MapReduce in Dataflow (Python)

Start Lab 02:00:00

MapReduce in Dataflow (Python)

2 hours Free ★★★★☆ Rate Lab

Duration is 1 min

In this lab, you learn how to use pipeline options and carry out Map and Reduce operations in Dataflow.

What you need

You must have completed Lab 0 and have the following:

- Logged into GCP Console with your Qwiklabs generated account

What you learn

In this lab, you learn how to:

- Use pipeline options in Dataflow
- Carry out mapping transformations

Introduction

Duration is 1 min

The goal of this lab is to learn how to write MapReduce operations using Dataflow.

Setup

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

2. Note the lab's access time (for example, 02:00:00 and make sure you can finish in that time block.

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. You will use them to sign in to the Google Cloud Console.

Open Google 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

Password
TG959yrKDX

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.

If you use other credentials, you'll get errors or **incur charges**.

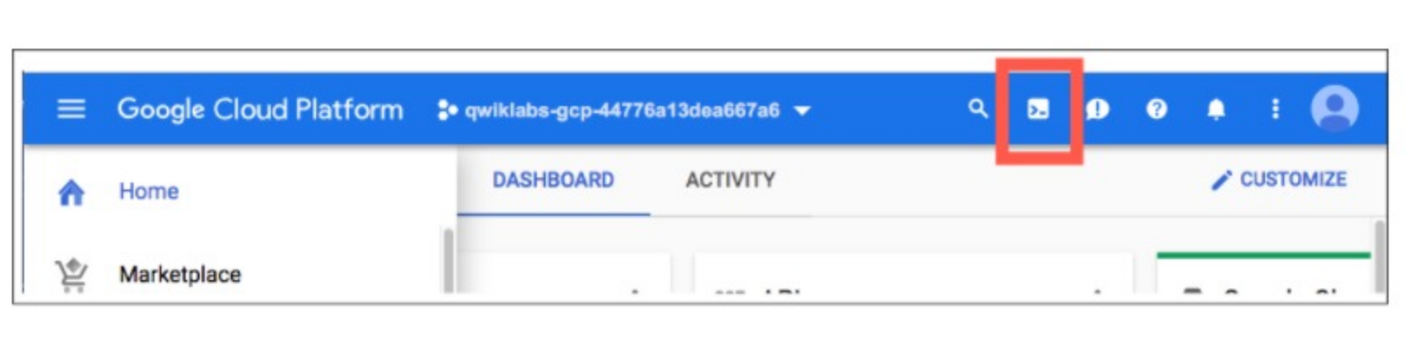
7. Accept the terms and skip the recovery resource page.

Do not click **End Lab** unless you are finished with the lab or want to restart it. This clears your work and removes the project.

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.



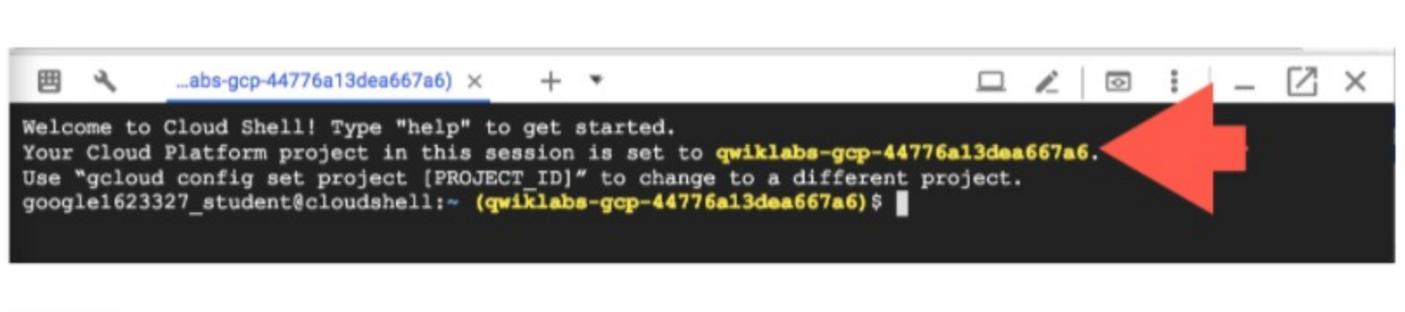
Click **Continue**.

