

# MapReduce in Beam (Python) 2.5

## 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.

## 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**.

- 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 "qwklabs-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 <span>?</span>	Inheritance
<input type="checkbox"/>	96496971506-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor Owner		
<input type="checkbox"/>	admiral@qwklabs-services-prod.iam.gserviceaccount.com		Owner		
<input type="checkbox"/>	qwklabs-gcp-00-3f97701829bb@qwklabs-gcp-00-3f97701829bb.iam.gserviceaccount.com	Qwiklabs User Service Account	BigQuery Admin Owner Storage Admin		
<input type="checkbox"/>	student-03-93dbfa673ace@qwklabs.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.

- In the Google Cloud console, on the **Navigation menu**, click **Cloud Overview > Dashboard**.
- Copy the project number (e.g. 729328892908).
- On the **Navigation menu**, select **IAM & Admin > IAM**.
- At the top of the roles table, below **View by Principals**, click **Grant Access**.
- For **New principals**, type:

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


- 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.

Open the SSH terminal and connect to the training VM

You will be running all code from a curated training VM.

1. In the Console, on the **Navigation menu** () , click **Compute Engine > VM instances**.
2. Locate the line with the instance called **training-vm**.
3. On the far right, under **Connect**, click on **SSH** to open a terminal window.
4. In this lab, you will enter CLI commands on the **training-vm**.

Clone the training github repository

- In the **training-vm** SSH terminal enter the following command:

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

## Task 2. Identify map and reduce operations

- Return to the **training-vm** SSH terminal and navigate to the directory `/training-data-analyst/courses/data_analysis/lab2/python` and view the file `is_popular.py` with Nano. **Do not make any changes to the code**. Press **Ctrl+X** to exit Nano.

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

```
nano is_popular.py
```

Can you answer these questions about the file `is_popular.py`?

- 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?
- 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. In the **training-vm** SSH terminal, run the pipeline locally:

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

2. Identify the output file. It should be **output**<suffix> and could be a sharded file:

```
ls -al /tmp
```

3. Examine the output file, replacing '\*' with the appropriate suffix:

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

#### **Task 4. Use command line parameters**

1. In the **training-vm** SSH terminal, change the output prefix from the default value:

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

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

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

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

#### **End your lab**