Simple Dataflow Pipeline (Python) 2.5

configuration(creating training vm , create bucket)

- ✓ Check project permissions, **Enable Dataflow API (see appendix)**
- ✓ **Open the SSH terminal and connect to the training VM**
  Compute Engine > **VM instances > training-vm > Connect**
- ✓ In **training-vm** SSH terminal Download Code Repository
  git clone https://github.com/GoogleCloudPlatform/training-data-analyst
- ✓ **Create a Cloud Storage bucket**
  Cloud Storage > **Browser > Create Bucket**
  Name :<your unique bucket name (Project ID)>
  Location type：Multi-Region
  Location ： <Your location>
- ✓ In **training-vm** SSH terminal init bucket variable

```
BUCKET="<your unique bucket name (Project ID)>"
echo $BUCKET
```

Pipeline filtering

- ✓ **In training-vm** SSH terminal change directory and show code source and than Press **Ctrl+X** to exit Nano.

  ```
  cd ~/training-data-analyst/courses/data_analysis/lab2/python
  nano grep.py
  ```

- ✓ Can you answer these questions about the file grep.py?
  - •What files are being read?
  - •What is the search term?
  - •Where does the output go?

  There are three transforms in the pipeline:
  - •What does the transform do?
  - •What does the second transform do?
  - •Where does its input come from?
  - •What does it do with this input?
  - •What does it write to its output?
  - •Where does the output go to?
  - •What does the third transform do?

Execute the pipeline locally

1.In the **training-vm** SSH terminal, locally execute grep.py.

python3 grep.py

The output file will be output.txt. If the output is large enough, it will be sharded into separate parts with names like: output-00000-of-00001.

2.Locate the correct file by examining the file's time.

ls -al /tmp

3.Examine the output file(s).

4.You can replace "-*" below with the appropriate suffix.

cat /tmp/output-*

Does the output seem logical?

Execute the pipeline on the cloud

1.Copy some Java files to the cloud. In the **training-vm** SSH terminal, enter the following commmand:
gsutil cp ../javahelp/src/main/java/com/google/cloud/training/dataanalyst/javahelp/*.java gs://$BUCKET/javahelp
2.Using Nano, edit the Dataflow pipeline in grepc.py.
nano grepc.py
3.Replace PROJECT and BUCKET with your Project ID and Bucket name.

PROJECT='qwiklabs-gcp-your-value' BUCKET='qwiklabs-gcp-your-value'
qwiklabs-gcp-04-4491a9a7c668
Save the file and close Nano by pressing the **CTRL+X** key, then press **Y**, and **Enter**.
4.Submit the Dataflow job to the cloud:
python3 grepc.py
Note: You may ignore the message: **WARNING:root:Make sure that locally built Python SDK docker image has Python 3.7 interpreter.** Your Dataflow job will start successfully.Because this is such a small job, running on the cloud will take significantly longer than running it locally (on the order of 7-10 minutes).
5. Monitor job
6. **Cloud Storage > Browser** > **javahelp folder >** output.txt

Monitor job in dataflow

Job summary

Job name	examplejob2
Job ID	2018-02-06_12_47_44-6148155460441137914
Region	us-central1
Job status	Succeeded
SDK version	Google Cloud Dataflow SDK for Python 2.2.0
Job type	Batch
Start time	Feb 6, 2018, 3:47:45 PM
Elapsed time	4 min 58 sec

Autoscaling

Workers	0
Current state	Stopping worker pool.
	Feb 6, 2018 3:47 PM

Lab : MapReduce in Dataflow

Preparations, identify map and reduce

- ✓ Open the SSH terminal and connect to the training VM
  Compute Engine > VM instances > training-vm > Connect
- ✓ In the **training-vm** SSH terminal (Clone the training github repository)
  git clone https://github.com/GoogleCloudPlatform/training-data-analyst
- ✓ **Identify Map and Reduce operations**

In training-vm SSH terminal and navigate to the directory /training-data-analyst/courses/data_analysis/lab2/python than is_popular.py with Nano than **Ctrl+X**
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?