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

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
```

```
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)

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

Launch Google Cloud Shell Code Editor

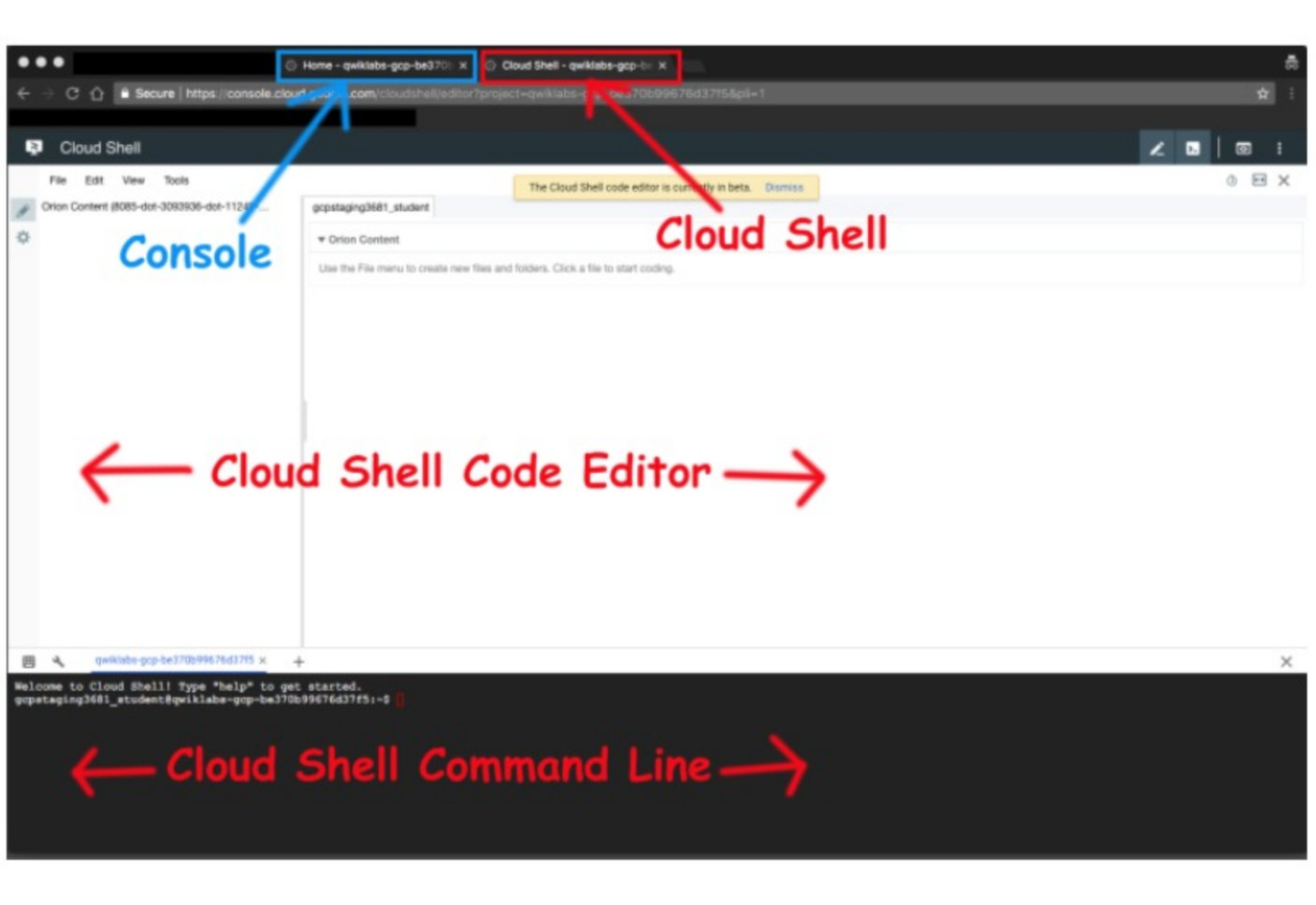
Use the Google Cloud Shell Code Editor to easily create and edit directories and files in the Cloud Shell instance.

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

Open Editor

You now have three interfaces available:

- The Cloud Shell Code Editor



Identify Map and Reduce operations

Duration is 5 min

Step 1

In CloudShell clone the source repo which has starter scripts for this lab:

