

Dataflow: Qwik Start - Python | Google Cloud Skills Boost

Qwiklabs : 10-12 minutes

GSP207



Google Cloud Self-Paced Labs

Overview

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 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.

To complete this lab, you need:

- Access to a standard internet browser (Chrome browser recommended).

Note: Use an Incognito or private browser window to run this lab. This prevents any conflicts between your personal account and the Student account, which may cause extra charges incurred to your personal account.

- Time to complete the lab---remember, once you start, you cannot pause a lab.

Note: If you already have your own personal Google Cloud account or project, do not use it for this lab to avoid extra charges to your account.

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 the **Lab Details** panel with the following:
 - The **Open Google Console** button
 - Time remaining
 - The temporary credentials that you must use for this lab

- Other information, if needed, to step through this lab
2. Click **Open Google Console**. The lab spins up resources, and then opens another tab that shows the **Sign in** page.

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

Note: If you see the **Choose an account** dialog, click **Use Another Account**.

3. If necessary, copy the **Username** from the **Lab Details** panel and paste it into the **Sign in** dialog. Click **Next**.

4. Copy the **Password** from the **Lab Details** panel and paste it into the **Welcome** dialog. Click **Next**.

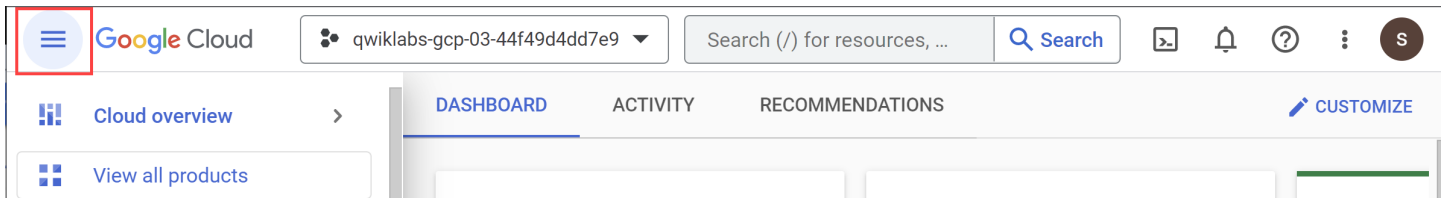
Important: You must use the credentials from the left panel. Do not use your Google Cloud Skills Boost credentials. **Note:** Using your own Google Cloud account for this lab may incur extra charges.

5. 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.

1. Click **Activate Cloud Shell**  at the top of the Google Cloud console.

When you are connected, you are already authenticated, and the project is set to your **PROJECT_ID**. The output contains a line that declares the **PROJECT_ID** for this session:

Your Cloud Platform project in this session is set to **YOUR_PROJECT_ID**

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

2. (Optional) You can list the active account name with this command:

gcloud auth list

3. Click **Authorize**.

4. Your output should now look like this:

Output:

ACTIVE: * ACCOUNT: student-01-xxxxxxxxxxxx@qwiklabs.net To set the active account, run: \$ gcloud config set account `ACCOUNT`

5. (Optional) 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
```

Note: For full documentation of gcloud, in Google Cloud, refer to [the gcloud CLI overview guide](#).

Set the region

- In Cloud Shell, run the following command to set the project region for this lab:

```
gcloud config set compute/region {{{project_0.default_region | "REGION"}}
```

Ensure that the Dataflow API is successfully enabled

To ensure access to the necessary API, restart the connection to the Dataflow API.

1. In the Cloud Console, enter "Dataflow API" in the top search bar. Click on the result for **Dataflow API**.
2. Click **Manage**.
3. Click **Disable API**.

If asked to confirm, click **Disable**.

4. Click **Enable**.

When the API has been enabled again, the page will show the option to disable.

Task 1. Create a Cloud Storage bucket

1. On the **Navigation menu** (≡), click **Cloud Storage > Buckets**.
2. Click **Create bucket**.
3. In the **Create bucket** dialog, specify the following attributes:

- **Name:** To ensure a unique bucket name, use the following name: **-bucket**. Note that this name does not include sensitive information in the bucket name, as the bucket namespace is global and publicly

visible.

- **Location type:** Multi-region
- **Location:** us
- A location where bucket data will be stored.

4. Click **Create**.

5. If Prompted Public access will be prevented, click **Confirm**.

Test completed task

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

Create a Cloud Storage bucket.

Task 2. Install pip and the Cloud Dataflow SDK

1. The latest Cloud Dataflow SDK for Python requires a Python version ≥ 3.7 .

To ensure you are running the process with the correct version, run the Python3.9 Docker Image:

```
docker run -it -e DEVSHELL_PROJECT_ID=$DEVSHELL_PROJECT_ID python:3.9 /bin/bash
```

This command pulls a Docker container with the latest stable version of Python 3.9 and then opens up a command shell for you to run the following commands inside your container.

2. After the container is running, install the latest version of the Apache Beam for Python by running the following command from a virtual environment:

```
pip install 'apache-beam[gcp]==2.42.0'
```

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

3. 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](https://github.com/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.

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

ls

5. 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.

Task 3. Run an example pipeline remotely

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

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

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

```
python -m apache_beam.examples.wordcount --project $DEVSHIELD_PROJECT_ID \ --runner  
DataflowRunner \ --staging_location $BUCKET/staging \ --temp_location $BUCKET/temp \ --output  
$BUCKET/results/output \ --region {{{project_0.default_region | "filled in at lab start"}}
```

In your output, wait until you see the message:

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

Then continue with the lab.

Task 4. Check that your job succeeded

1. Open the Navigation menu and click **Dataflow** from the list of services.

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

2. 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**.

Test completed task

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

Run an Example Pipeline Remotely.

3. Click **Navigation menu > Cloud Storage** in the Cloud Console.
4. Click on the name of your bucket. In your bucket, you should see the **results** and **staging** directories.
5. Click on the **results** folder and you should see the output files that your job created:
6. Click on a file to see the word counts it contains.

Task 5. Test your understanding

Below is a multiple choice question to reinforce your understanding of this lab's concepts. Answer it to the best of your abilities.

Congratulations!

Finish your quest

This self-paced lab is part of the [Baseline: Data, ML, AI](#) quest. A quest is a series of related labs that form a learning path. Completing this quest earns you a badge 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](#) or any quest that contains this lab and get immediate completion credit. See the [Google Cloud Skills Boost catalog](#) to see all available 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 [Google Cloud Skills Boost 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 and 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: May 4, 2023

Lab Last Tested: May 4, 2023

Copyright 2023 Google LLC All rights reserved. Google and the Google logo are trademarks of Google LLC. All other company and product names may be trademarks of the respective companies with which they are associated.