Dataproc: Qwik Start - Command Line

GSP104

Overview

Cloud Dataproc is a fast, easy-to-use, fully-managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, more cost-efficient way. Operations that used to take hours or days take seconds or minutes instead. Create Cloud Dataproc clusters quickly and resize them at any time, so you don't have to worry about your data pipelines outgrowing your clusters.
This lab shows you how to use gcloud on the Google Cloud to create a Google Cloud Dataproc cluster, run a simple Apache Spark job in the cluster, then modify the number of workers in the cluster.

Setup and Requirements

Before you click the Start Lab button

Read these instructions. Labs are timed and you cannot pause them. The timer, which starts when you click **Start Lab**, shows how long Google Cloud resources will be made available to you.

This Qwiklabs hands-on lab lets you do the lab activities yourself in a real cloud environment, not in a simulation or demo environment. It does so by giving you new, temporary credentials that you use to sign in and access Google Cloud for the duration of the lab.

What you need

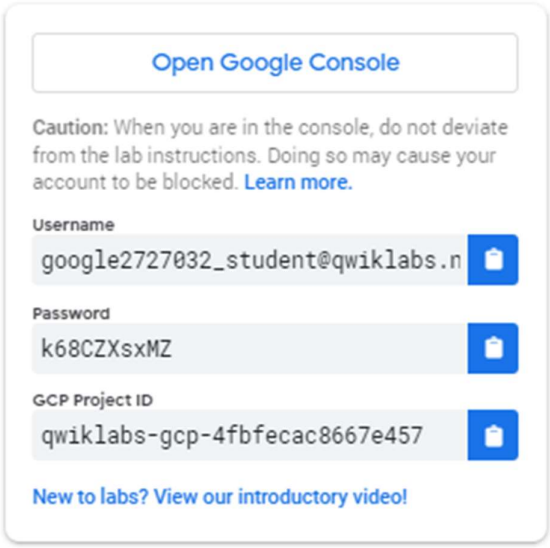
To complete this lab, you need:

- Access to a standard internet browser (Chrome browser recommended).
- Time to complete the lab.
  Note: If you already have your own personal Google Cloud account or project, do not use it for this lab.

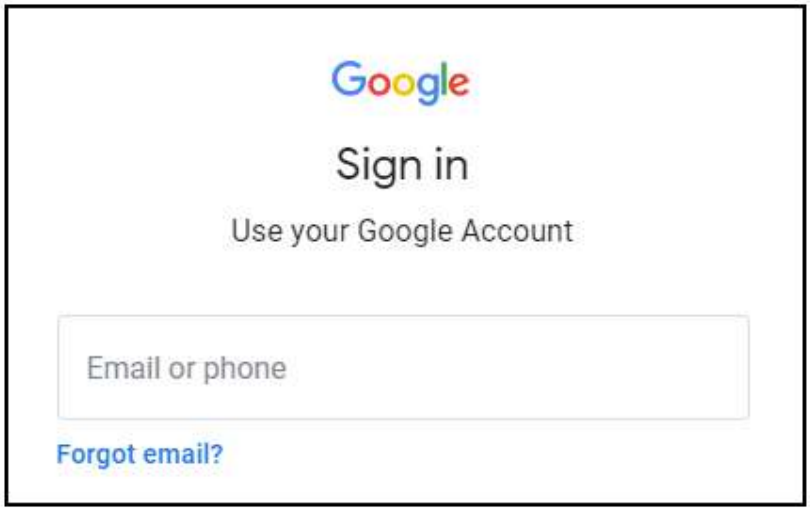
Note: If you are using a Pixelbook, open an Incognito window to run this lab.

How to start your lab and sign in to the Google Cloud Console

1. Click the **Start Lab** button. If you need to pay for the lab, a pop-up opens for you to select your payment method. On the left is a panel populated with the temporary credentials that you must use for this lab.

