Serverless Data Analysis with Beam: MapReduce in Beam (Java)

1 hour 30 minutes No cost

Overview

In this lab, you will identify Map and Reduce operations, execute the pipeline, and 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 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.

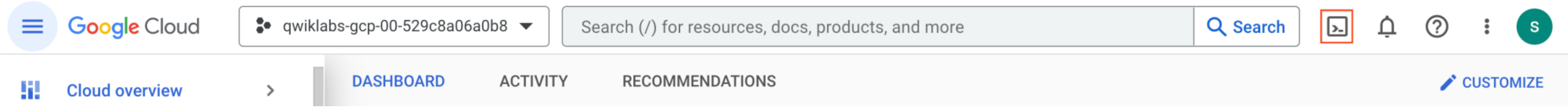
Note: Do not click **End Lab** unless you have finished the lab or want to restart it. This clears your work and removes the project.

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.

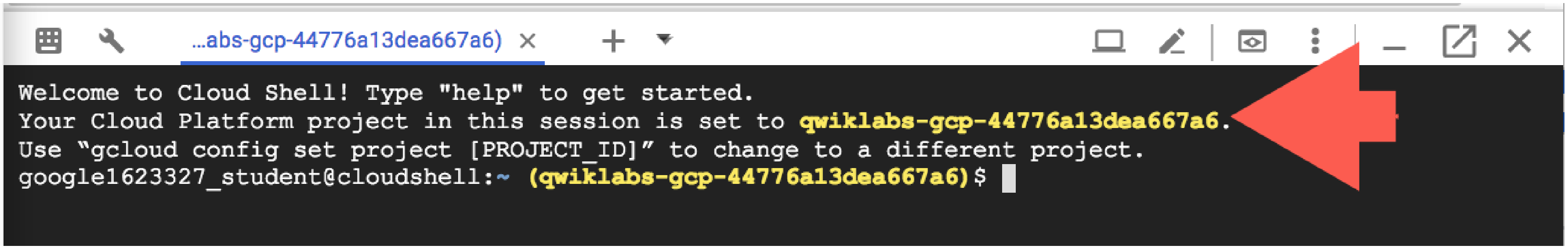
Google Cloud Shell provides command-line access to your Google Cloud resources.

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:

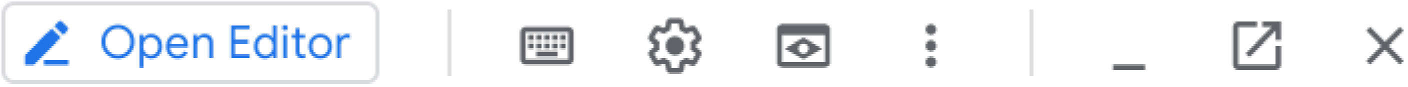
```
[core] project = qwiklabs-gcp-44776a13dea667a6
```

Note: Full documentation of **gcloud** is available in the [gcloud CLI overview guide].

