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

## 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

### **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.



1. 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:



**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

Copied!

content\_copy

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

Copied!

content\_copy

Output:

[core]

project = <project\_ID>

Example output:

[core]

project = qwiklabs-gcp-44776a13dea667a6

Full documentation of **gcloud** is available on [Google Cloud gcloud Overview](https://cloud.google.com/sdk/gcloud).

### **Launch Google Cloud Shell Code Editor**

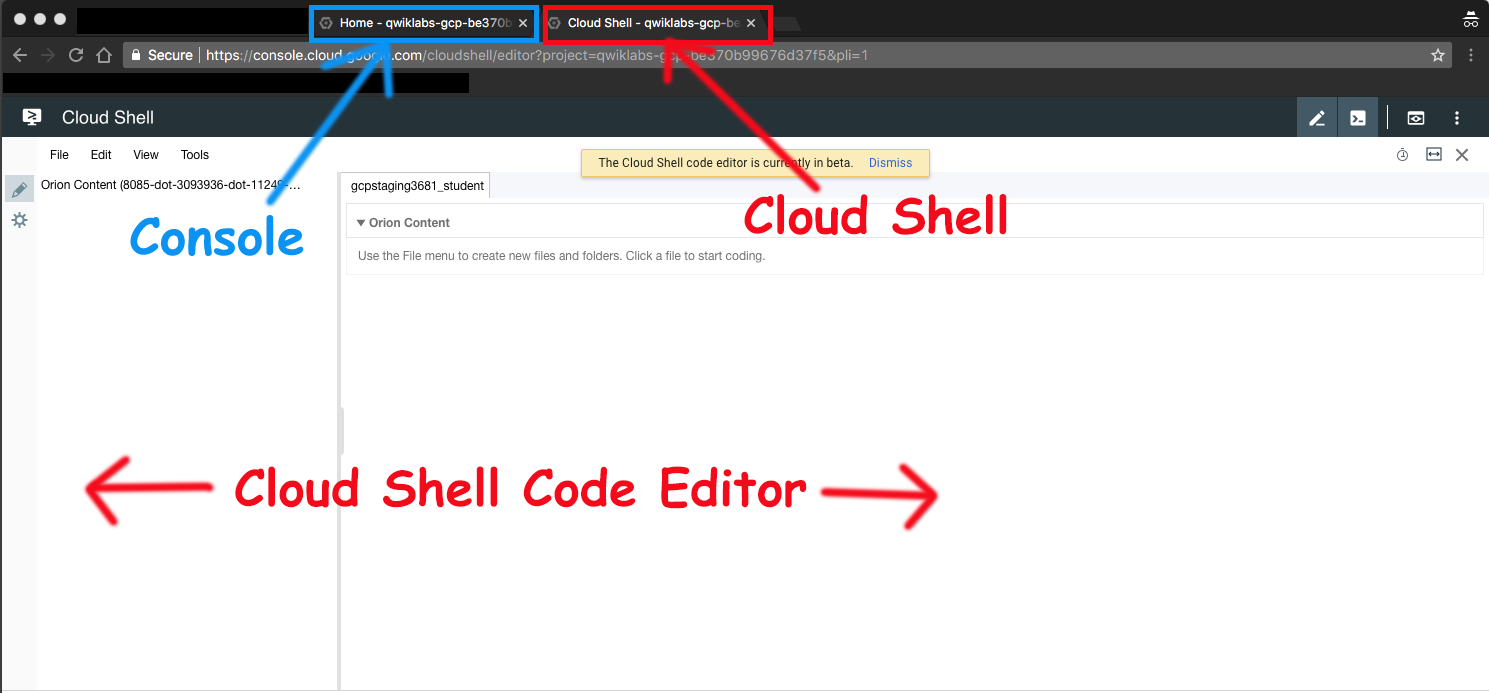
Use the Google Cloud Shell Code Editor to easily create and edit directories and files in the Cloud Shell instance.

