Dataflow: Qwik Start - Python

GSP207

Google Cloud Self-Paced Labs

In this lab you will set up your Python development environment, get the Cloud Dataflow SDK for Python, and run an example pipeline using the Cloud Console.

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

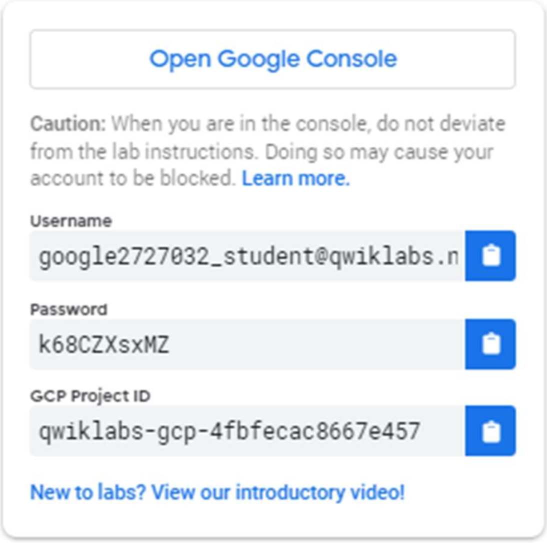
To complete this lab, you need:

- 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.



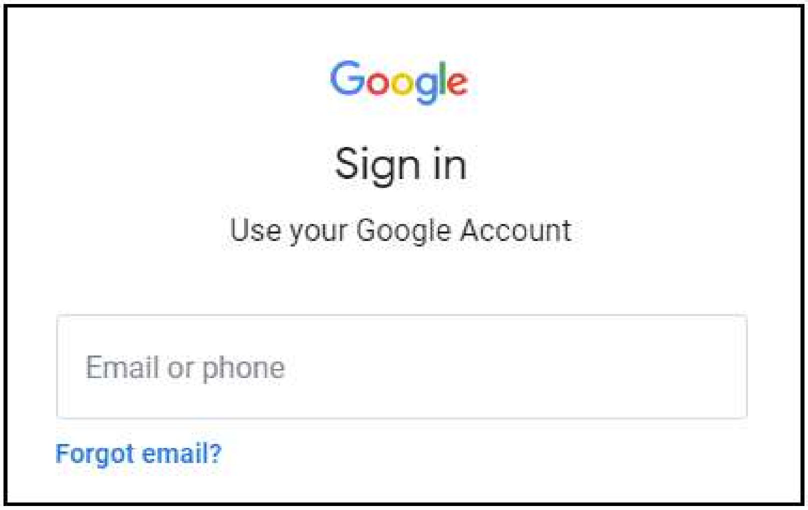
Caution: When you are in the console, do not deviate from the lab instructions. Doing so may cause your account to be blocked. Learn more.

Username
google2727032_student@qwiklabs.n

Password
k68CZXsxMZ

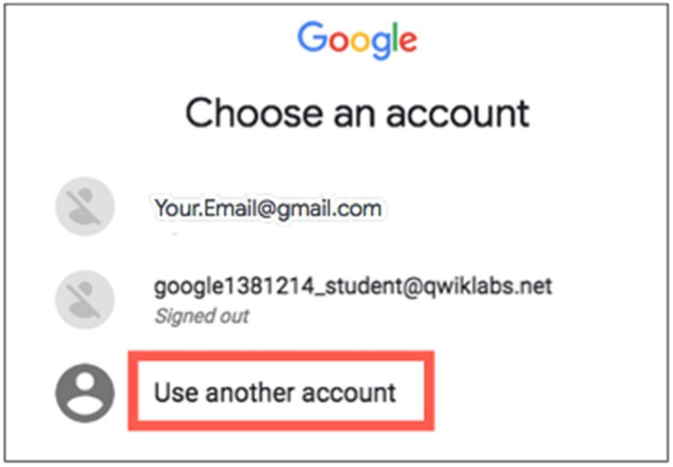
GCP Project ID
qwiklabs-gcp-4fbfecac8667e457

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.