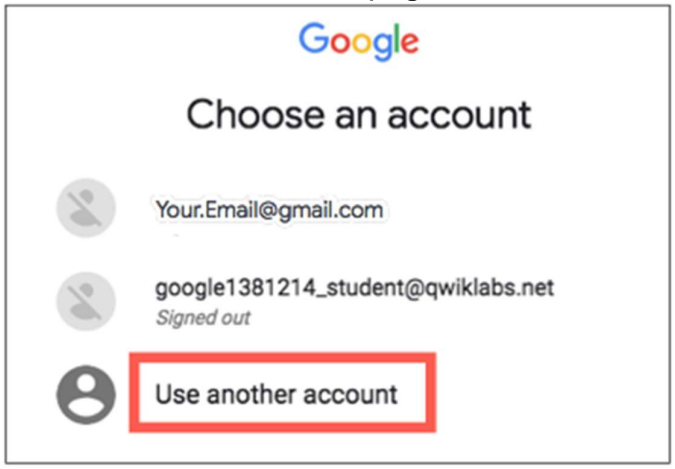


2. Copy the username, and then click **Open Google Console**. The lab spins up resources, and then opens another tab that shows the **Sign in** page.



Tip: Open the tabs in separate windows, side-by-side.

If you see the **Choose an account** page, click **Use Another Account**.



3. In the **Sign in** page, paste the username that you copied from the Connection Details panel. Then copy and paste the password.

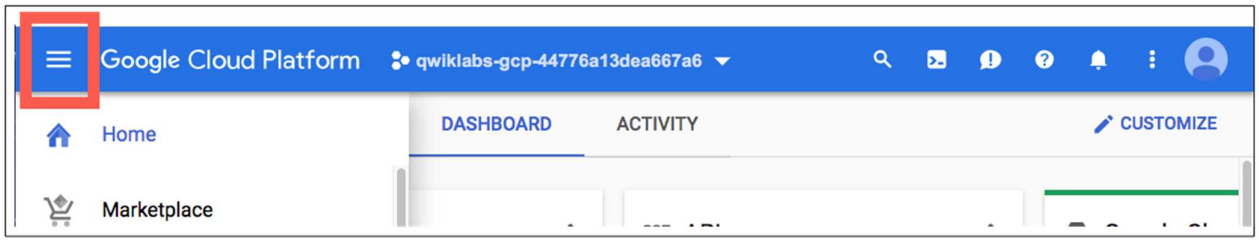
   Important: You must use the credentials from the Connection Details panel. Do not use your Qwiklabs credentials. If you have your own Google Cloud account, do not use it for this lab (avoids incurring charges).

4. Click through the subsequent pages:

   - Accept the terms and conditions.
   - Do not add recovery options or two-factor authentication (because this is a temporary account).
   - Do not sign up for free trials.

After a few moments, the Cloud Console opens in this tab.

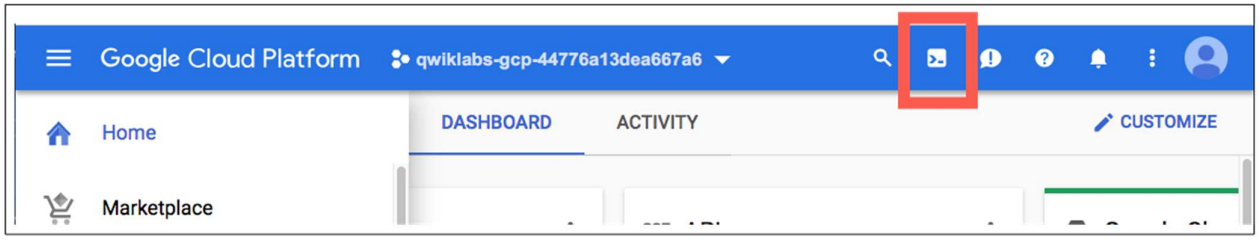
Note: You can view the menu with a list of Google Cloud Products and Services by clicking the **Navigation menu** at the top-
left.



