



Fall 2023- CS 3300

## GCP Setup Guide (Borrowed from materials in CSE 6242)

### Getting Started

A video tutorial has been created to help walk through the steps of GCP setup.

You can watch the video for:

- Creating Project
- Creating Bucket
- Creating Cluster

#### NOTE:

1. Ensure you are using personal GMAIL account. Do not use GMAIL account associated with any organization or university.
2. This setup guide is designed to help students quick start GCP Setup and assumes the GMAIL account used is setting up GCP account for the first time. The screens might vary based on your GCP account and this is excepted. Use this guide as a starting point and as a heuristic to identify and setup the GCP Storage buckets and Dataproc Cluster on your own based on the vagaries of your Google Account.

(Students are expected to setup the storage buckets and the clusters on their own. This guide is for reference only, and is not meant to serve as a step-by-step guide due to the variations in individual google accounts)

Instructions to setup GCP Storage buckets and Dataproc Cluster are provided below:

Please make sure you have a Google account and if you don't have one, you would need to create a new account.

1. Login into your account using this [link](#)

2. Access your GCP coupon through Ed discussion and redeem credits.

The screenshot shows the Google Cloud 'GCP credit application' page. At the top, there's a navigation bar with the Google Cloud logo, a 'Select a project' dropdown, and a search bar. Below this, the title 'GCP credit application' is followed by a note: 'Fill in the following information below to apply GCP credits to your account listed below.' The form contains several input fields: 'First name \*' with the value 'Danrong', 'Last name \*' with the value 'Zhang', 'Account email' with a placeholder '@gmail.com', and 'Coupon code \*'. Below the form is a 'Terms and conditions' section with a link to 'Terms of Service' and an 'ACCEPT AND CONTINUE' button. A footnote states '\* Indicates required'.

3. When you enter the educational credits, Google automatically creates a new billing account named “Billing Account for Education” and activate the “free trial option” at the top right corner of the console.

**Note: You might not need to activate if you have previously used GCP with educational credits.**



The screenshot displays the 'Billing Account Overview' page in the Google Cloud console. The left sidebar shows the 'Billing' menu with options like 'Overview', 'Reports', 'Cost table', 'Cost breakdown', 'Budgets & alerts', 'Billing export', 'Cost optimization', 'Committed use discounts...', and 'Release Notes'. The main content area is titled 'BILLING ACCOUNT OVERVIEW' and shows details for the 'Current month' (September 1 – 27, 2022). It includes a 'Month-to-date total cost' of \$0.00 and an 'End-of-month total cost (forecasted)' of 'Not enough historical data to project cost'. A 'View report' link is provided. Below this, the 'Cost trend' section shows the 'Average monthly total cost' as \$0.00 for the period from September 1, 2021, to September 30, 2022. On the right, there's a 'Billing account' section with a 'Manage' link, the account ID '01F4B4-725ACE-9F90ED', and the 'Organization' listed as 'No organization'. At the bottom right, the 'Billing health checks' section shows a status of 0 warnings, 1 info, and 1 success.

4. You will need to enter default billing credit card information if you are using GCP for first time. This is to confirm it is a human and you may need to give your personal credit card details for billing (money would not be deducted from your card). Please make sure to use the billing account named “**Billing Account for Education**” is linked to a new project which will be used to create your storage and clusters (details in step 7). **Don't use your own personal billing account otherwise the free educational credits will not be applied.**


Try Google Cloud Platform for free

## Step 2 of 2

### Customer info

 Account type ⓘ 

Individual ▼

 Name and address ⓘ

Name

Name is required

Address line 1

Address line 1 is required

Address line 2

City

City is required

State ▼ ZIP code ⓘ

State is required ZIP code is required

Phone number (optional)