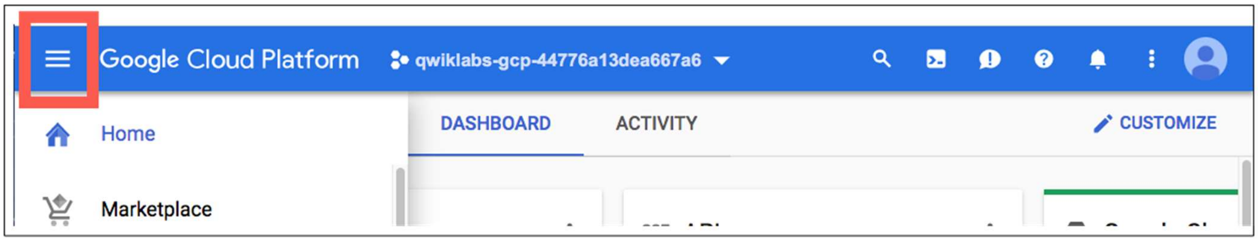
   Important: You must use the credentials from the Connection Details panel. Do not use your Qwiklabs credentials. If you have your own Google Cloud account, do not use it for this lab (avoids incurring charges).

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.

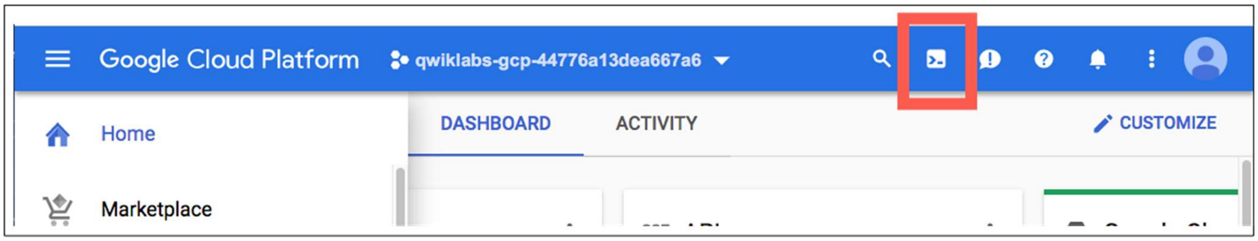
Note: You can view the menu with a list of Google Cloud Products and Services by clicking the **Navigation menu** at the top-left.



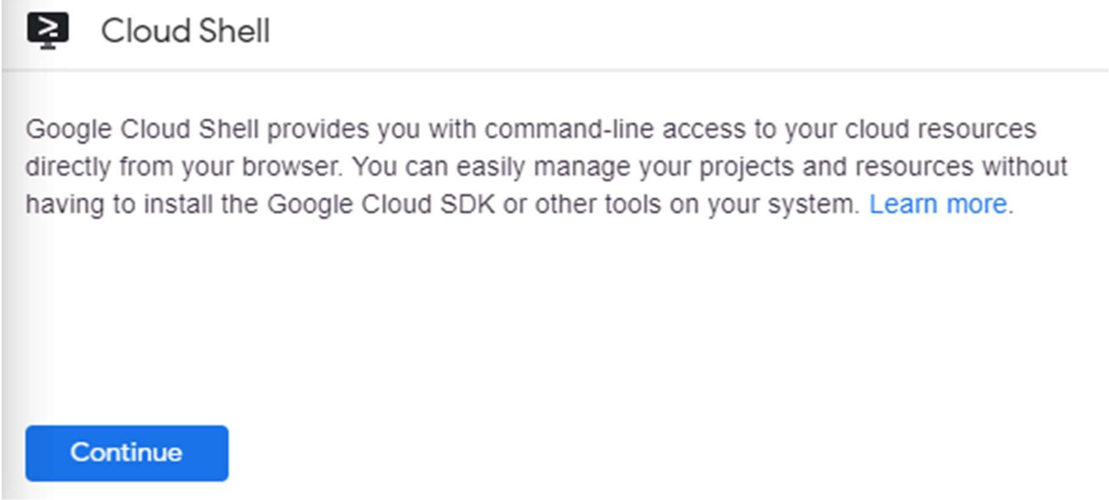
Activate Cloud Shell

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.