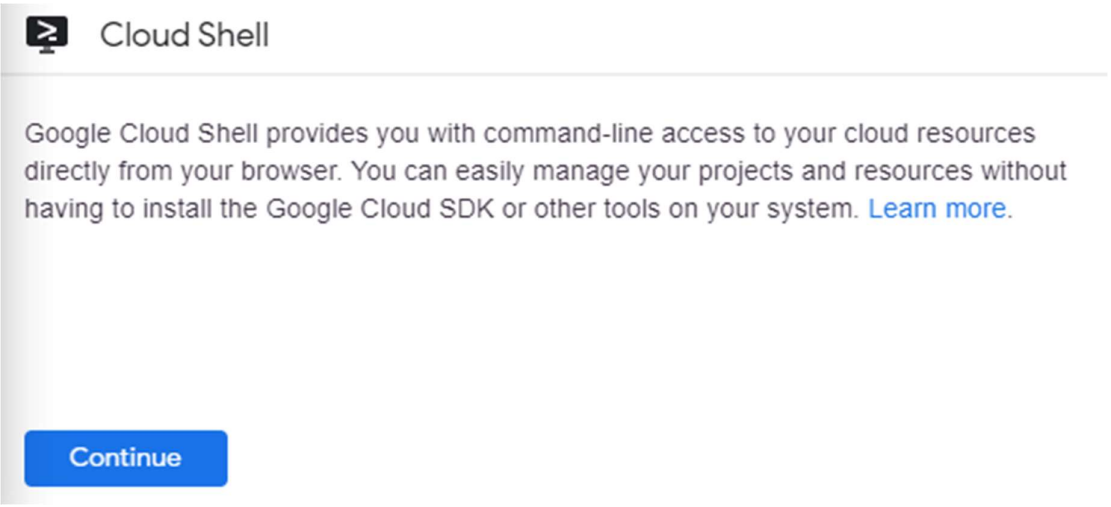
Activate Cloud Shell

Cloud Shell is a virtual machine that is loaded with development tools. It offers a persistent 5GB home directory and runs on the Google Cloud. Cloud Shell provides command-line access to your Google Cloud resources.

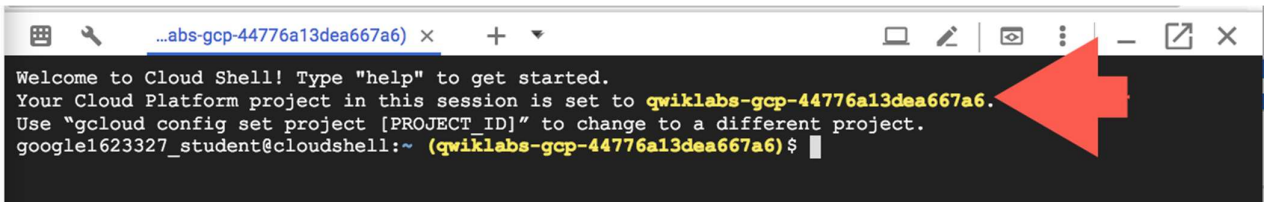
In the Cloud Console, in the top right toolbar, click the **Activate Cloud Shell** button.



Click **Continue**.



It takes a few moments to provision and connect to the environment. When you are connected, you are already authenticated, and the project is set to your *PROJECT_ID*. For example:



`gcloud` is the command-line tool for Google Cloud. It comes pre-installed on Cloud Shell and supports tab-completion.

You can list the active account name with this command:

```
gcloud auth list
```

(Output)

```
Credentialed accounts:
 - <myaccount>@<mydomain>.com (active)
```

(Example output)

```
Credentialed accounts:
 - google1623327_student@qwiklabs.net
```

You can list the project ID with this command:

```
gcloud config list project
```

