Streaming Data Processing: Streaming Data Pipelines into Bigtable | Qwiklabs

Monday, December 7, 2020   6:13 PM

Clipped from: https://googlecourses.qwiklabs.com/course_sessions/77404/labs/25811

Overview

In this lab you will use Dataflow to collect traffic events from simulated traffic sensor data made available through Google Cloud PubSub, and write them into a Bigtable table.

Objectives

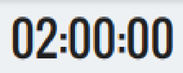
In this lab, you will perform the following tasks:

- Launch Dataflow pipeline to read from Pub/Sub and write into Bigtable
- Open an HBase shell to query the Bigtable database

Setup

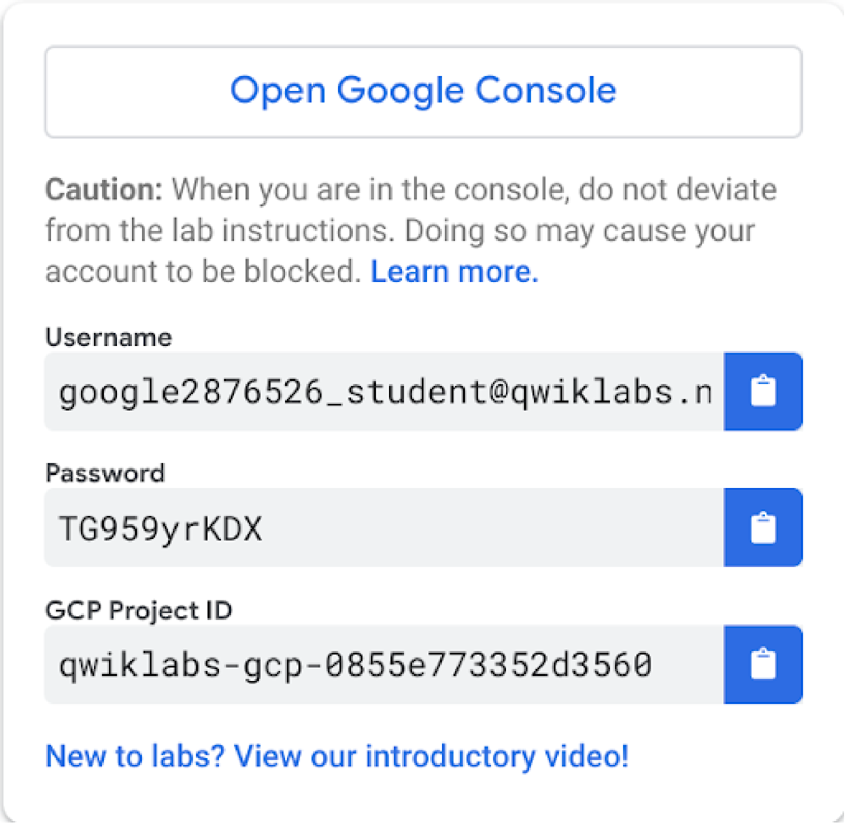
For each lab, you get a new GCP project and set of resources for a fixed time at no cost.

1. Make sure you signed into Qwiklabs using an **incognito window**.
2. Note the lab's access time (for example,

   and make sure you can finish in that time block.
3. When ready, click

   .
4. Note your lab credentials. You will use them to sign in to Cloud Platform 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

   google2876526_student@qwiklabs.n

   Password

   TG959yrKDX

   GCP Project ID

   qwiklabs-gcp-0855e773352d3560

   New to labs? View our introductory video!
5. Click **Open Google Console**.

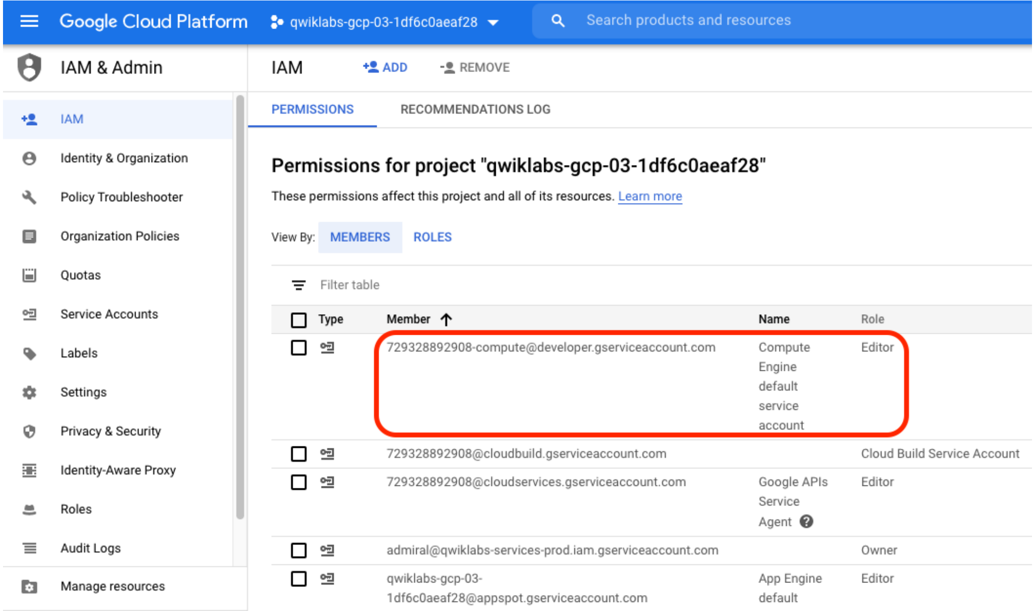
6. Click **Use another account** and copy/paste credentials for **this** lab into the prompts.