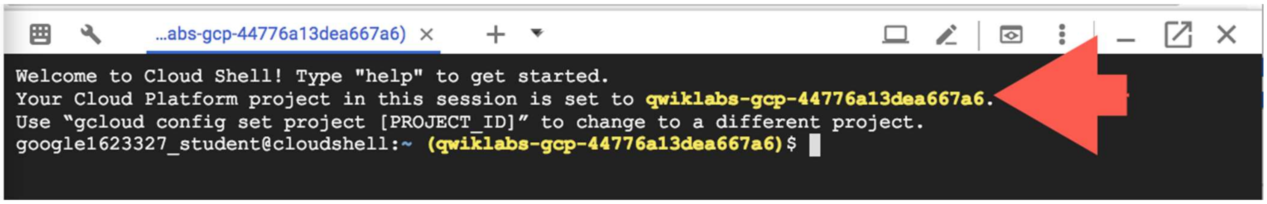
Click **Continue**.



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. 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)

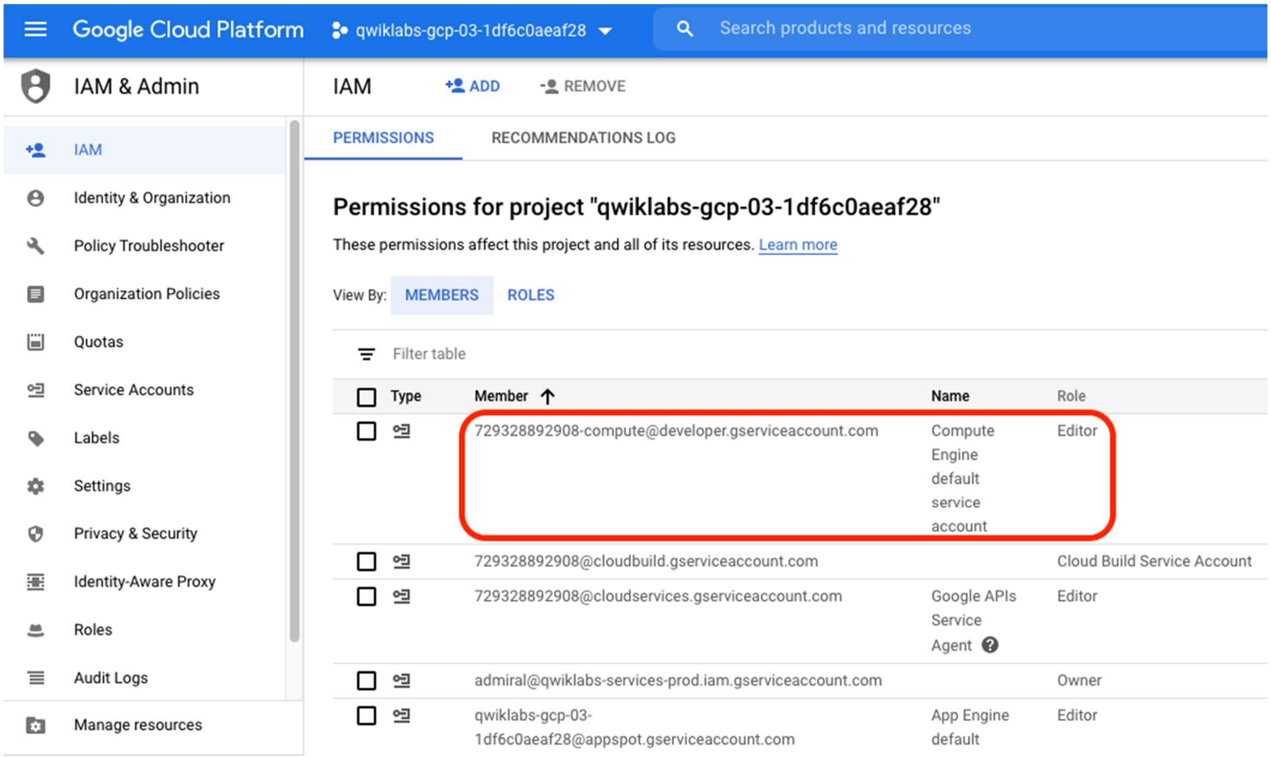
```
[core]
project = qwiklabs-gcp-44776a13dea667a6
```