### Access to all Cloud Platform Products

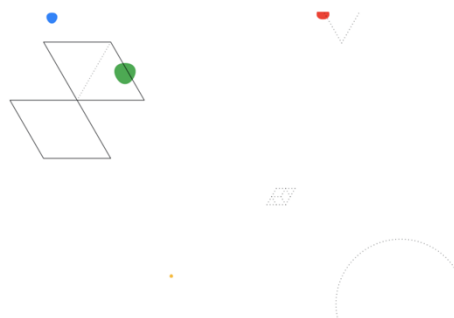
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

### \$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.



State ▼ ZIP code ⓘ

State is required ZIP code is required


Phone number (optional)

### How you pay

 Automatic payments

You pay for this service only after you accrue costs, via an automatic charge when you reach your billing threshold or 30 days after your last automatic payment, whichever comes first.

### Payment method ⓘ

 Add credit or debit card ▼

#

Card number MM YY CVC

Card number is required Month is required Year is required CVC is required

Cardholder name

☒ Credit or debit card address is same as above

### Access to all Cloud Platform Products

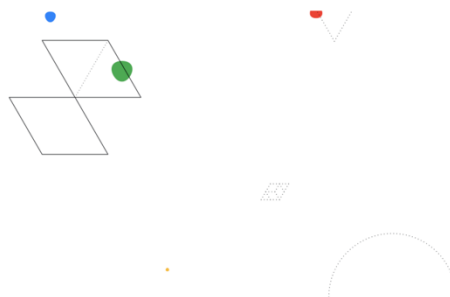
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

### \$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

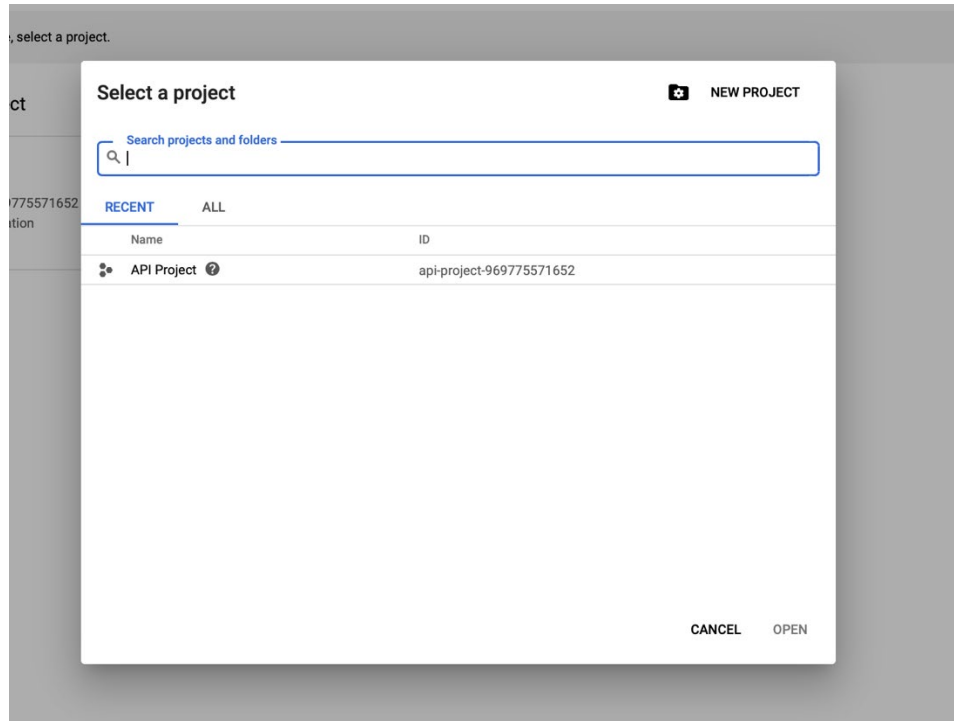
We ask you for your credit card to make sure you are not a robot. You won't be charged unless you manually upgrade to a paid account.