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

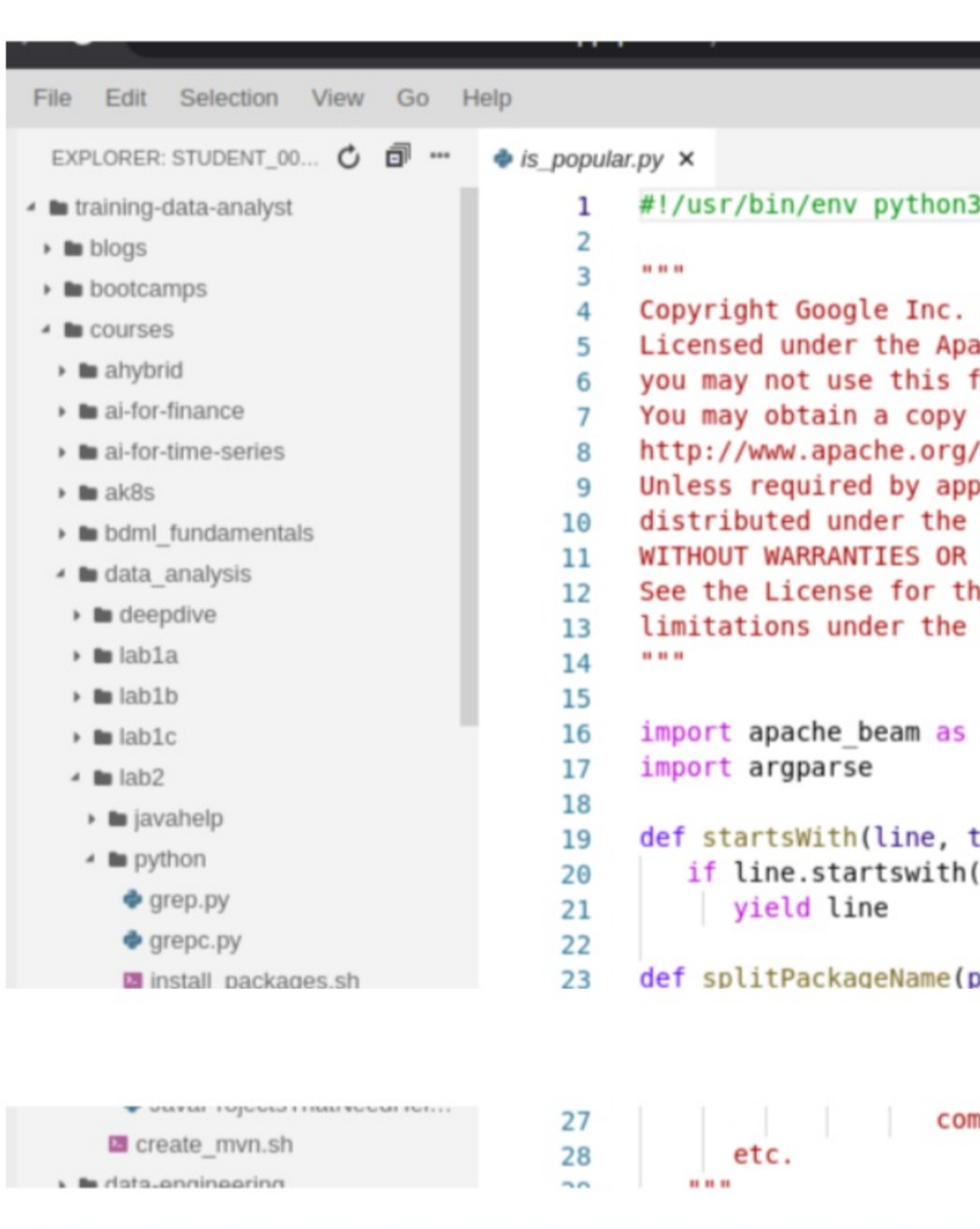
Then navigate to the code for this lab.

```
cd training-data-analyst/courses/data_analysis/lab2/python
```

Step 2

Click on the **Refresh** icon.

browser editor or with the command line using nano:




```
nano is_popular.py
```

Step 3

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? ____________________

Step 4

What are the key steps in the pipeline?

Which of these steps happen in parallel? ____________________

Which of these steps are aggregations? ____________________

Execute the pipeline

Duration is 2 min

Step 1

Install the necessary dependencies for Python dataflow:

```
sudo ./install_packages.sh
```

```
pip3 -V
```

If not, open a new CloudShell tab and it should pick up the updated pip.

Step 2

Run the pipeline locally:

```
python3 ./is_popular.py
```

Note: If you see an error that says "`No handlers could be found for logger "oauth2client.contrib.multistore_file"`, you may ignore it. The error is simply saying that logging from the oauth2 library will go to stderr.

Step 3

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

Use command line parameters

Duration is 2 min

Step 1

Change the output prefix from the default value:

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

What will be the name of the new file that is written out?

Step 2

Note that we now have a new file in the /tmp directory:

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

What you learned

Duration is 1 min

In this lab, you:

- Used pipeline options in Dataflow
- Identified Map and Reduce operations in the Dataflow pipeline

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.