Launch Google Cloud Shell Code Editor

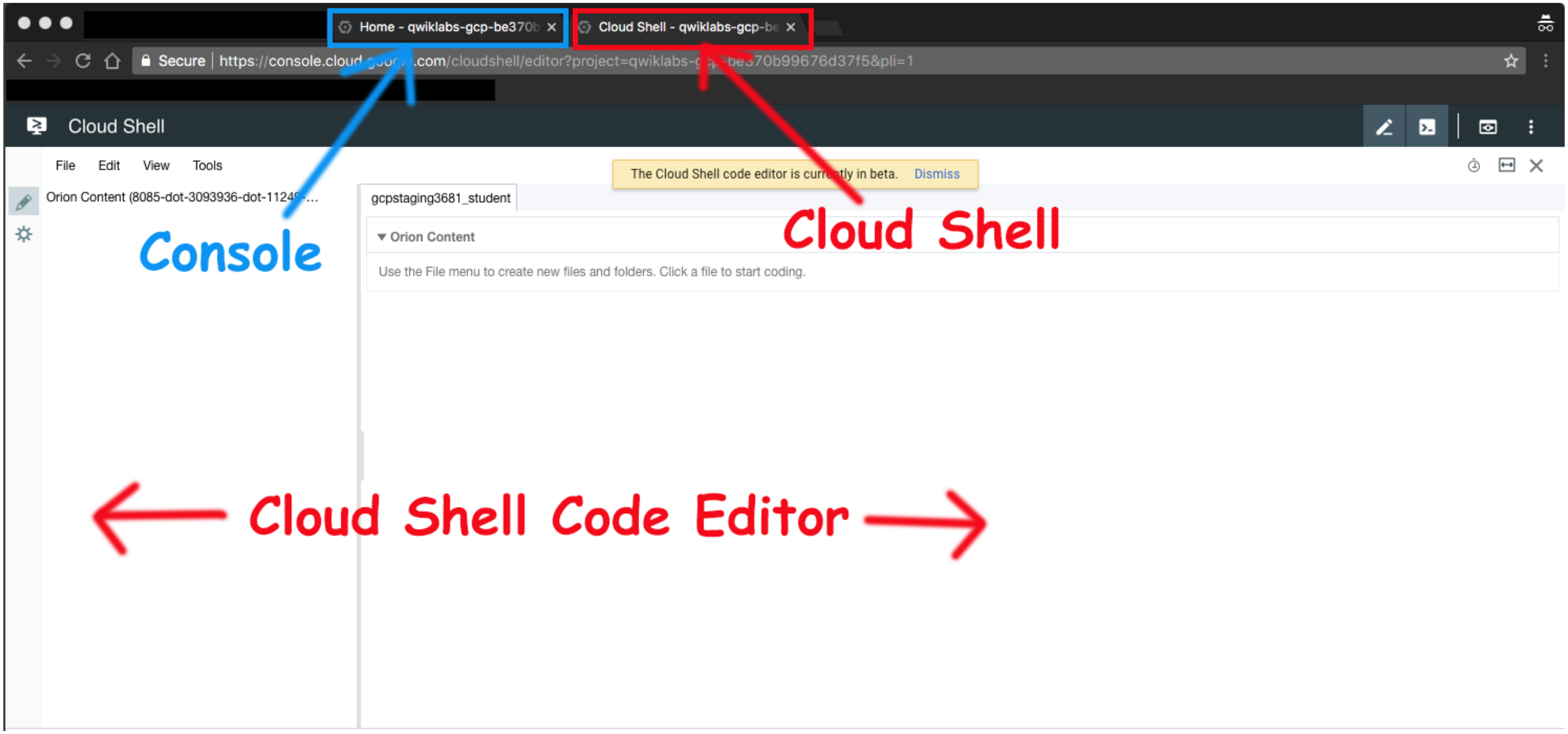
Use the Google Cloud Shell Code Editor to easily create and edit directories and files in the Cloud Shell instance.

- Once you activate the Google Cloud Shell, click **Open editor** 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]
Home - qwiklabs-gcp-be370b
Cloud Shell - qwiklabs-gcp-be
Secure | https://console.cloud.google.com/cloudshell/editor?project=qwiklabs-gcp-be370b99676d37f5&pli=1
[redacted]
Cloud Shell
File
Edit
View
Tools
Orion Content (8085-dot-3093936-dot-1124...
gcpstaging3681_student
Dismiss
Orion Content
Console
Cloud Shell
Cloud Shell Code Editor

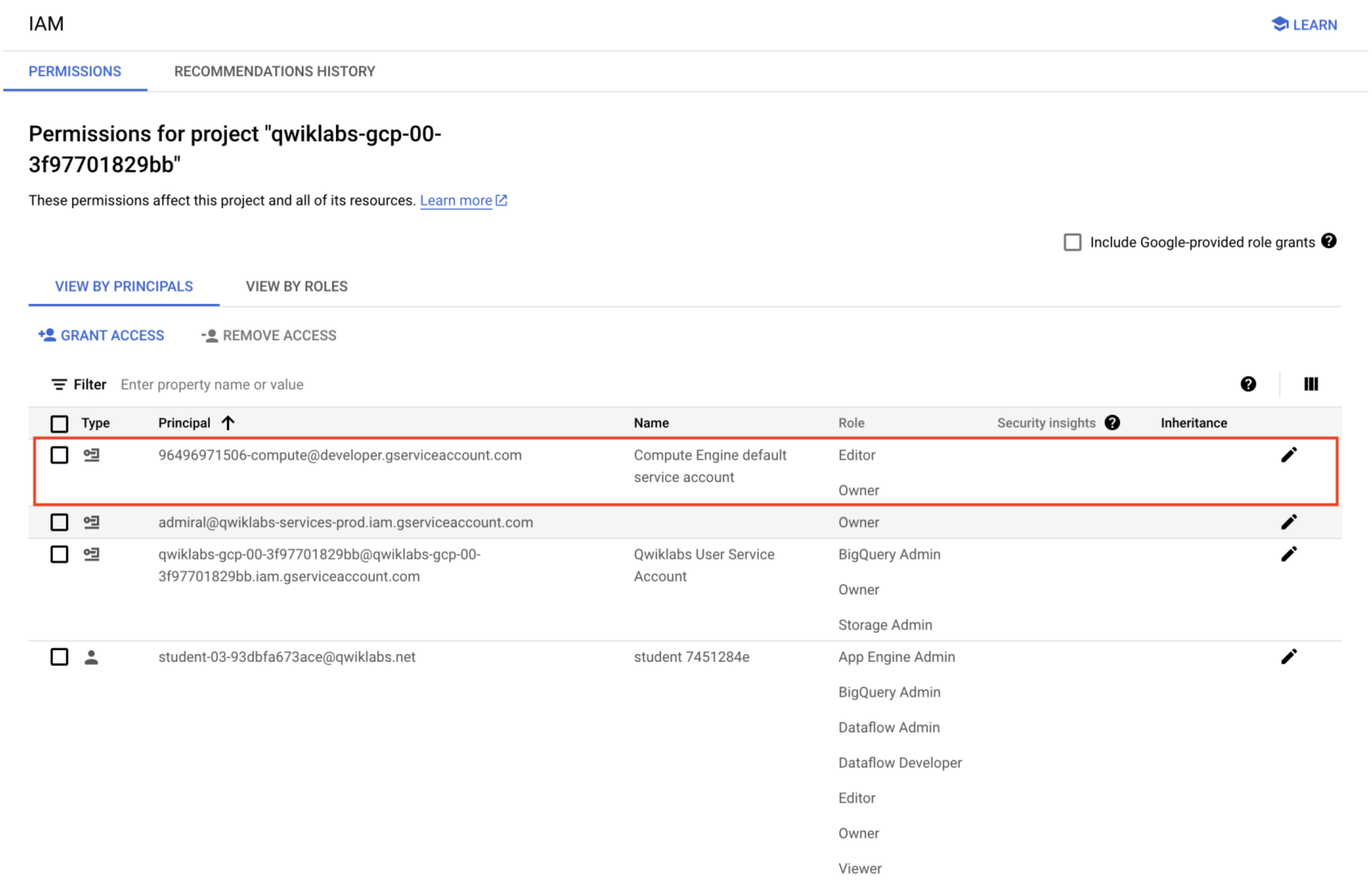
The Cloud Shell code editor is currently in beta.