(Output)

```
[core]
project = <project_ID>
```

(Example output)

```
[core]
project = qwiklabs-gcp-44776a13dea667a6
```

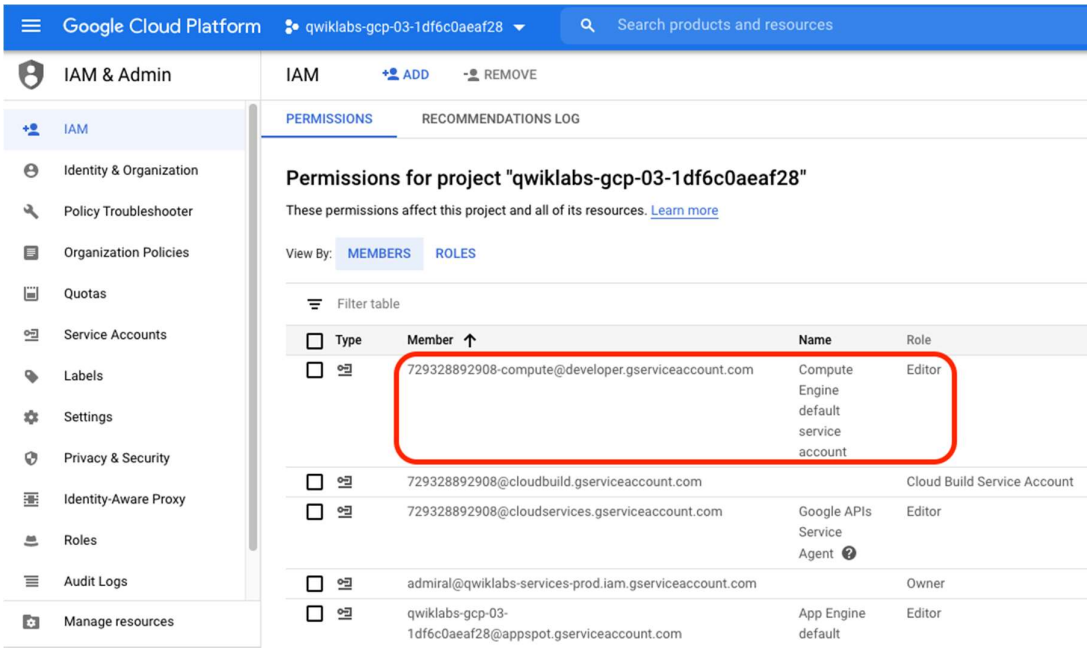
For full documentation of `gcloud` see the gcloud command-line tool overview.

Check project permissions

Before you begin your work on Google Cloud, you need to ensure that your project has the correct permissions within Identity and Access Management (IAM).

1. In the Google Cloud console, on the **Navigation menu** (), click **IAM & Admin** > **IAM**.

2. Confirm that the default compute Service Account `{project-number}-compute@developer.gserviceaccount.com` is present and has the `editor` role assigned. The account prefix is the project number, which you can find on **Navigation menu** > **Home**.



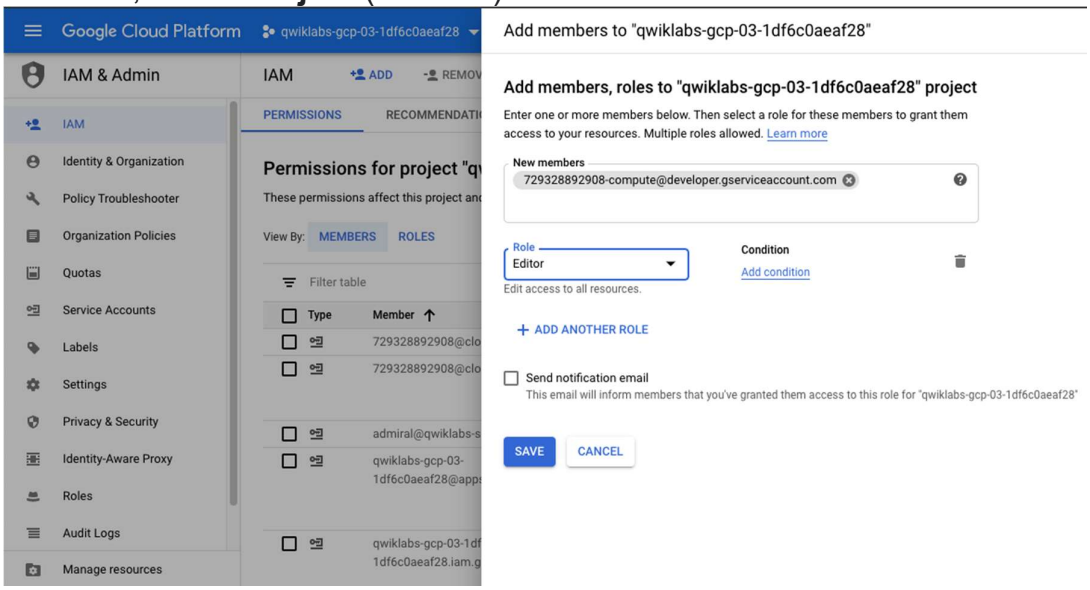
If the account is not present in IAM or does not have the `editor` role, follow the steps below to assign the required role.