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

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**.



Permissions for project "qwiklabs-gcp-03-1df6c0aeaf28"

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

□ Type	Member ↑	Name	Role
□	729328892908-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor
□	729328892908@cloudbuild.gserviceaccount.com		Cloud Build Service Account
□	729328892908@cloudservices.gserviceaccount.com	Google APIs Service Agent	Editor
□	admiral@qwiklabs-services-prod.iam.gserviceaccount.com		Owner
□	qwiklabs-gcp-03-1df6c0aeaf28@appspot.gserviceaccount.com	App Engine default	Editor

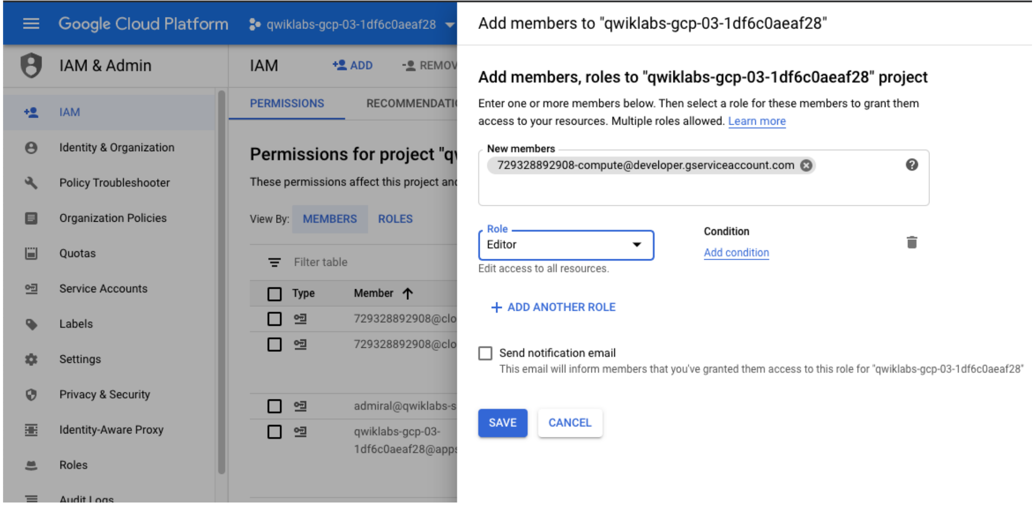
If the account is not present in IAM or does not have the `editor` role, follow the steps below to assign the required role.

- 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 members**, type:

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

Replace `{project-number}` with your project number.

- For **Role**, select **Project** > **Editor**. Click **Save**.



Add members to "qwiklabs-gcp-03-1df6c0aeaf28"

Add members, roles to "qwiklabs-gcp-03-1df6c0aeaf28" project

Enter one or more members below. Then select a role for these members to grant them access to your resources. Multiple roles allowed. Learn more

New members: 729328892908-compute@developer.gserviceaccount.com

Role: Editor

Edit access to all resources.

Condition: Add condition

+ ADD ANOTHER ROLE

□ Send notification email

This email will inform members that you've granted them access to this role for "qwiklabs-gcp-03-1df6c0aeaf28"

SAVE CANCEL

Audit Logs

Manage resources

qwiklabs-gcp-03-1df
1df6c0aeaf28.iam.g

Task 1: Preparation

You will be running a sensor simulator from the training VM. There are several files and some setup of the environment required.

Open the SSH terminal and connect to the training VM

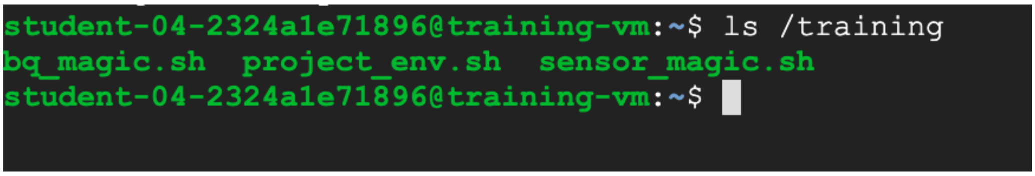
1. In the Console, on the **Navigation menu** ( ), click **Compute Engine** > **VM instances**.
2. Locate the line with the instance called **training-vm**.
3. On the far right, under **Connect** column, Click on **SSH** to open a terminal window.
4. In this lab you will enter CLI commands on the **training-vm**.

