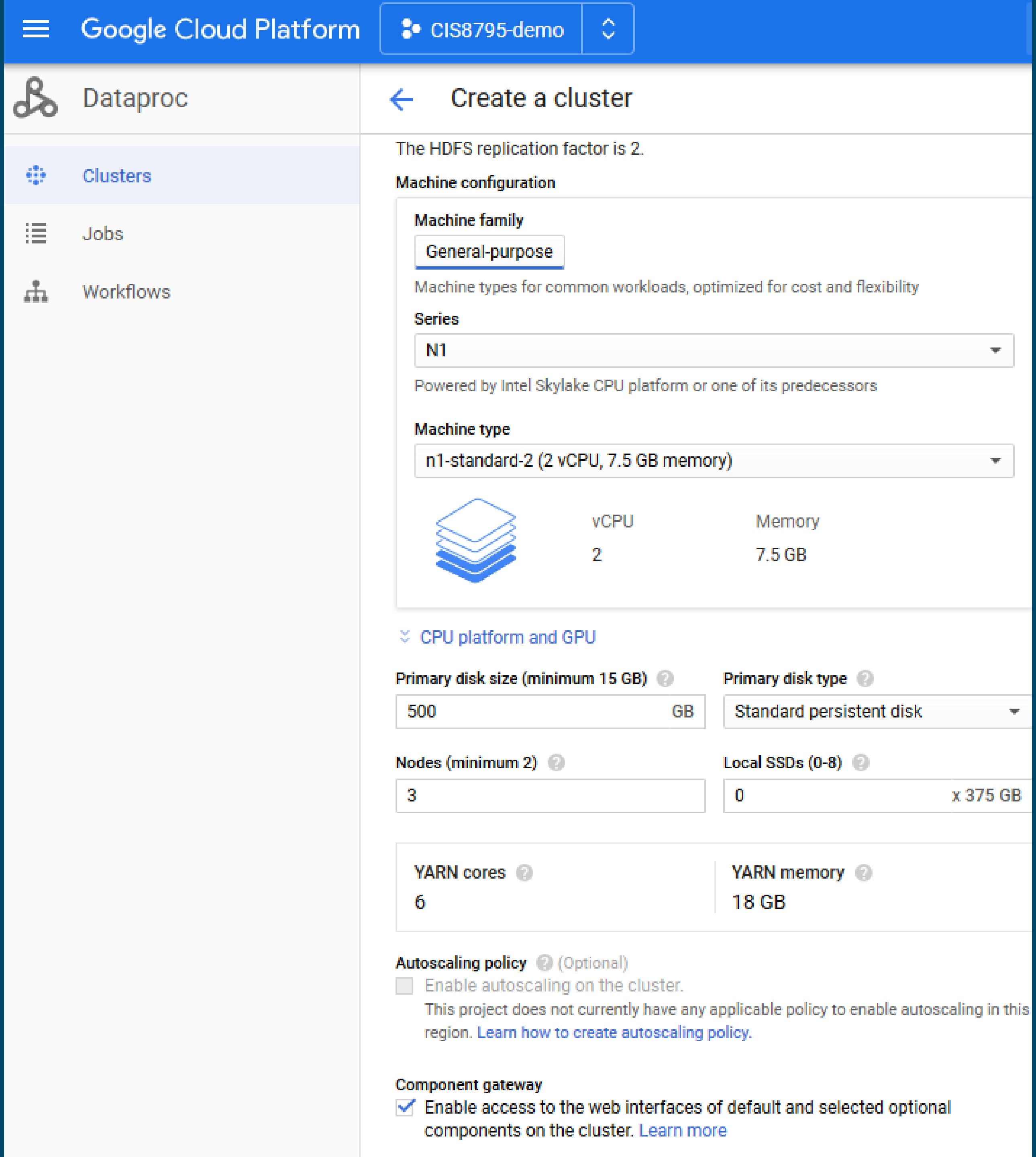
Machine types for common workloads, optimized for cost and flexibility

Series
N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type
n1-standard-2 (2 vCPU, 7.5 GB memory)

vCPU	Memory
2	7.5 GB



Create a cluster

The HDFS replication factor is 2.