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**.



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

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

☐ 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**.

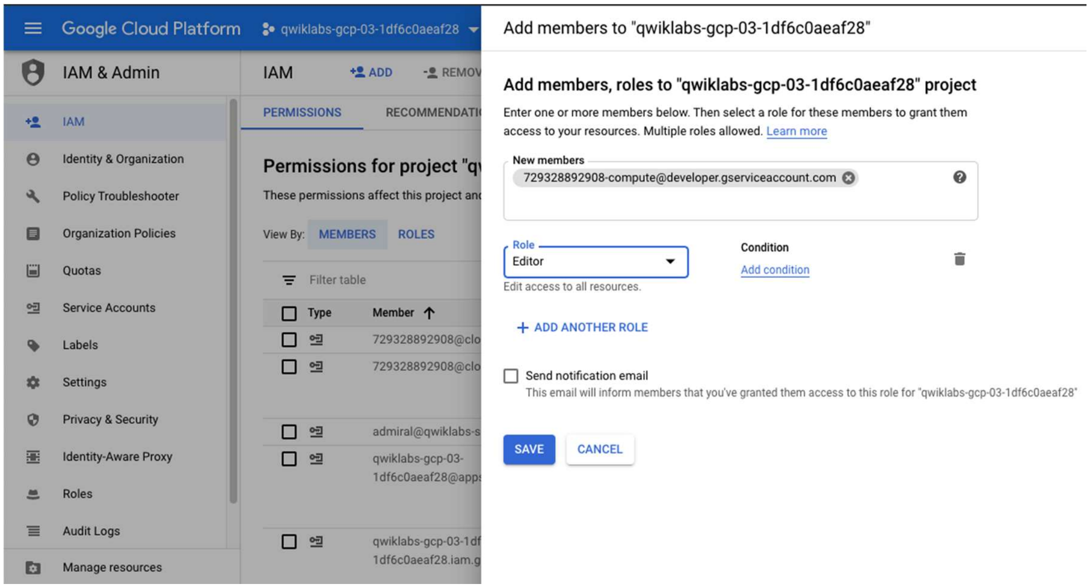
- At the top of the **IAM** page, click **Add**.

- For **New members**, type:

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

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

- For **Role**, select **Project** (or Basic) > **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

+ ADD ANOTHER ROLE

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

SAVE CANCEL

Create a Cloud Storage bucket

1. In the Cloud Console, click on **Navigation menu** and then click on **Storage**.

2. Click **Create bucket**.

3. In the **Create bucket** dialog, specify the following attributes:

- ***Name***: A unique bucket name. Do not include sensitive information in the bucket name, as the bucket namespace is global and publicly visible.
- ***Location type***: Region
- ***Location***: `us-central1`
- A location where bucket data will be stored.

4. Click **Create**.

Test Completed Task

Click **Check my progress** to verify your performed task. If you have completed the task successfully you will granted with an assessment score.

Create a Cloud Storage bucket.

