Serverless Data Analysis with Dataflow: MapReduce in Dataflow (Python) | Qwiklabs

Tuesday, November 17, 2020 3:41 PM

Clipped from: https://googlecourses.qwiklabs.com/course_sessions/71782/labs/11648

Overview

In this lab, you will identify Map and Reduce operations, execute the pipeline, use command line parameters.

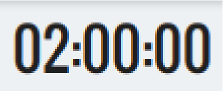
Objective

- Identify Map and Reduce operations
- Execute the pipeline
- Use command line parameters

Setup

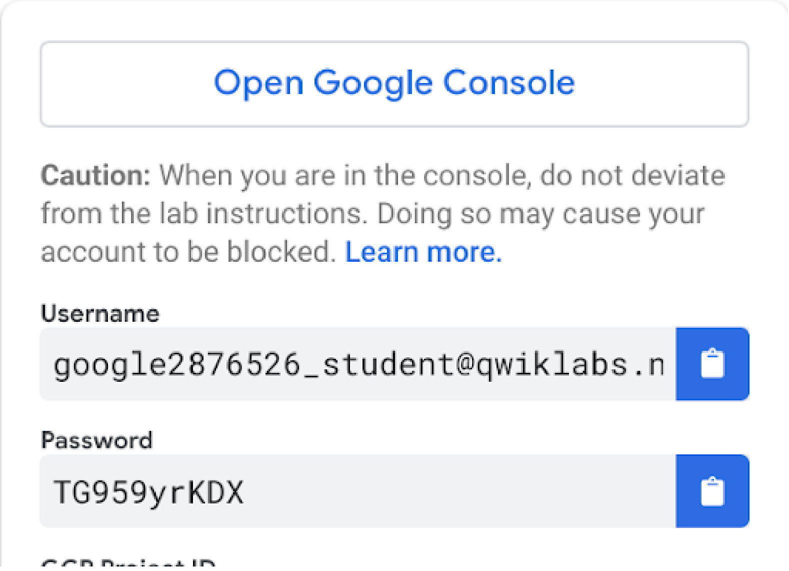
For each lab, you get a new GCP project and set of resources for a fixed time at no cost.

1. Make sure you signed into Qwiklabs using an **incognito window**.
2. Note the lab's access time (for example,

   
   02:00:00

   and make sure you can finish in that time block.
3. When ready, click

   .
4. Note your lab credentials. You will use them to sign in to Cloud Platform Console.



Caution: When you are in the console, do not deviate from the lab instructions. Doing so may cause your account to be blocked. Learn more.

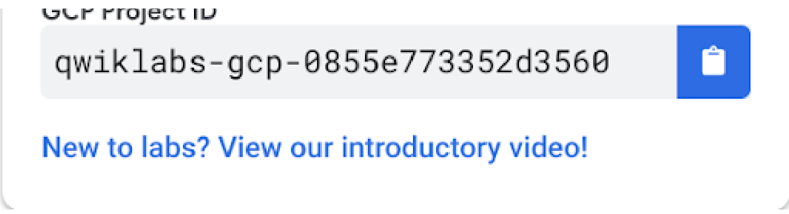
Username

google2876526_student@qwiklabs.n

Password

TG959yrKDX

GCP Project ID



GCP Project ID

qwiklabs-gcp-0855e773352d3560

New to labs? View our introductory video!

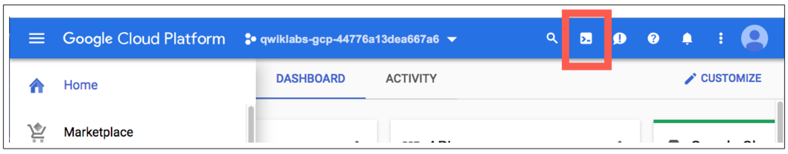
5. Click **Open Google Console**.
6. Click **Use another account** and copy/paste credentials for **this** lab into the prompts.

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

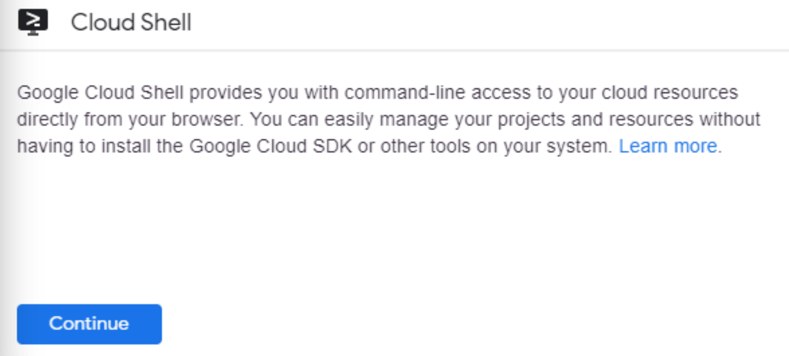
Activate Google Cloud Shell

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 GCP resources.