Machine configuration

Machine family

General-purpose

Machine types for common workloads, optimized for cost and flexibility

Series

N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type

n1-standard-2 (2 vCPU, 7.5 GB memory)

vCPU	Memory
2	7.5 GB

Primary disk size (minimum 15 GB): 500 GB

Primary disk type: Standard persistent disk

Nodes (minimum 2): 3

Local SSDs (0-8): 0 x 375 GB

YARN cores	YARN memory
6	18 GB

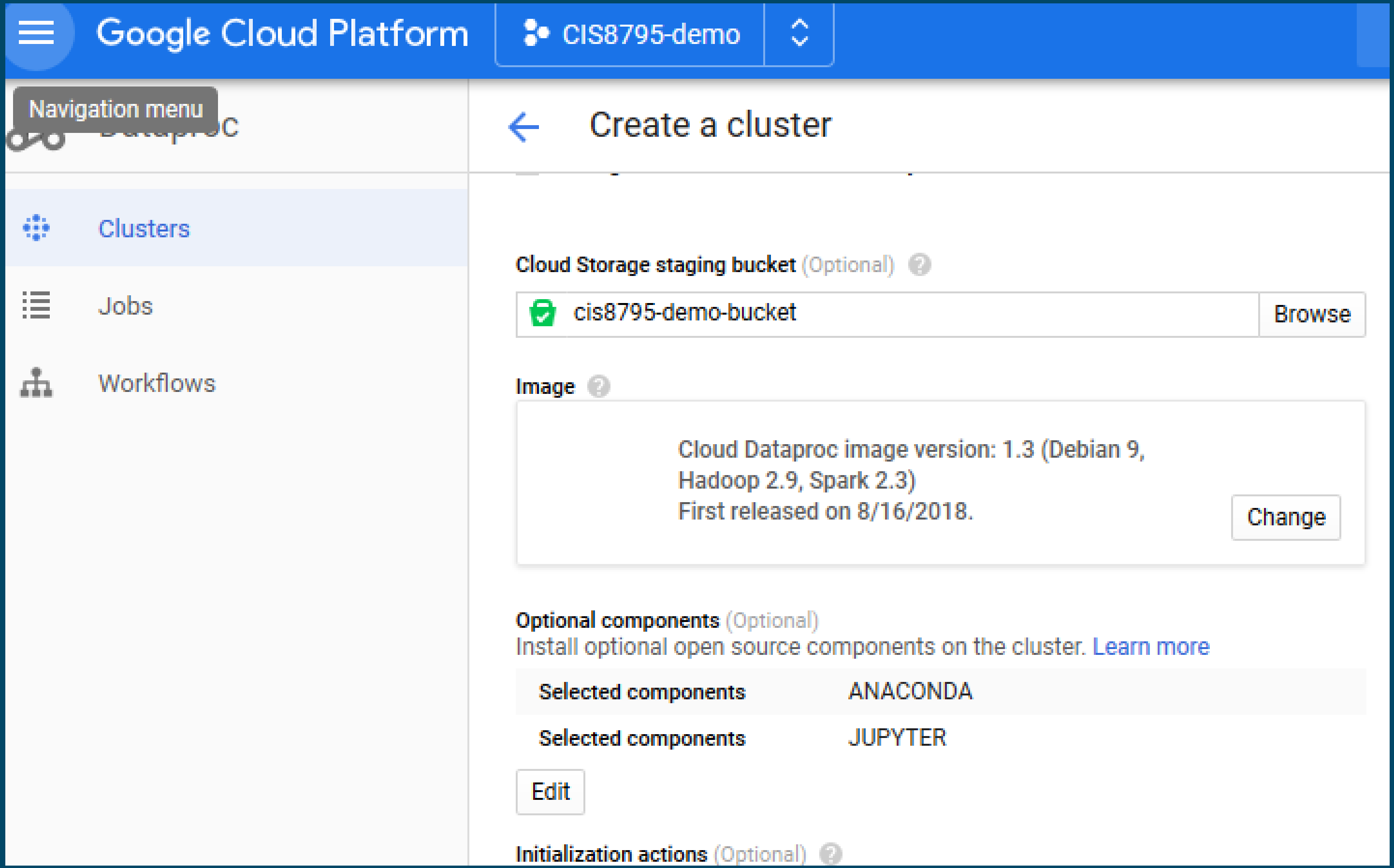
Autoscaling policy (Optional)

□ Enable autoscaling on the cluster.

This project does not currently have any applicable policy to enable autoscaling in this region. Learn how to create autoscaling policy.