Execute the pipeline , Use command line parameters

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-*

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*

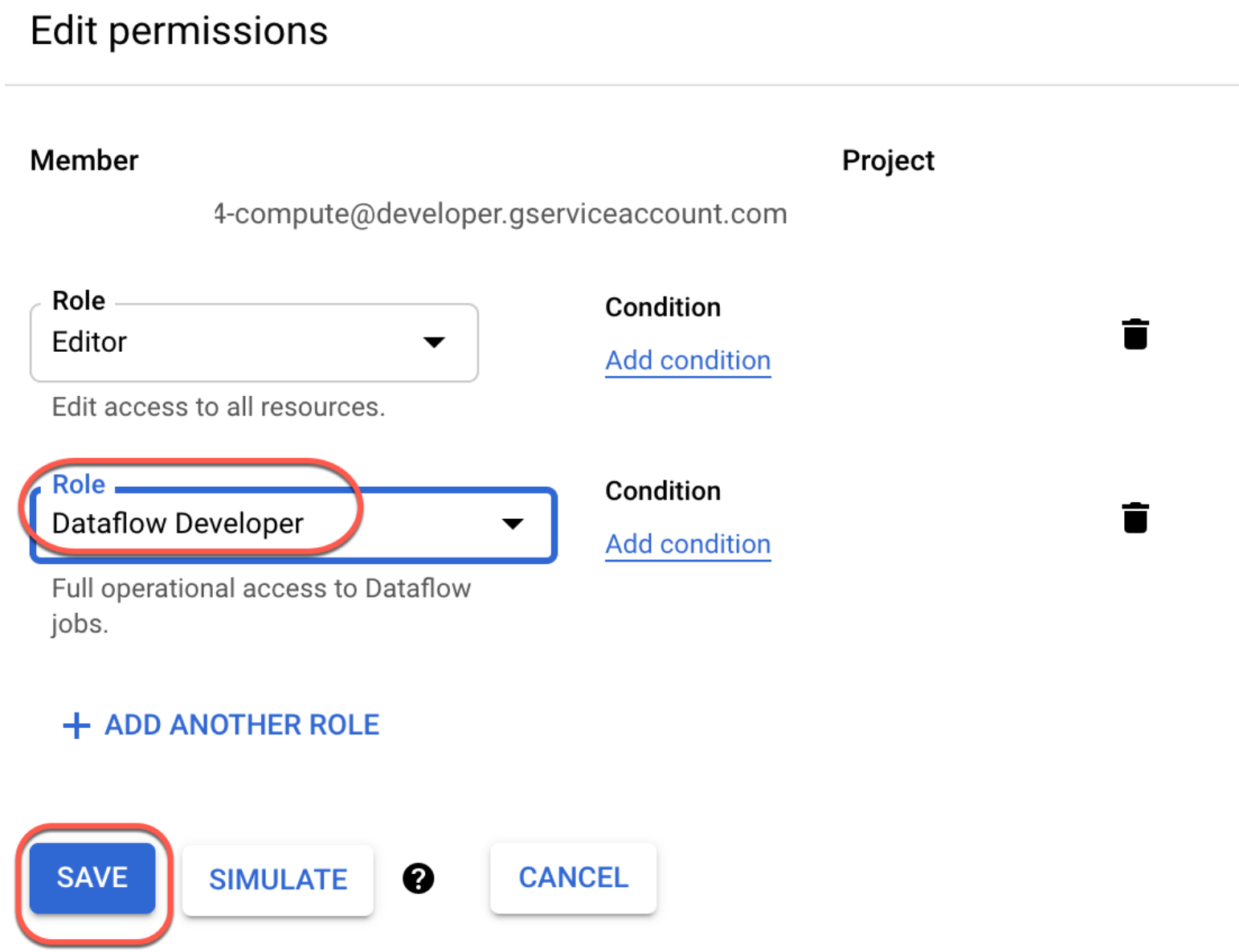
Lab : Practicing Pipeline Side Inputs

Preparation(add permissions, enable dataflow API)

Assign the Dataflow Developer Role

If the account does not have the Dataflow Developer role, follow the steps below to assign the required role.

- On the **Navigation menu**, click **IAM & Admin** > **IAM**.
- Select the default compute Service Account {project-number}-compute@developer.gserviceaccount.com.
- Select the **Edit** option (the pencil on the far right).
- Click **Add Another Role**
- Click inside the box for **Select a Role**. In the **Type to filter** selector, type and choose **Dataflow Developer**.
- Click **Save**.



Preparations, identify map and reduce

- ✓ Open the SSH terminal and connect to the training VM
  Compute Engine > **VM instances** > **training-vm** > **Connect**
- ✓ In the **training-vm** SSH terminal (Clone the training github repository)
  git clone https://github.com/GoogleCloudPlatform/training-data-analyst
- ✓ **Create a Cloud Storage bucket**
  Cloud Storage > **Browser** > **Create Bucket**
  Name :<your unique bucket name (Project ID)>
  Location type：Multi-Region
  Location ： <Your location>
- ✓ In **training-vm** SSH terminal init bucket , project variable

```
BUCKET="<your unique bucket name (Project ID)>"
 echo $BUCKET
PROJECT="<your unique project name (Project ID)>"
echo $PROJECT
```