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



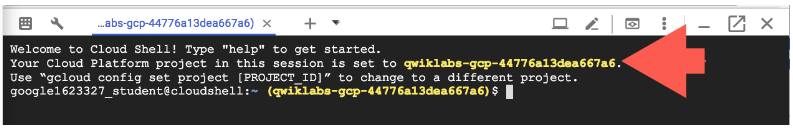
2. Click **Continue**.



Cloud Shell

Google Cloud Shell provides you with command-line access to your cloud resources directly from your browser. You can easily manage your projects and resources without having to install the Google Cloud SDK or other tools on your system. Learn more.

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 Platform. 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:

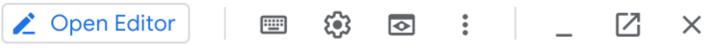
```
[core]
project = qwiklabs-gcp-44776a13dea667a6
```

Full documentation of **gcloud** is available on [Google Cloud gcloud Overview]().

Launch Google Cloud Shell Code Editor

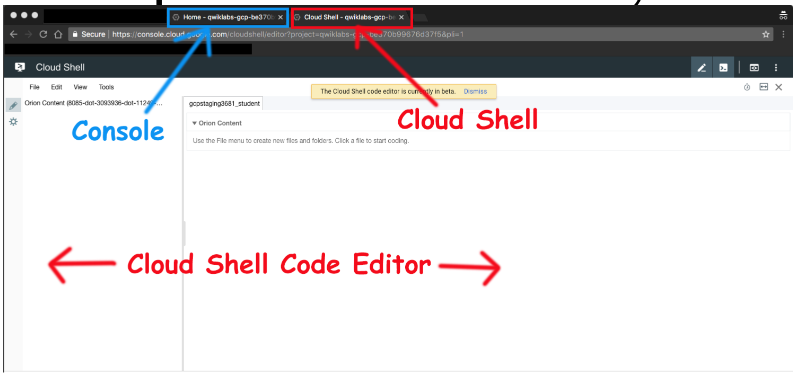
Use the Google Cloud Shell Code Editor to easily create and edit directories and files in the Cloud Shell instance.

Once you activate the Google Cloud Shell, click the **Open editor** button to open the Cloud Shell Code Editor.



You now have three interfaces available:

- The Cloud Shell Code Editor
- Console (By clicking on the tab). You can switch back and forth between the Console and Cloud Shell by clicking on the tab.
- The Cloud Shell Command Line (By clicking on **Open Terminal** in the Console)


[redacted]

[redacted]

The Cloud Shell code editor is currently in beta.

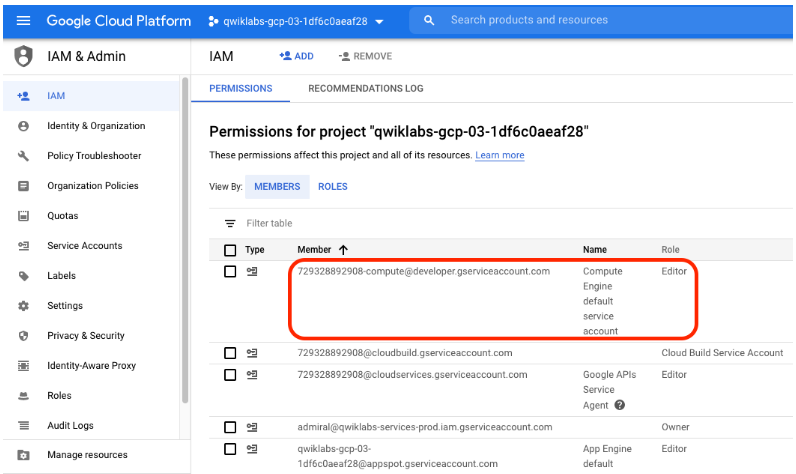
Use the File menu to create new files and folders. Click a file to start coding.

Home - qwiklabs-gcp-be370b
Cloud Shell - qwiklabs-gcp-be
Secure | https://console.cloud.google.com/cloudshell/editor?project=qwiklabs-gcp-be370b99676d37f5&pli=1
Cloud Shell
File
Edit
View
Tools
Dismiss
Orion Content (8085-dot-3093936-dot-1124...
gcpstaging3681_student
Orion Content
Console
Cloud Shell
Cloud Shell Code Editor

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**.