START MY FREE TRIAL

5. Go to [Google Console Home](#)

- On the top left menu, click on dropdown to “select a project” (or if you already have existing projects, it will list the latest project. Click on the drop down in this case as well)
- Click on New Project on the top right corner



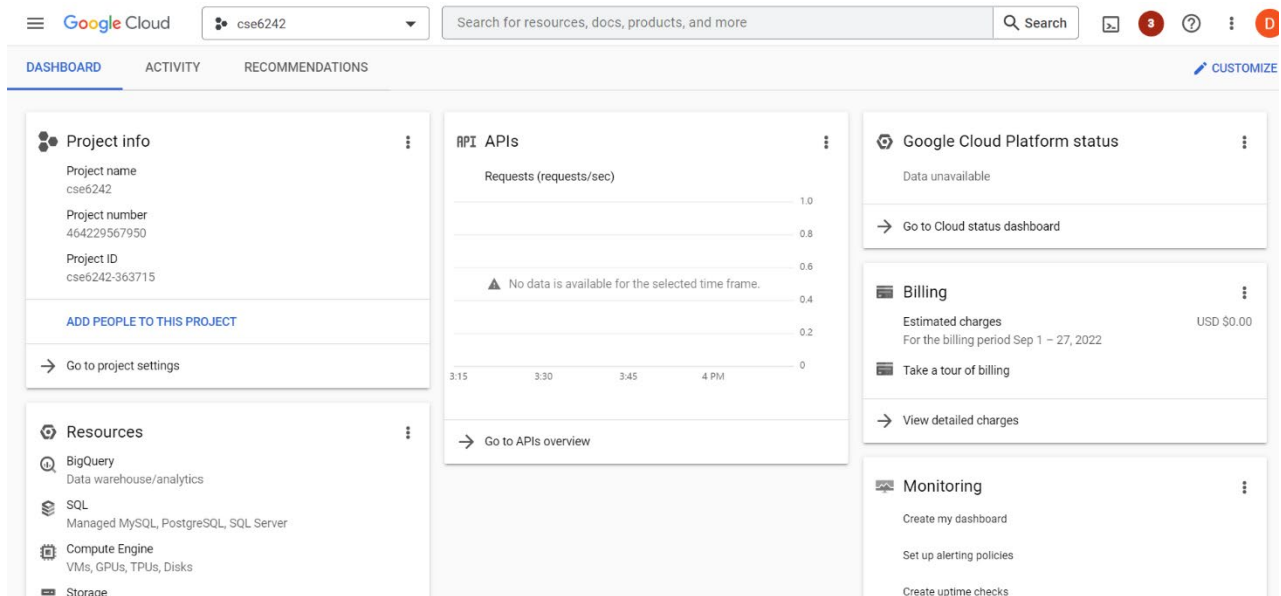
- Enter CS3300 (or any other default name) and create a new project.

The screenshot shows the 'New Project' page. At the top is the Google Cloud header with a search bar. Below it is the 'New Project' title. A warning box states: 'You have 21 projects remaining in your quota. Request an increase or delete projects. [Learn more](#)'. Below this is a 'MANAGE QUOTAS' link. The form has four main sections: 'Project name \*' with the value 'CSE6242' and an 'EDIT' link; 'Project ID: cse6242-363820. It cannot be changed later.'; 'Billing account \*' with a dropdown menu showing 'Billing Account for Education'; and 'Location \*' with a dropdown menu showing 'No organization' and a 'BROWSE' link. At the bottom are 'CREATE' and 'CANCEL' buttons.

**Note: The new Project CS3300 should be linked to the Billing Account “Billing Account for Education”. This step can be achieved while creating a new project in step 7 as listed above. If you have only one billing account (namely Billing Account “Billing Account for**

Education”, it will automatically link your project to Billing Account “Billing Account for Education”.