- In the Google Cloud console, on the **Navigation menu**, click **Home**.
- Copy the project number (e.g. `729328892908`).
- On the **Navigation menu**, click **IAM & Admin** > **IAM**.
- At the top of the **IAM** page, click **Add**.
- For **New members**, type:

```
{project-number}-compute@developer.gserviceaccount.com
```

Replace `{project-number}` with your project number.

- For **Role**, select **Project** (or Basic) > **Editor**. Click **Save**.



Create a cluster

In Cloud Shell, run the following command to set the Region:

```
gcloud config set dataproc/region us-central1
```

Run the following command to create a cluster called `example-cluster` with default Cloud Dataproc settings:

```
gcloud dataproc clusters create example-cluster --worker-boot-disk-size 500
```

If asked to confirm a zone for you cluster. Enter **Y**.

Your cluster will build for a couple of minutes.

```
Waiting for cluster creation operation...done.
Created [... example-cluster]
```

When you see a "Created" message, you're ready to move on.

Test Completed Task

Click **Check my progress** to verify your performed task. If you have successfully created a Dataproc cluster, you will see an assessment score.

Create a Dataproc cluster

Check my progress

Submit a job

Run this command to submit a sample Spark job that calculates a rough value for pi:

```
gcloud dataproc jobs submit spark --cluster example-cluster \
  --class org.apache.spark.examples.SparkPi \
  --jars file:///usr/lib/spark/examples/jars/spark-examples.jar -- 1000
```

The command specifies:

- That you want to run a spark job on the `example-cluster` cluster
- The `class` containing the main method for the job's pi-calculating application
- The location of the jar file containing your job's code
- The parameters you want to pass to the job—in this case, the number of tasks, which is `1000`

Parameters passed to the job must follow a double dash (--). See the gcloud documentation for more information.

The job's running and final output is displayed in the terminal window:

```
Waiting for job output...
...
Pi is roughly 3.14118528
...
state: FINISHED
```

Test Completed Task

Click **Check my progress** to verify your performed task. If you have successfully submitted a job, you will see an assessment score.

Submit a job

Check my progress

Update a cluster

To change the number of workers in the cluster to four, run the following command:

```
gcloud dataproc clusters update example-cluster --num-workers 4
```

Your cluster's updated details are displayed in the command's output:

```
Waiting on operation [projects/qwiklabs-gcp-
7f7aa0829e65200f/regions/global/operations/b86892cc-e71d-4e7b-aa5e-6030c945ea67].
Waiting for cluster update operation...done.
```

You can use the same command to decrease the number of worker nodes:

```
gcloud dataproc clusters update example-cluster --num-workers 2
```

Now you can create a Dataproc cluster and adjust the number of workers from the `gcloud` command line on the Google Cloud.

Test your Understanding

Below are multiple-choice questions to reinforce your understanding of this lab's concepts. Answer them to the best of your abilities.

Clusters can be created and scaled quickly with a variety of virtual machine types, disk sizes, and number of nodes.

True

False

Congratulations!

Finish Your Quest

Continue your Quest with Baseline: Data, ML, AI. A Quest is a series of related labs that form a learning path. Completing this Quest earns you the badge above, to recognize your achievement. You can make your badge (or badges) public and link to them in your online resume or social media account. Enroll in this Quest and get immediate completion credit if you've taken this lab. See other available Qwiklabs Quests.

Next Steps / Learn More

This lab is also part of a series of labs called Qwik Starts. These labs are designed to give you a little taste of the many features available with Google Cloud. Search for "Qwik Starts" in the lab catalog to find the next lab you'd like to take!

Google Cloud Training & Certification

...helps you make the most of Google Cloud technologies. Our classes include technical skills and best practices to help you get up to speed quickly and continue your learning journey. We offer fundamental to advanced level training, with on-demand, live, and virtual options to suit your busy schedule. Certifications help you validate and prove your skill and expertise in Google Cloud technologies.
Manual Last Updated March 01, 2021
Lab Last Tested March 01, 2021