Check my progress

Install pip and the Cloud Dataflow SDK

1. The Cloud Dataflow SDK for Python requires Python version 3.7. Check that you are using Python version 3.7 by running:

```
python3 --version
```

2. Install pip, Python's package manager. Check if you already have pip installed by running:

```
pip3 --version
```

3. After installation, check that you have pip version 7.0.0 or newer. To update pip, run the following command:

```
sudo pip3 install -U pip
```

This step is optional but highly recommended. Install and create a Python virtual environment for initial experiments:
If you do not have virtualenv version 13.1.0 or newer, install it by running:

```
sudo pip3 install --upgrade virtualenv
```

A virtual environment is a directory tree containing its own Python distribution. To create a virtual environment, run the following:

```
virtualenv -p python3.7 env
```

A virtual environment needs to be activated for each shell that will use it. Activating it sets some environment variables that point to the virtual environment's directories. To activate a virtual environment in Bash, run:

```
source env/bin/activate
```

This command sources the script bin/activate under the virtual environment directory you created. For instructions using other shells, see the virtualenv documentation.

4. Install the latest version of the Apache Beam for Python by running the following command from a virtual environment:

```
pip install apache-beam[gcp]
```

You will see some warnings returned that are related to dependencies. It is safe to ignore them for this lab.

5. Run the `wordcount.py` example locally by running the following command:

```
python -m apache_beam.examples.wordcount --output OUTPUT_FILE
```

Note: You installed `google-cloud-dataflow` but are executing `wordcount` with `Apache_beam`. The reason for this is that Cloud Dataflow is a distribution of Apache Beam

You may see a message similar to the following:

```
INFO:root:Missing pipeline option (runner). Executing pipeline using the default
runner: DirectRunner.
INFO:oauth2client.client:Attempting refresh to obtain initial access_token
```

This message can be ignored.

You can now list the files that are on your local cloud environment to get the name of the `OUTPUT_FILE`:

```
ls
```

Copy the name of the `OUTPUT_FILE` and `cat` into it:

```
cat <file name>
```

Your results show each word in the file and how many times it appears.

Run an Example Pipeline Remotely

Set the BUCKET environment variable to the bucket you created earlier.

```
BUCKET=gs://<bucket name provided earlier>
```

Now you'll run the `wordcount.py` example remotely:

```
python -m apache_beam.examples.wordcount --project $DEVSHELL_PROJECT_ID \
  --runner DataflowRunner \
  --staging_location $BUCKET/staging \
  --temp_location $BUCKET/temp \
  --output $BUCKET/results/output \
  --region us-central1
```

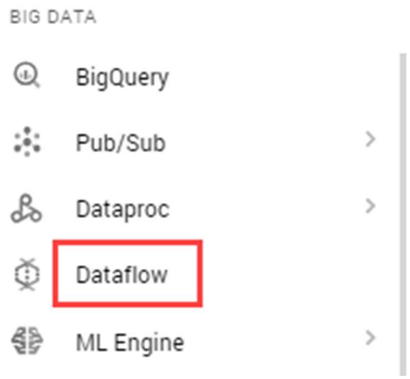
In your output, wait until you see the message:

```
JOB_MESSAGE_DETAILED: Workers have started successfully.
```

Then continue with the lab.

Check that your job succeeded

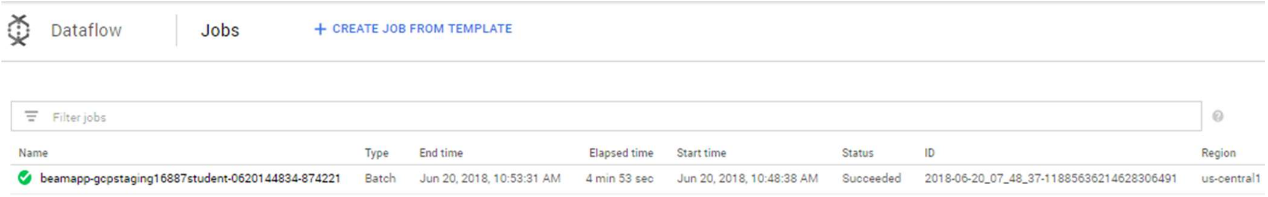
Open the navigation menu and select **Dataflow** from the list of services:



You should see your **wordcount** job with a **status** of **Running** at first.