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

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** > **Cloud Overview** > **Dashboard**.



IAM

LEARN

PERMISSIONS RECOMMENDATIONS HISTORY

Permissions for project "qwiklabs-gcp-00-3f97701829bb"

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

□ Include Google-provided role grants

VIEW BY PRINCIPALS VIEW BY ROLES

GRANT ACCESS REMOVE ACCESS

Filter Enter property name or value

□ Type	Principal ↑	Name	Role	Security insights	Inheritance
□	96496971506-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor Owner		
□	admiral@qwiklabs-services-prod.iam.gserviceaccount.com		Owner		
□	qwiklabs-gcp-00-3f97701829bb@qwiklabs-gcp-00-3f97701829bb.iam.gserviceaccount.com	Qwiklabs User Service Account	BigQuery Admin Owner Storage Admin		
□	student-03-93dbfa673ace@qwiklabs.net	student 7451284e	App Engine Admin BigQuery Admin Dataflow Admin Dataflow Developer Editor Owner Viewer		

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 **Cloud Overview** > **Dashboard**.
2. Copy the project number (e.g. `729328892908`).
3. On the **Navigation menu**, select **IAM & Admin** > **IAM**.
4. At the top of the roles table, below **View by Principals**, click **Grant Access**.
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**.

Task 1. Lab preparations

Specific steps must be completed to successfully execute this lab:

1. Create Cloud Storage bucket (which was completed for you automatically when the lab environment started).
2. On the Google Cloud Console title bar, click **Activate Cloud Shell**. If prompted, click **Continue**. Clone the lab code github repository using the following command:

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

Task 2. Identify Map and Reduce operations

- In the Cloud Shell code editor navigate to the directory `/training-data-analyst/courses/data_analysis/lab2/javahelp/src/main/java/com/google/cloud/training/dataanalyst/javahelp` and view the file `IsPopular.java` in the Cloud Shell editor.

Note: Do not make any changes to the code.

Alternatively, you could view the file with nano:

Note: Do not make any changes to the code. cd ~/training-data-analyst/courses/data_analysis/lab2/javahelp/src/main/java/com/google/cloud/training/dataanalyst/javahelp nano IsPopular.java **Note:** Normally, you would develop this Java code in an Integrated Development Environment such as Eclipse or IntelliJ (not in CloudShell).

Can you answer these questions about the file `IsPopular.java`?

- What getX() methods are present in the class MyOptions?
- What is the default output prefix?
- How is the variable outputPrefix in main() 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. Copy and paste the following Maven command in Cloud Shell:

```
export PATH=/usr/lib/jvm/java-8-openjdk-amd64/bin/:$PATH cd ~/training-data-analyst/courses/data_analysis/lab2/javahelp mvn compile -e exec:java \ -Dexec.mainClass=com.google.cloud.training.dataanalyst.javahelp.IsPopular
```

Note: It will take 4-5 mintues to complete the process.

2. Examine the output file:

```
cat /tmp/output.csv
```

Task 4. Use command line parameters

1. Change the output prefix from the default value:

```
mvn compile -e exec:java \ -Dexec.mainClass=com.google.cloud.training.dataanalyst.javahelp.IsPopular \ -Dexec.args="--outputPrefix=/tmp/myoutput"
```

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

```
ls -lrt /tmp/*.csv
```

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.