www.cloudskillsboost.google /course_sessions/3062553/labs/343602

PDE Prep: Cloud Dataproc Cluster Operations and Maintenance | Google Cloud Skills Boost

Qwiklabs ⁝ 9-11 minutes

Overview

This is a Challenge lab where you must complete a series of tasks within a limited time period. Instead of following step-by-step instructions, you are presented with general objectives. An automated scoring system (shown on this page) will provide feedback on whether you have completed each task correctly.

To score 100% the challenge, you must complete all tasks within the time period!

This lab does not teach GCP concepts. Instead, it is a test of your Data Engineering skills. This lab is only recommended for students who have Cloud Dataproc skills.

Objectives

- Create a cluster with access to a Cloud Storage staging bucket.
- Run PySpark jobs with input arguments.
- Upgrade the master node configuration on an existing cluster.
- Resolve a cluster capacity performance issue.
- Upgrade the number of worker nodes on an existing cluster.

Setup and requirements

For each lab, you get a new Google Cloud project and set of resources for a fixed time at no cost.

1. Sign in to Qwiklabs using an **incognito window**.
2. Note the lab's access time (for example, `1:15:00`), and make sure you can finish within that time. There is no pause feature. You can restart if needed, but you have to start at the beginning.
3. When ready, click **Start lab**.
4. Note your lab credentials (**Username** and **Password**). You will use them to sign in to the Google Cloud Console.
5. Click **Open Google Console**.

6. Click **Use another account** and copy/paste credentials for **this** lab into the prompts.
   If you use other credentials, you'll receive errors or **incur charges**.

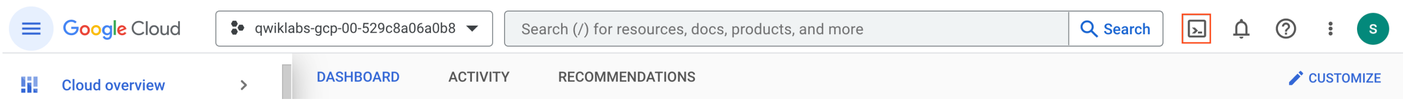
7. Accept the terms and skip the recovery resource page.

Activate Google Cloud Shell

Google 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.

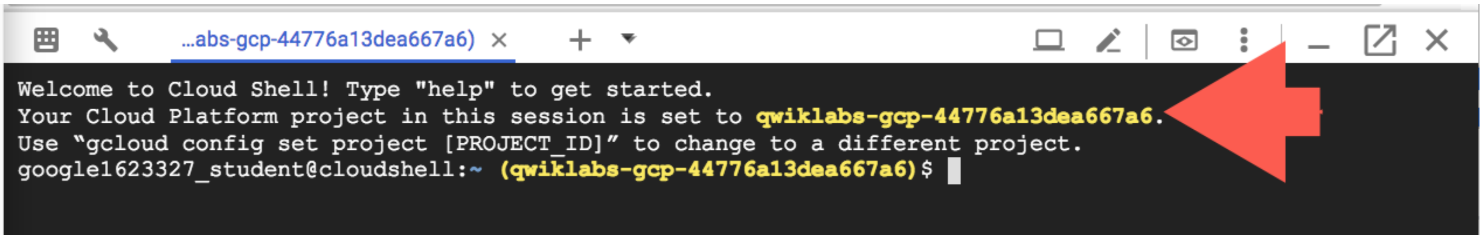
Google Cloud Shell provides command-line access to your Google Cloud resources.

1. In Cloud console, on the top right toolbar, click the Open Cloud Shell button.


Google Cloud
qwiklabs-gcp-00-529c8a06a0b8
Search (/) for resources, docs, products, and more
Search
S
Cloud overview
DASHBOARD
ACTIVITY
RECOMMENDATIONS
CUSTOMIZE

2. 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:



```
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to qwiklabs-gcp-44776a13dea667a6.
Use "gcloud config set project [PROJECT_ID]" to change to a different project.
google1623327_student@cloudshell:~ (qwiklabs-gcp-44776a13dea667a6)$
```

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: - @.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 =

Example output:

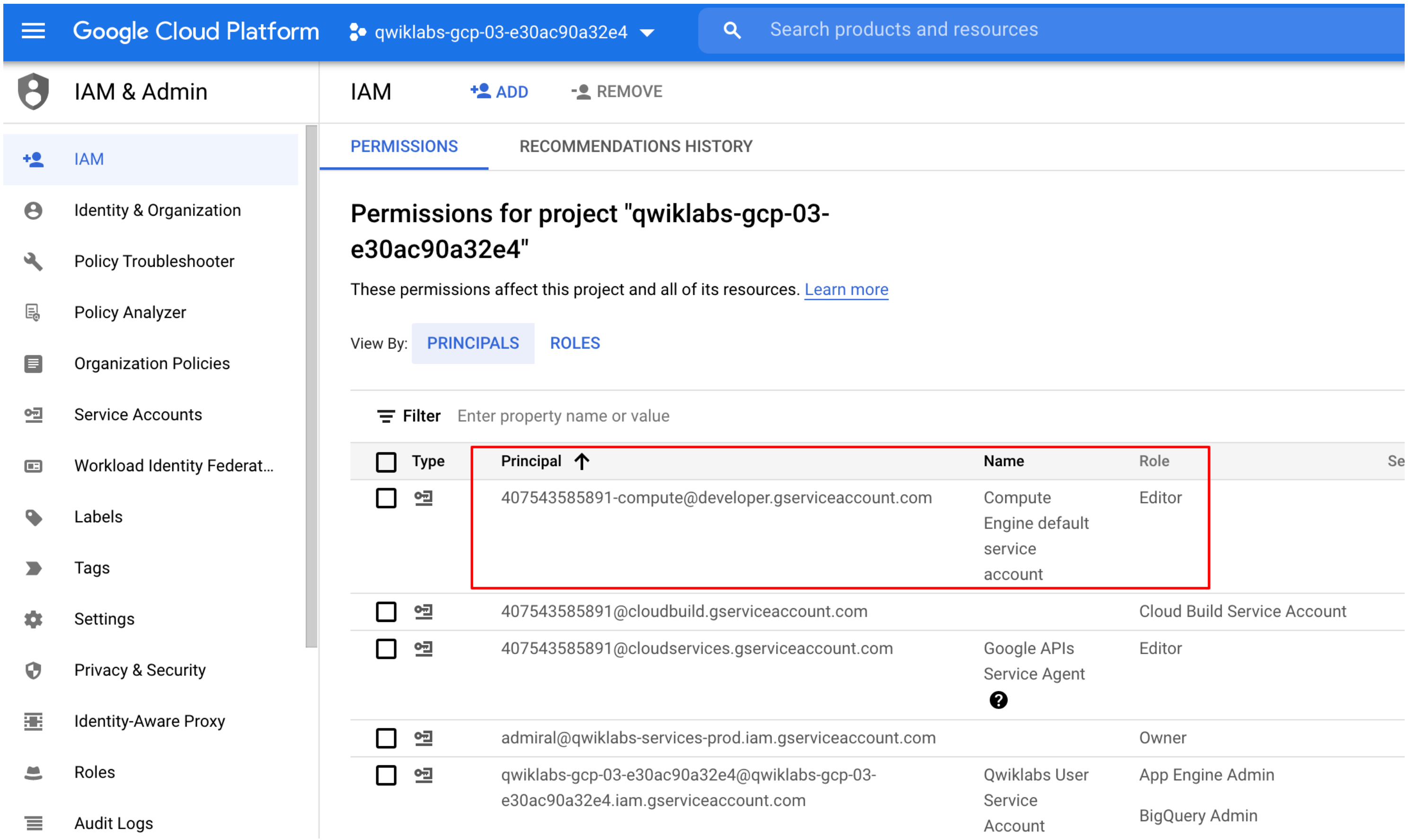
[core] project = qwiklabs-gcp-44776a13dea667a6 **Note:** Full documentation of **gcloud** is available in the gcloud CLI overview guide .

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** (☰), select **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**.



Permissions for project "qwiklabs-gcp-03-e30ac90a32e4"

These permissions affect this project and all of its resources. Learn more

□	Type	Principal ↑	Name	Role	Se
□		407543585891-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor	
□		407543585891@cloudbuild.gserviceaccount.com		Cloud Build Service Account	
□		407543585891@cloudservices.gserviceaccount.com	Google APIs Service Agent ❓	Editor	
□		admiral@qwiklabs-services-prod.iam.gserviceaccount.com		Owner	
□		qwiklabs-gcp-03-e30ac90a32e4@qwiklabs-gcp-03-e30ac90a32e4.iam.gserviceaccount.com	Qwiklabs User Service Account	App Engine Admin BigQuery Admin	

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

1. In the Google Cloud console, on the **Navigation menu**, click **Home**.

2. Copy the project number (e.g. `729328892908`).

3. On the **Navigation menu**, select **IAM & Admin** > **IAM**.

4. At the top of the **IAM** page, click **Add**.

5. For **New principals**, type:

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

6. Replace `{project-number}` with your project number.
7. For **Role**, select **Project** (or Basic) > **Editor**.
8. Click **Save**.

Challenge scenario

For this scenario, you work as a Data Engineer for MJTelco. (Congratulations on the new job!)