9. Wait till the new project is created.

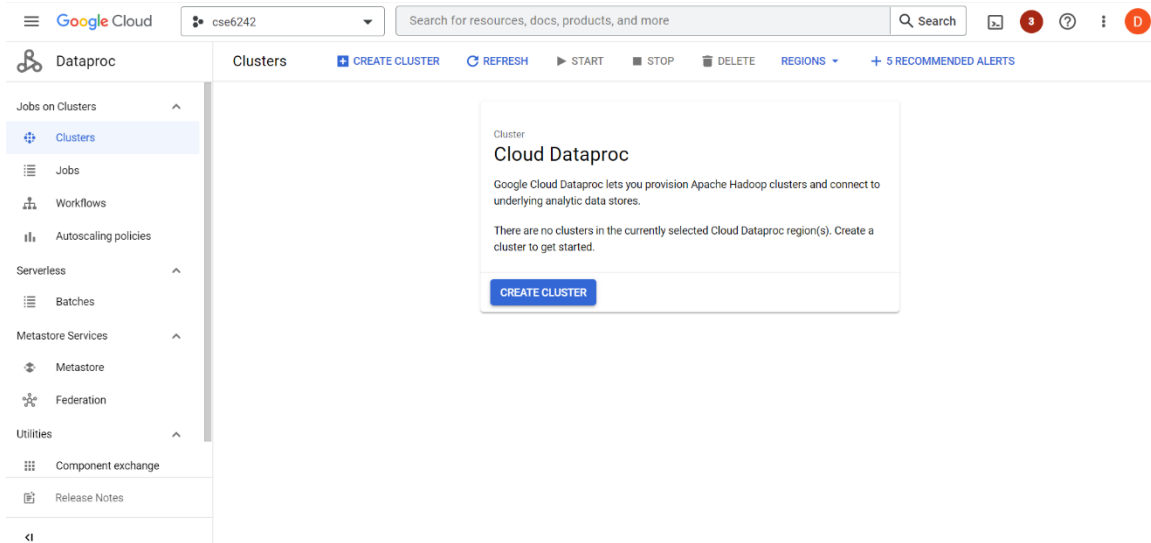


## Creating Clusters in Google DataProc (With Spark and Jupyter Notebook Components)

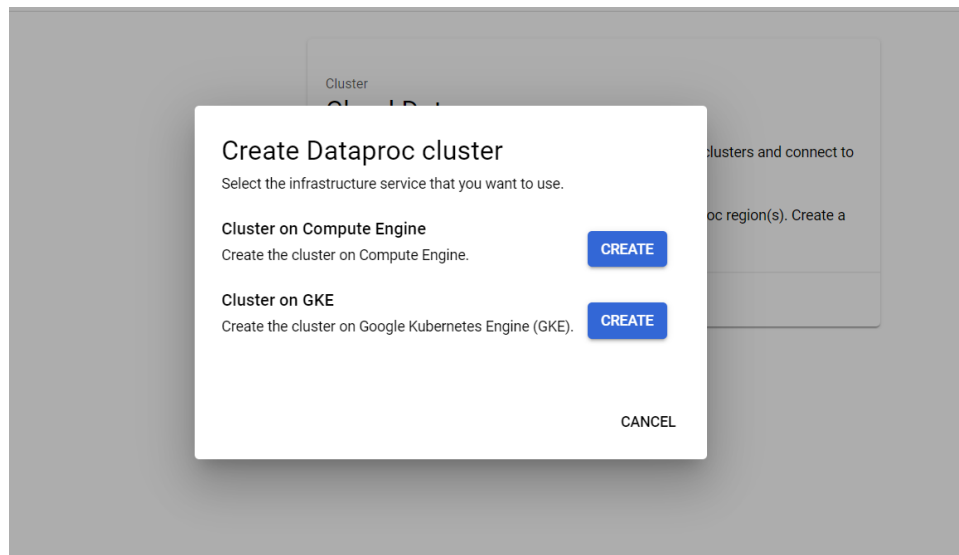
Google DataProc can be used to create an Apache Spark distribution cluster. This means that it handles large amounts of data on demand. The next step is to use Google’s DataProc web-based management tool to create a Linux cluster. Follow the recommended steps shown to create a new cluster (or see full documentation here on [GCP site](#)).