Bigquery , Explore the pipeline code

- ✓ First query

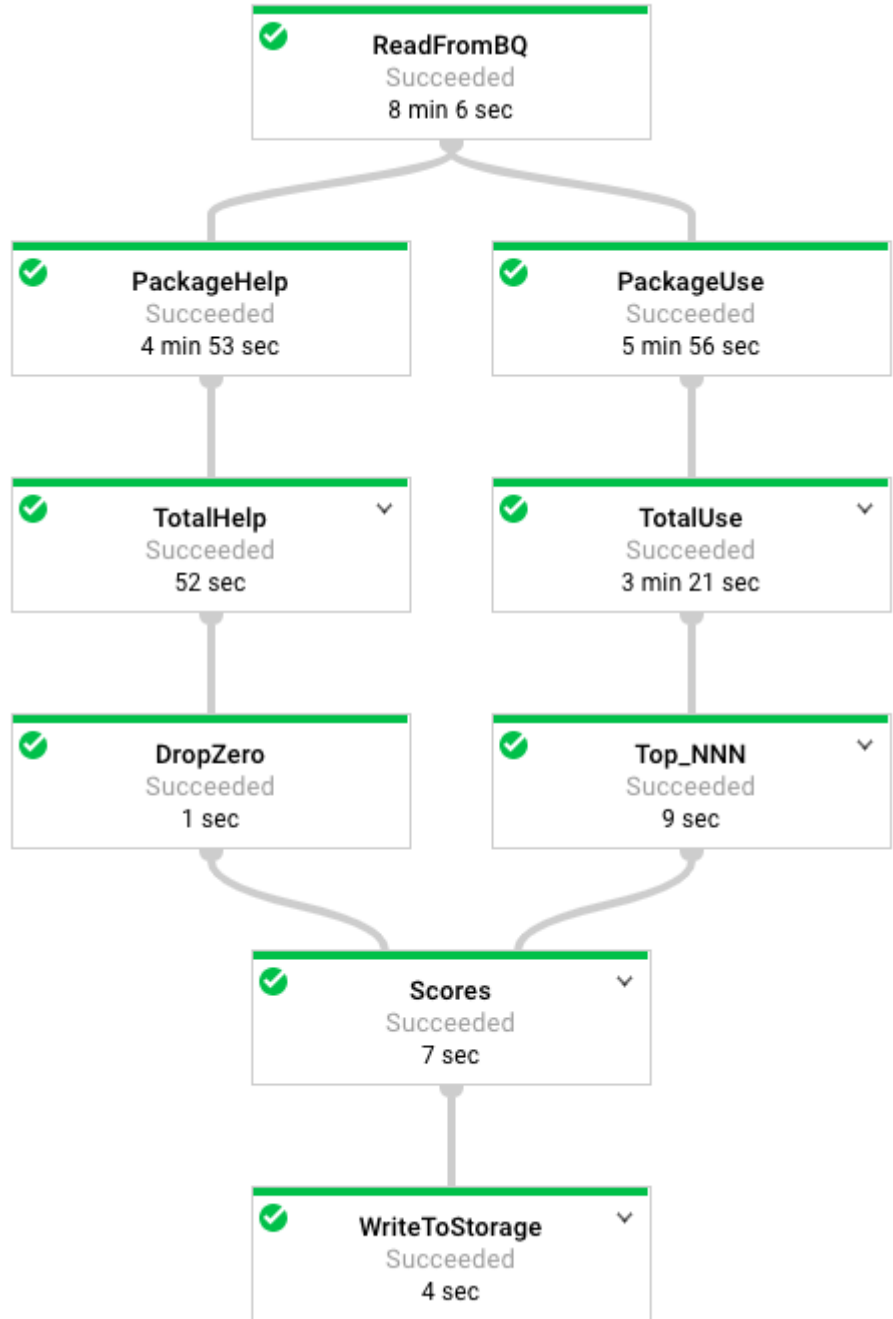
```
SELECT
  content
FROM
  `fh-bigquery.github_extracts.contents_java_2016`
LIMIT
  10
```

- ✓ Second query

```
SELECT
  COUNT(*)
FROM
  `fh-bigquery.github_extracts.contents_java_2016`
```

- ✓ In VM Terminal

```
cd ~/training-data-analyst/courses/data_analysis/lab2/python
nano JavaProjectsThatNeedHelp.py
Ctrl+X
```


ReadFromBQ
Succeeded
8 min 6 sec
PackageHelp
Succeeded
4 min 53 sec
PackageUse
Succeeded
5 min 56 sec
TotalHelp
Succeeded
52 sec
TotalUse
Succeeded
3 min 21 sec
DropZero
Succeeded
1 sec
Top_NNN
Succeeded
9 sec
Scores
Succeeded
7 sec
WriteToStorage
Succeeded
4 sec

Execute the pipeline

1.The program requires BUCKET and PROJECT values and choosing whether to run the pipeline locally using --DirectRunner or on the cloud using --DataFlowRunner