Component gateway

☑ Enable access to the web interfaces of default and selected optional components on the cluster. Learn more



Create a cluster

Cloud Storage staging bucket (Optional)

cis8795-demo-bucket

Image

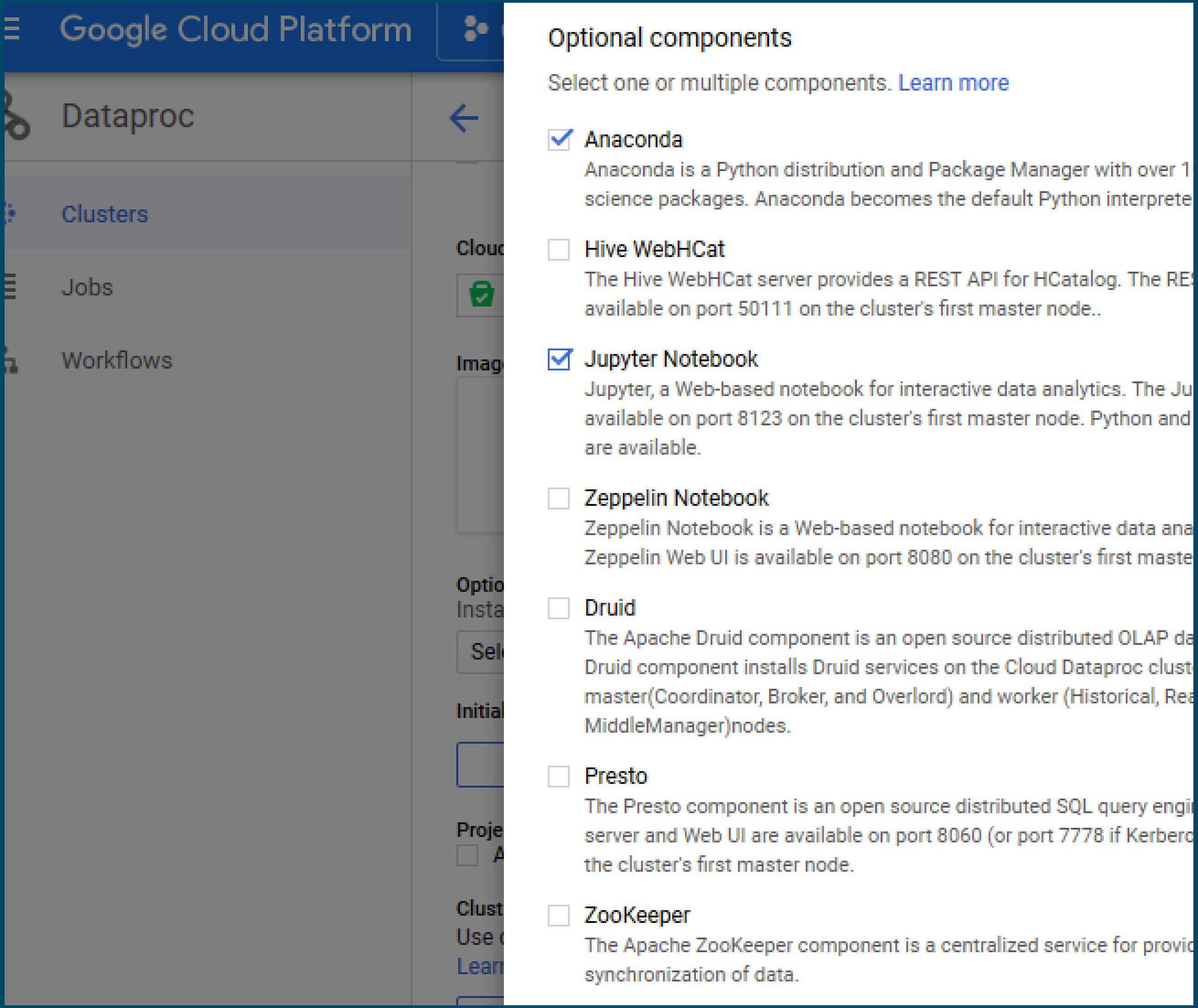
Cloud Dataproc image version: 1.3 (Debian 9, Hadoop 2.9, Spark 2.3)
First released on 8/16/2018.