MJTelco's Data Scientist team plans to port an existing predictive machine learning application to a Cloud Dataproc cluster. The application is written in Python and runs on Spark. The application takes a long time to run, even with sample data.

So the Data Scientists have established a benchmark by running a program in their data center. They have attempted to run the benchmark on a Cloud Dataproc cluster, but it is taking longer than they would like. Your job is to run the benchmark program on Cloud Dataproc and make adjustments to the cluster configuration to meet their requirements.

Note: The benchmark program is a PySpark application. It calculates the value of PI. The input value determines how many iterations are used in the calculation.

Requirements

If the benchmark program is given an input value of **220**, and the job completes in under **75** seconds, the requirements will be met.

At this time, when the benchmark program is submitted with the input value of **20**, the job completes in under **75** seconds. When it is submitted with the required input value of **220**, the job takes about **2** minutes to run, which does not meet the requirement.

Objective 1

Your first job is to duplicate the cluster that the Data Scientists are using and then run the benchmark job.

Task 1. Stage the benchmark PySpark application

1. Create a Cloud Storage bucket for use by your Cloud Dataproc cluster.
2. Give the bucket the same name as your **project**.
3. Copy the benchmark Python Spark application to the bucket in your project.

The benchmark application has been shared with you from a Cloud Storage bucket: `gs://cloud-training/preppde/benchmark.py`.

Click *Check my progress* to verify the objective. Create a Cloud Storage bucket

Task 2. Create a Cloud Dataproc Cluster that matches the Data Analyst's configuration

The Data Scientists are using a minimal Cloud Dataproc cluster consisting of one master node and two worker nodes. All the instances are of type **n1-standard-2**.

1. Create a Cloud Dataproc cluster named `mjtelco` using version **2.0 (Debian 10, Hadoop 3.2, Spark 3.1)** with a master node of **n1-standard-2** and two worker nodes of **n1-standard-2** in **us-east1** region and **us-east1-b** zone.
2. Use the default settings on everything else. Remember to set advanced options to give the cluster access to your Cloud Storage staging bucket.

Click *Check my progress* to verify the objective. Create a Dataproc cluster

Task 3. Demonstrate the successful benchmark job without the required input value

1. Submit the python job to the cluster, and give the job the name `mjtelco-test-1`.
2. Give the job the input argument of **20**.
3. For **Max restarts per hour**, enter **1**.

The job should take under 75 seconds to run and should succeed.

Click *Check my progress* to verify the objective. Demonstrate the successful benchmark job

Task 4. Demonstrate the slower benchmark job with the required input value

1. Submit the python job to the cluster, and give the job the name `mjtelco-test-2`.
2. Give the job the input argument of **220**.
3. For **Max restarts per hour**, enter `1`.

The job should take between 1 minute 45 seconds and 3 minutes to run and should succeed.

Click *Check my progress* to verify the objective. Demonstrate the slower benchmark job

Objective 2

Your second job is to improve the performance of the cluster and to reduce the time it takes to run the benchmark job.

Task 5. Upgrade the master node

- Upgrade the master node to a 4-CPU instance, **n1-standard-4**.

Task 6. Demonstrate that the benchmark job completes in less time

1. After the upgraded master node is running, submit the python job again to the cluster.
2. Give the job the name `mjtelco-test-3`.
3. Give the job the input argument of **220**.
4. For **Max restarts per hour**, enter `1`.

The job should take about 2 minutes to run and should succeed.

Click *Check my progress* to verify the objective. Demonstrate the faster benchmark job

Task 7. Grow the cluster

You are getting closer but the job still does not complete in under the required time (under 75 seconds) when given the input value of **220**.

- Upgrade the cluster by adding three more **n1-standard-2** worker nodes for a total of five workers.

Task 8. Submit the job and verify improved performance

1. After the additional nodes are running, submit the job again.
2. Submit the python job to the cluster, and give the job the name `mjtelco-test-4`.
3. Give the job the input argument of **220**.
4. For **Max restarts per hour**, enter `1`.

The benchmark job should now complete within the required time (under 75 seconds).

Click *Check my progress* to verify the objective. Grow the cluster and submit the job

End your lab

When you have completed your lab, click **End Lab**. Google Cloud Skills Boost removes the resources you’ve used and cleans the account for you.

You will be given an opportunity to rate the lab experience. Select the applicable number of stars, type a comment, and then click **Submit**.

The number of stars indicates the following:

- 1 star = Very dissatisfied
- 2 stars = Dissatisfied
- 3 stars = Neutral
- 4 stars = Satisfied
- 5 stars = Very satisfied

You can close the dialog box if you don't want to provide feedback.

For feedback, suggestions, or corrections, please use the **Support** tab.