Permissions for project "qwiklabs-gcp-03-1df6c0aeaf28"

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

View By: MEMBERS ROLES

Filter table

□ Type	Member ↑	Name	Role
□	729328892908-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor
□	729328892908@cloudbuild.gserviceaccount.com		Cloud Build Service Account
□	729328892908@cloudservices.gserviceaccount.com	Google APIs Service Agent	Editor
□	admiral@qwiklabs-services-prod.iam.gserviceaccount.com		Owner
□	qwiklabs-gcp-03-1df6c0aeaf28@appspot.gserviceaccount.com	App Engine default	Editor

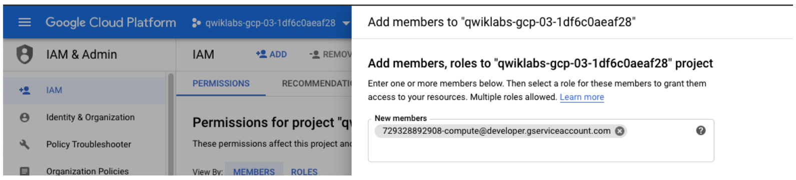
If the account is not present in IAM or does not have the `editor` role, follow the steps below to assign the required role.

- 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** > **Editor**. Click **Save**.

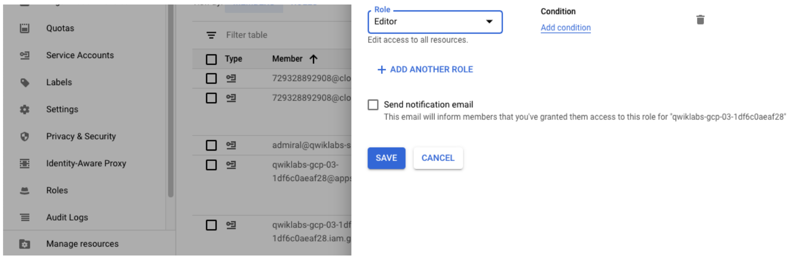


Add members to "qwiklabs-gcp-03-1df6c0aeaf28"

Add members, roles to "qwiklabs-gcp-03-1df6c0aeaf28" project

Enter one or more members below. Then select a role for these members to grant them access to your resources. Multiple roles allowed. Learn more

New members: 729328892908-compute@developer.gserviceaccount.com



Role: Editor

Edit access to all resources.

Condition
Add condition

□ Send notification email
This email will inform members that you've granted them access to this role for "qwiklabs-gcp-03-1df6c0aeaf28"

Task 1. Review Preparations

These preparations should already be have been done:

- Create Cloud Storage bucket
- Clone github repository to Cloud Shell

```
git clone
https://github.com/GoogleCloudPlatform/training-data-analyst
```

- Upgrade packages and install Apache Beam

```
cd training-data-analyst/courses/data_analysis/lab2/python
sudo ./install_packages.sh
```

Task 2. Identify Map and Reduce operations

1. In the Cloud Shell code editor navigate to the directory `/training-data-analyst/courses/data_analysis/lab2/python` and view the file `is_popular.py` in the Cloud Shell editor. **Do not make any changes to the code.**

Alternatively, you could view the file with nano. **Do not make any changes to the code.**

```
cd ~/training-data-analyst/courses/data_analysis/lab2/python
nano is_popular.py
```

Can you answer these questions about the file `is_popular.py`?

- What custom arguments are defined?
- What is the default output prefix?
- How is the variable output_prefix in `main()` set?
- How are the pipeline arguments such as `--runner` set?
- What are the key steps in the pipeline?
- Which of these steps happen in parallel?
- Which of these steps are aggregations?

Task 3. Execute the pipeline

1. Run the pipeline locally:

```
cd ~/training-data-analyst/courses/data_analysis/lab2/python
python3 ./is_popular.py
```

1. Identify the output file. It should be **output**<suffix> and could be a sharded file.

```
ls -al /tmp
```

1. Examine the output file, replacing '-*' with the appropriate suffix.

```
cat /tmp/output-*
```

Task 4. Use command line parameters

1. Change the output prefix from the default value:

```
python3 ./is_popular.py --output_prefix=/tmp/myoutput
```

1. What will be the name of the new file that is written out?
2. Note that we now have a new file in the **/tmp** directory:

```
ls -lrt /tmp/myoutput*
```

End your lab

When you have completed your lab, click **End Lab**. Qwiklabs 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.

Last Updated Date: 2020-01-23

Last Tested Date: 2020-01-23