Optional components (Optional)

Install optional open source components on the cluster. Learn more

Selected components	ANACONDA
Selected components	JUPYTER

Initialization actions (Optional)



Optional components

Select one or multiple components. Learn more

☑ **Anaconda**
Anaconda is a Python distribution and Package Manager with over 1
science packages. Anaconda becomes the default Python interprete

□ **Hive WebHCat**
The Hive WebHCat server provides a REST API for HCatalog. The RE
available on port 50111 on the cluster's first master node..

☑ **Jupyter Notebook**
Jupyter, a Web-based notebook for interactive data analytics. The Ju
available on port 8123 on the cluster's first master node. Python and
are available.

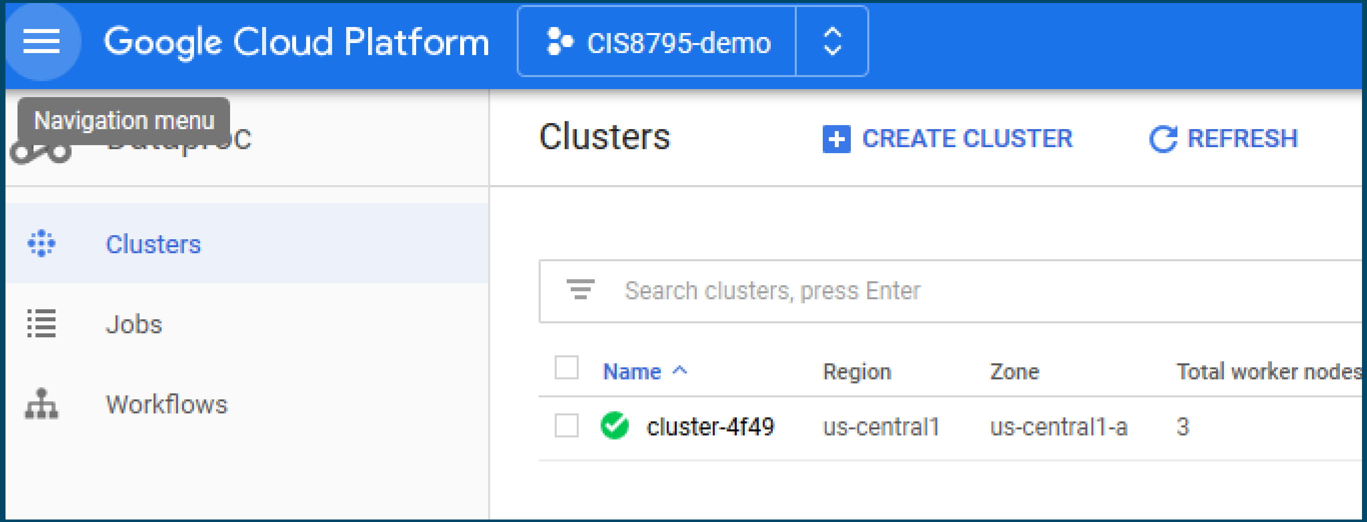
□ **Zeppelin Notebook**
Zeppelin Notebook is a Web-based notebook for interactive data ana
Zeppelin Web UI is available on port 8080 on the cluster's first maste

□ **Druid**
The Apache Druid component is an open source distributed OLAP da
Druid component installs Druid services on the Cloud Dataproc clust
master(Coordinator, Broker, and Overlord) and worker (Historical, Rea
MiddleManager)nodes.