Follow the steps in the below link to create the cluster:

1. Go to [Google Data Proc home page](#)
2. On the top left corner, make sure the project is CSE6242.



3. Enable API and Click CREATE for 'Cluster on Compute Engine'. (Note that you might not need to enable API if you previously used Cloud Dataproc before)



4. Choose your GT Username as the cluster name (or you can use the default name)
5. Choose the closest region and location. For example, us-east4-a or us-east4-b or us-east4-c for students in east coast
6. Choose Cluster type as 'Standard'(1 master, N workers)
7. Leave the autoscaling policy as None.

The screenshot shows the Google Cloud console interface for creating a Dataproc cluster. The left sidebar contains navigation links for Jobs on Clusters, Clusters, Jobs, Workflows, Autoscaling policies, Serverless, Batches, Metastore Services, Metastore, Federation, Utilities, Component exchange, Workbench, and Release Notes. The main content area is titled 'Create a Dataproc cluster on Compute Engine' and shows the 'Set up cluster' step. The 'Name' field is 'dzhang373'. The 'Location' is set to 'us-east4' region and 'us-east4-c' zone. The 'Cluster type' is 'Standard (1 master, N workers)'. The 'Autoscaling' policy is 'None'. The 'Enhanced Flexibility Mode' is checked. There are 'CREATE' and 'CANCEL' buttons, and a link to 'EQUIVALENT COMMAND LINE'.

8. Check the default Image version under Image type and version, if your default version is 2.0-debian 10, click change and **2.11**.
9. Check the Component gateway checkbox (Enable access to the web interfaces of default and selected optional components on the cluster.)
10. Under “Optional components”, click on “Select Component” button and select the "Jupyter Notebook" component. You can ignore the warning that will pop up on the bottom about installing anaconda additionally.

## 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 ?
- ☐ Trino ?
- ☐ ZooKeeper ?
- ☐ Ranger ?
- ☐ Flink ?
- ☐ Docker ?
- ☐ Solr ?
- ☐ Hudi ?

**i** Component Jupyter requires that component Anaconda is also selected if image version is not 2.0.

11. In 'Configure nodes', for both manager and worker nodes set the machine type as the base configuration (n2-standard-2 (2 vCPU, 8 GB memory))
12. Under 'Customize cluster', use the "Cloud Storage staging bucket" field to select your bucket- Browse the name of the bucket you created in prior steps (only specify the name of the bucket which would be your GT Username). Your notebooks will be stored in Cloud Storage under gs://bucket-name/notebooks/jupyter

### Cloud Storage staging bucket

Storage staging bucket  **BROWSE**

Cloud Storage staging bucket to be used for storing cluster job dependencies, job driver output, and cluster config files.

13. Click Create

It will take a few minutes for your Cluster to be up and running. Once the cluster is up and running:

1. Click on your **Cluster Name** to navigate into the cluster details screen



2. Click the **Web Interfaces** tab to display a list of Component Gateway links to the web interfaces of default and optional components installed on the cluster
3. Click the **Jupyter** link. The Jupyter notebook web UI opens in your local browser.
4. **Upload HW3 Q4 skeleton notebook** into your Jupyter notebook web UI (under GCS folder and not under local folder)
5. Open the notebook, set the kernel to PySpark and you will be ready to start working on the Q4 section.

## Delete Storage and Clusters

After you complete the Q4 section of this homework, please remember to delete both storage and clusters. The deletion process is easy.

**NOTE: Ensure to download all the notebooks from GCP cluster before deleting the GCP cluster (otherwise the work will be lost).**

1. On the Storage home page, check the bucket you want to delete and then click “DELETE.”
2. Wait till the bucket is deleted without any errors.
3. On the Cluster page, check the cluster you want to delete and then click “DELETE.”
4. Wait till the cluster is deleted without any errors.