Click on the name to watch the process. When all the boxes are checked off, you can continue watching the logs in Cloud Shell.

The process is complete when the status is **Succeeded**:



Name	Type	End time	Elapsed time	Start time	Status	ID	Region
✓ beamapp-gcpstaging16887student-0620144834-874221	Batch	Jun 20, 2018, 10:53:31 AM	4 min 53 sec	Jun 20, 2018, 10:48:38 AM	Succeeded	2018-06-20_07_48_37-11885636214628306491	us-central1

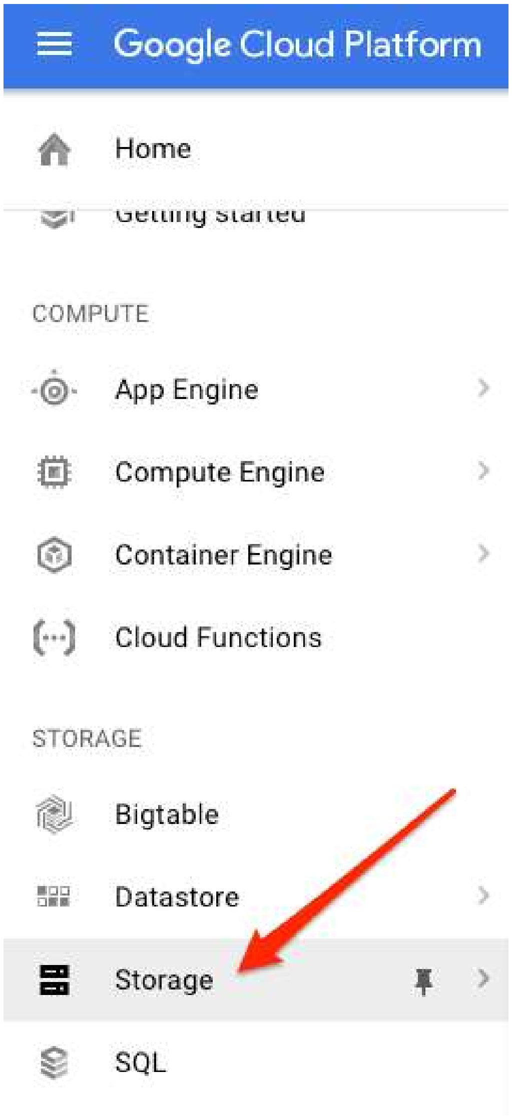
Test Completed Task

Click **Check my progress** to verify your performed task. If you have completed the task successfully you will granted with an assessment score.