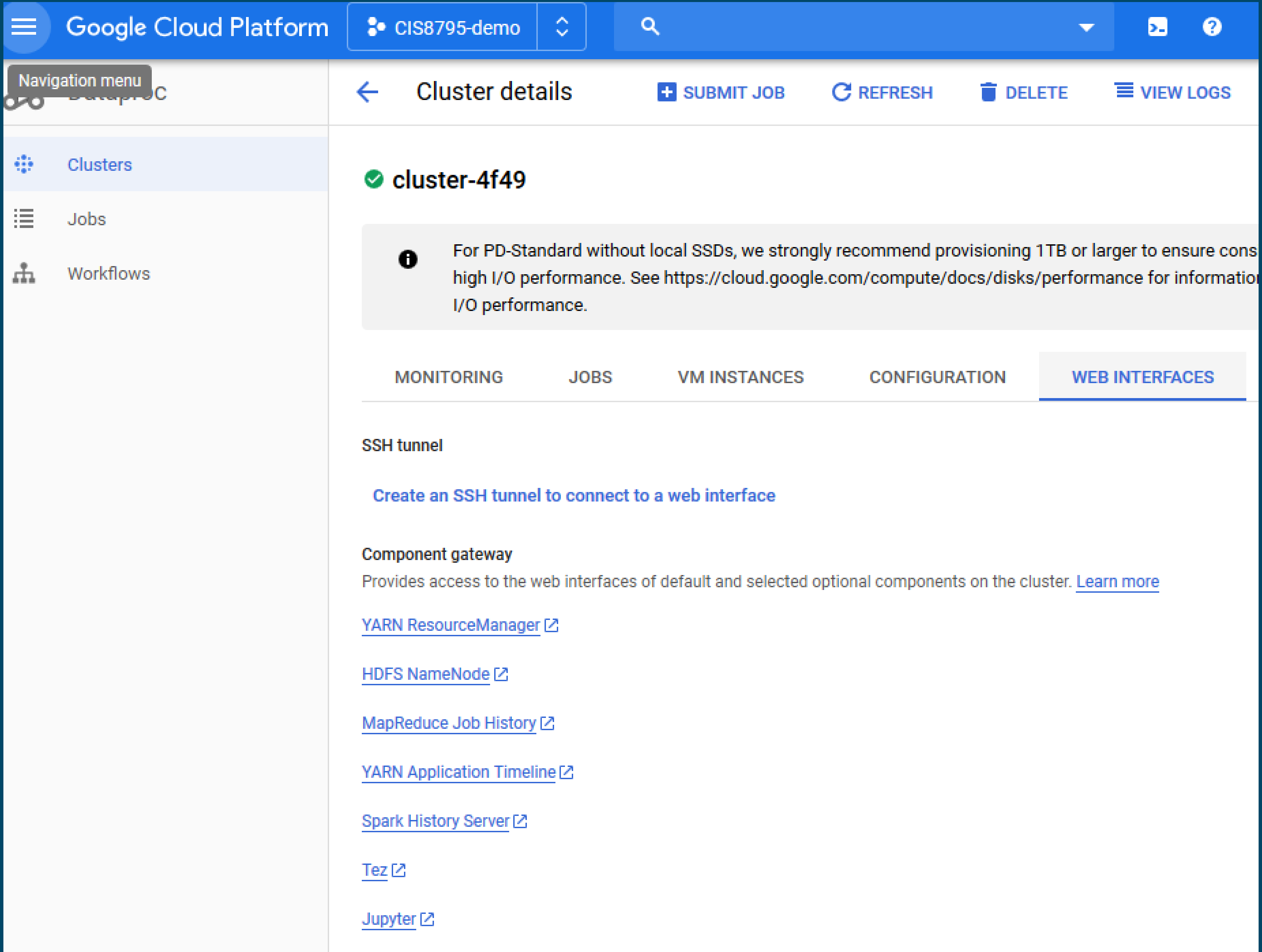
□ **Presto**
The Presto component is an open source distributed SQL query engi
server and Web UI are available on port 8060 (or port 7778 if Kerbero
the cluster's first master node.

□ **ZooKeeper**
The Apache ZooKeeper component is a centralized service for provic
synchronization of data.

After about 15 minutes…



□	Name ^	Region	Zone	Total worker nodes
□	✓ cluster-4f49	us-central1	us-central1-a	3



Cluster details

cluster-4f49

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

SSH tunnel

Create an SSH tunnel to connect to a web interface

Component gateway

Provides access to the web interfaces of default and selected optional components on the cluster. Learn more

YARN ResourceManager

HDFS NameNode

MapReduce Job History

YARN Application Timeline

Spark History Server

Tez

Jupyter

Opening the Jupyter Notebook

- Wait for 10 minutes or so
- Click on the cluster
- Navigate to “WEB INTERFACES”
- Select “Jupyter”
- Click on the link and navigate to the home page
- Select the .ipynb file
- Note: source files should be accessed in the below way

  Format : df = spark.read.csv('gs://<bucket-name>/<csv-file>', header=True, inferSchema=True)

  Example : df = spark.read.csv('gs://**newbucket-bdidemo**/cruise_ship_info.csv', header=True, inferSchema=True)

Important to remember

- Dataproc is managed infrastructure for spark clusters
  - Allows user to configure nodes
    - Number & size
    - Security & network options
  - Supports PySpark notebooks