2.Execute the pipeline locally by typing the following into the **training-vm** SSH terminal.

python3 JavaProjectsThatNeedHelp.py --bucket $BUCKET --project $PROJECT –DirectRunner

3. Cloud Storage > Browser > javahelp folder > **Result**

4.Execute the pipeline on the cloud by typing the following into the **training-vm** SSH terminal.

python3 JavaProjectsThatNeedHelp.py --bucket $BUCKET --project $PROJECT –DataFlowRunner

5. Monitor job in Dataflow

6. Cloud Storage > Browser > javahelp folder > **Result**

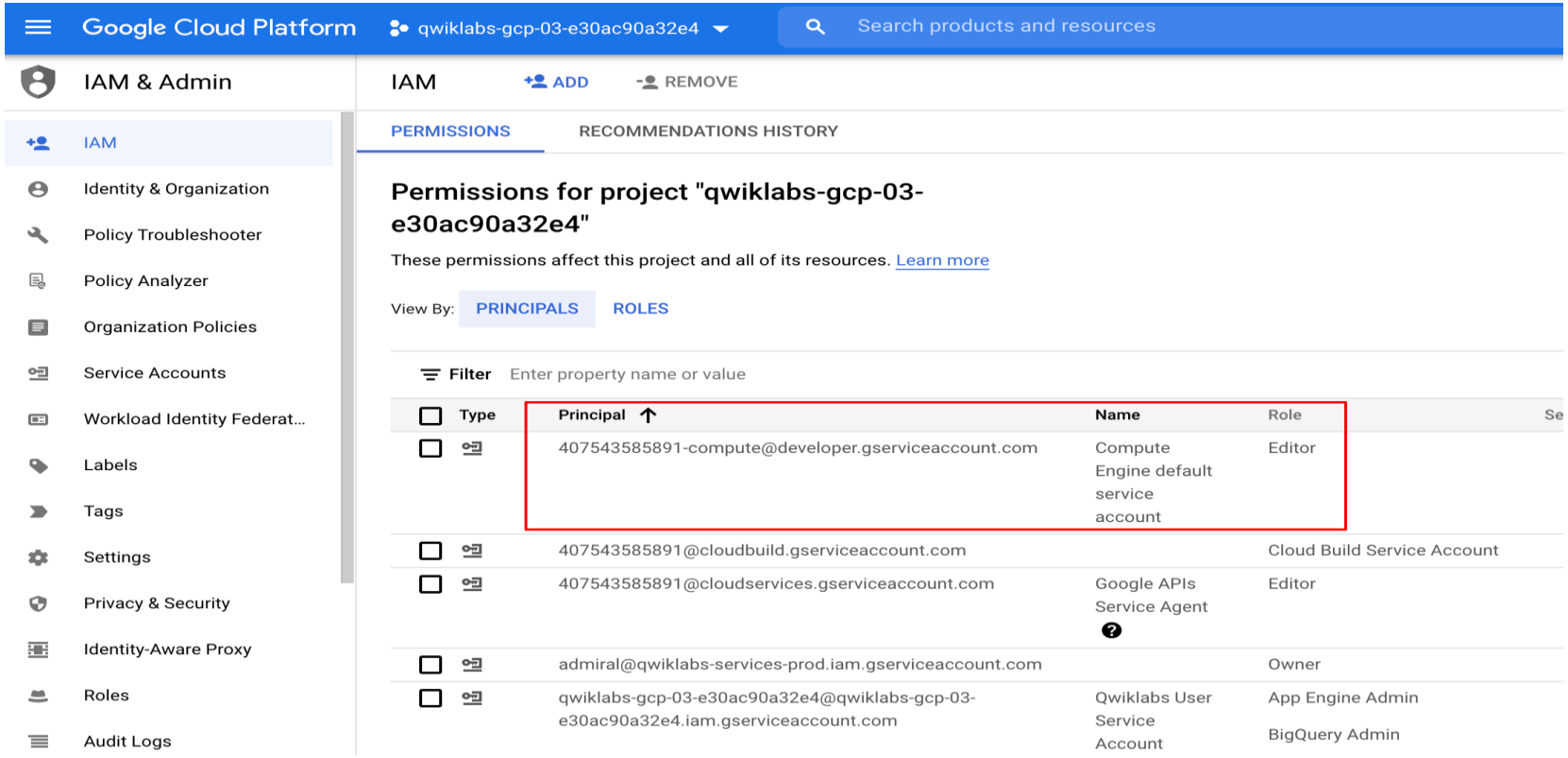
appendix

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** ( ), 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

Replace {project-number} with your project number.

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

Task 1. 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.
2. Click on the result for **Dataflow API**.
3. Click **Manage**.
4. Click **Disable API**.
5. If asked to confirm, click **Disable**.
6. Click **Enable**.