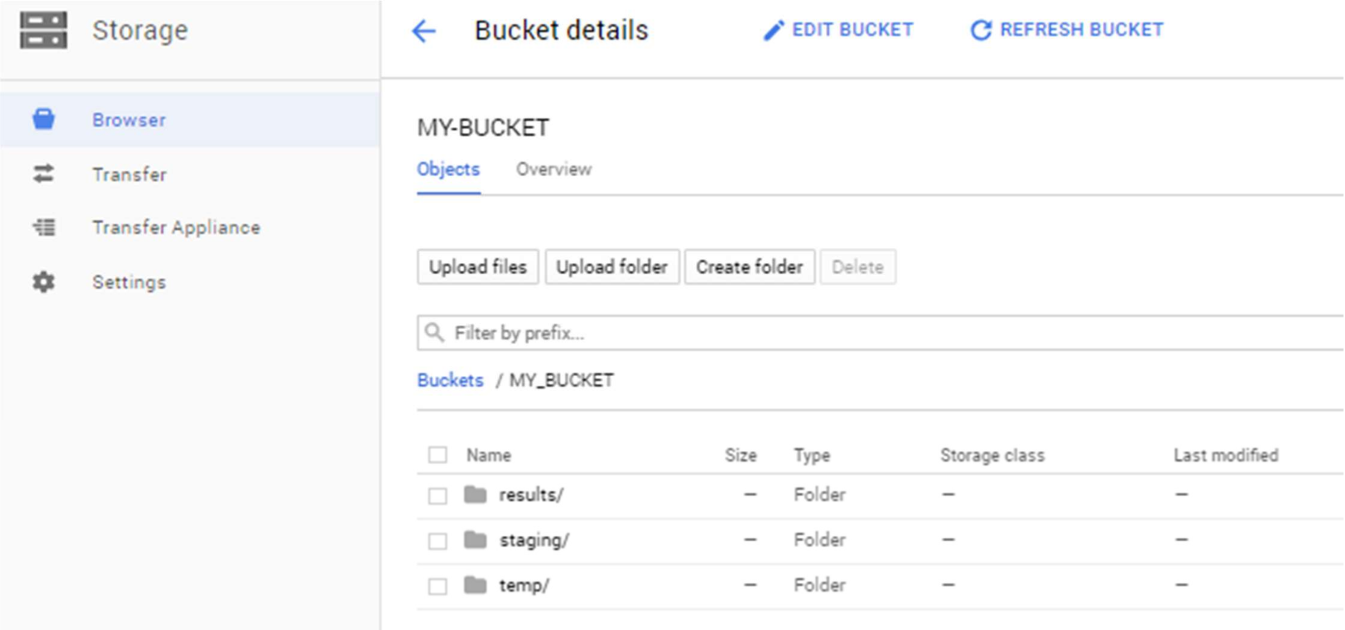
Run an Example Pipeline Remotely.

Check my progress

Click on **Storage** in the Console.

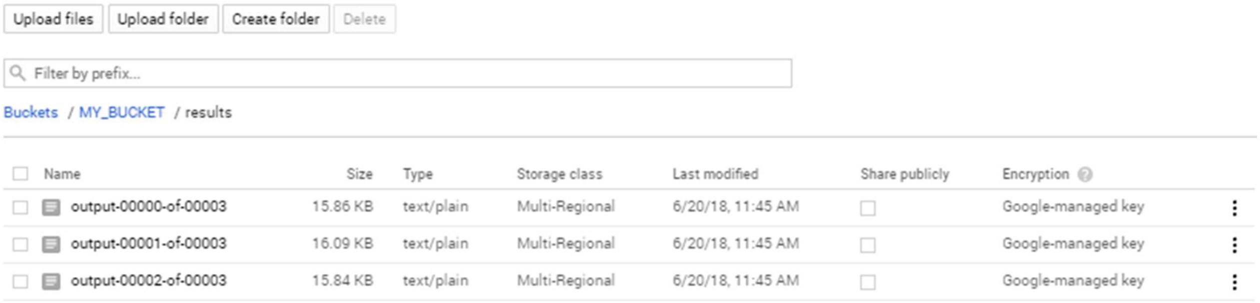


Click on the name of your bucket. In your bucket, you should see the **results** and **staging** directories:



□ Name	Size	Type	Storage class	Last modified
□ results/	–	Folder	–	–
□ staging/	–	Folder	–	–
□ temp/	–	Folder	–	–

Click on the **results** folder and you should see the output files that your job created:



Buckets / MY_BUCKET / results

□ Name	Size	Type	Storage class	Last modified	Share publicly	Encryption
□ output-00000-of-00003	15.86 KB	text/plain	Multi-Regional	6/20/18, 11:45 AM	□	Google-managed key
□ output-00001-of-00003	16.09 KB	text/plain	Multi-Regional	6/20/18, 11:45 AM	□	Google-managed key
□ output-00002-of-00003	15.84 KB	text/plain	Multi-Regional	6/20/18, 11:45 AM	□	Google-managed key

Click on a file to see the word counts it contains.

Test your Understanding

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

Dataflow temp_location must be a valid Cloud Storage URL.

○ 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 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!
To get your own copy of the book this lab is based on: Data Science on the Google Cloud Platform: O'Reilly Media, Inc.

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 October 12, 2020
Lab Last Tested October 12, 2020