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:

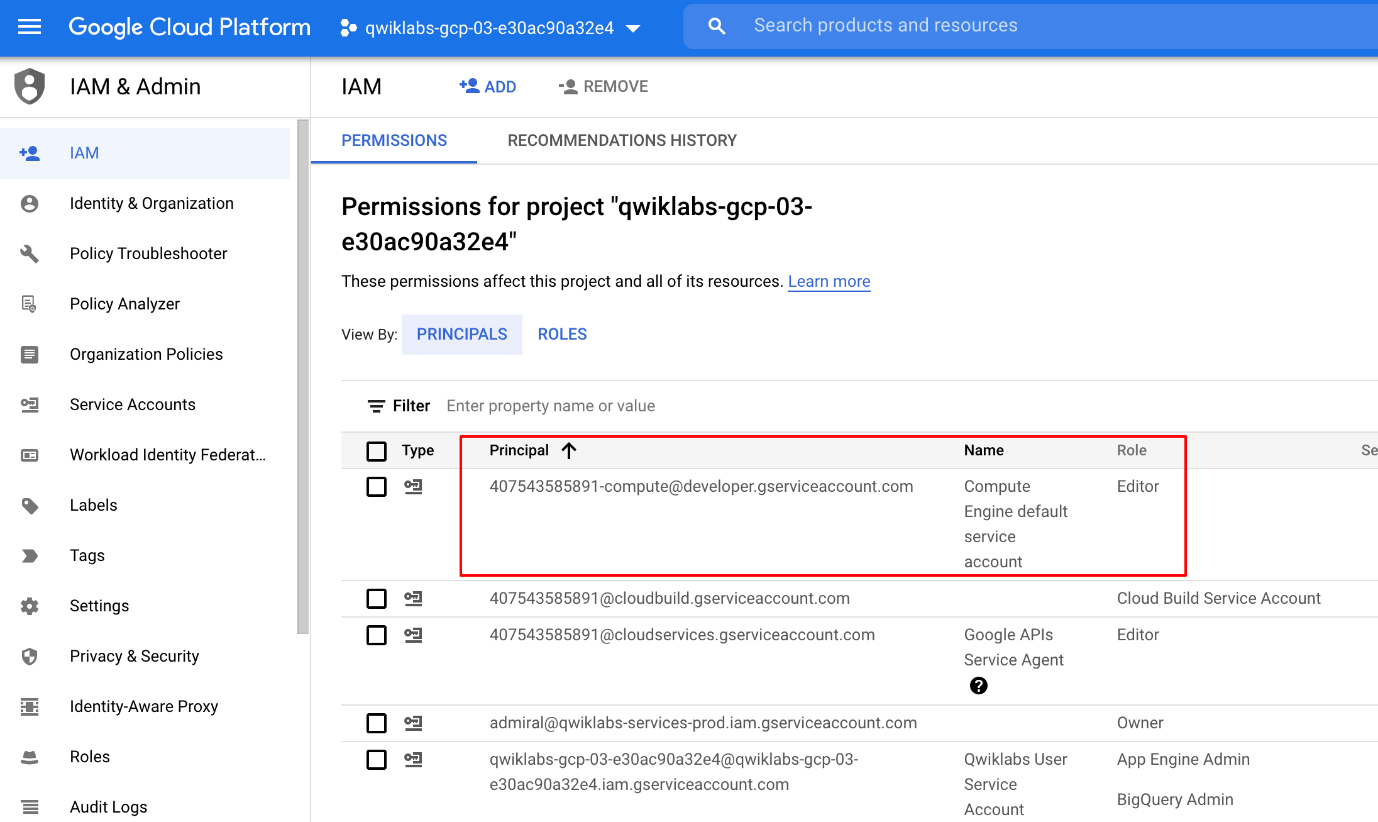
* The Cloud Shell Code Editor
* Console (By clicking on the tab). You can switch back and forth between the Console and Cloud Shell by clicking on the tab.
* The Cloud Shell Command Line (By clicking on **Open Terminal** in the Console)



### **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** (Navigation menu icon), 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**.



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 principals**, type:

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

Copied!

content\_copy

Replace {project-number} with your project number.

* For **Role**, select **Project** (or Basic) > **Editor**. Click **Save**.

## Task 1. Lab Preparations

Specific steps must be completed to successfully execute this lab:

* Create Cloud Storage bucket (which was complete for you automatically when the lab environment started)
* Clone github [repository](https://github.com/GoogleCloudPlatform/training-data-analyst) to Cloud Shell

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

Copied!

content\_copy

## 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/javahelp/src/main/java/com/google/cloud/training/dataanalyst/javahelp and view the file IsPopular.java 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/javahelp/src/main/java/com/google/cloud/training/dataanalyst/javahelp

nano IsPopular.java

Copied!

content\_copy

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

Copied!

content\_copy

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

1. Examine the output file:

cat /tmp/output.csv

Copied!

content\_copy

## 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"

Copied!

content\_copy

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

ls -lrt /tmp/\*.csv

Copied!

content\_copy

## End your lab