Verify initialization is complete

1. The **training-vm** is installing some software in the background. Verify that setup is complete by checking the contents of the new directory.

```
ls /training
```

The setup is complete when the result of your list (ls) command output appears as in the image below. If the full listing does not appear, wait a few minutes and try again. **Note**: It may take 2 to 3 minutes for all background actions to complete.



```
student-04-2324a1e71896@training-vm:~$ ls /training
bq_magic.sh  project_env.sh  sensor_magic.sh
student-04-2324a1e71896@training-vm:~$ 
```

Download Code Repository

1. Next you will download a code repository for use in this lab.

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

Set environment variables

1. On the **training-vm** SSH terminal enter the following:

```
source /training/project_env.sh
```

This script sets the __$$DEVSHELL\_PROJECT\_ID__ and __$$BUCKET__ environment variables.

Prepare HBase quickstart files

1. In the **training-vm** SSH terminal run the script to download and unzip the quickstart files (you will later use these to run the HBase shell.)

```
cd ~/training-data-analyst/courses/streaming/process/sandiego
./install_quickstart.sh
```

Click Check my progress to verify the objective. Copy sample files to the training_vm home directory

Task 2: Simulate traffic sensor data into Pub/Sub

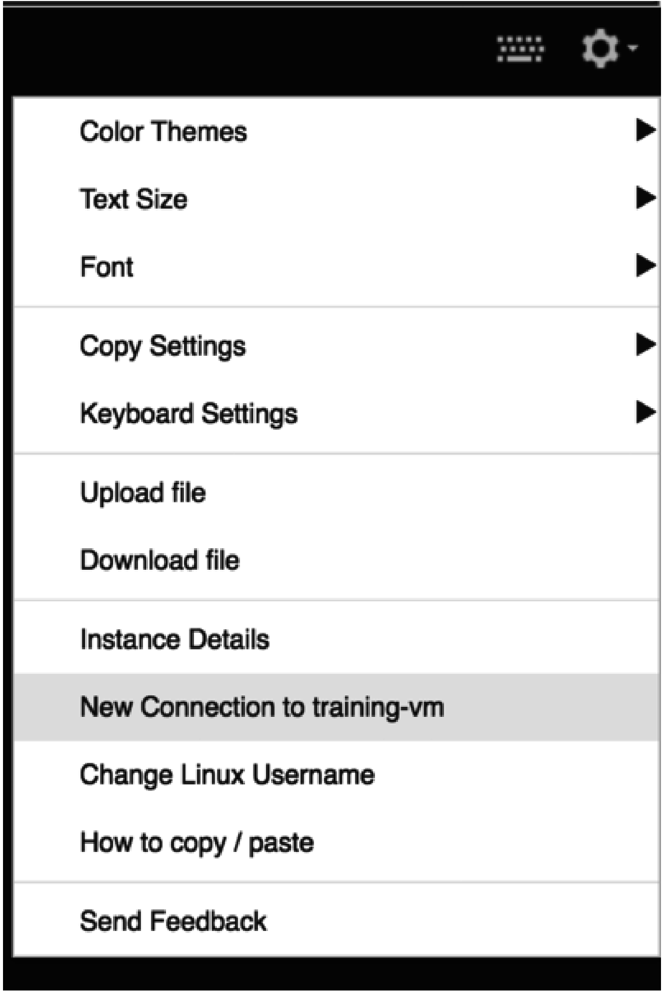
1. In the **training-vm** SSH terminal, start the sensor simulator. The script reads sample data from a csv file and publishes it to Pub/Sub.

```
/training/sensor_magic.sh
```

This command will send 1 hour of data in 1 minute. Let the script continue to run in the current terminal.

Open a second SSH terminal and connect to the training VM

1. In the upper right corner of the **training-vm** SSH terminal, click on the gear-shaped button (⚙)and select **New Connection to training-vm** from the drop-down menu. A new terminal window will open.



1. The new terminal session will not have the required environment variables. Run the following command to set them.
2. In the new **training-vm** SSH terminal enter the following:

```
source /training/project_env.sh
```

Click Check my progress to verify the objective. Simulate traffic sensor data into Pub/Sub

Task 3: Launch Dataflow Pipeline

1. In the second **training-vm** SSH terminal, navigate to the directory for this lab. Examine the script in Cloud Shell or using nano. **Do not make any changes to the code.**

```
cd ~/training-data-analyst/courses/streaming/process/sandiego

nano run_oncloud.sh
```

What does the script do?

1. The script takes 3 required arguments: project id, bucket name, classname and possibly a 4th argument: options. In this part of the lab, we will use the `--bigtable` option which will direct the pipeline to write into Cloud Bigtable.
2. Run the following script to create the Bigtable instance.

```
cd ~/training-data-analyst/courses/streaming/process/sandiego

./create_cbt.sh
```

1. Run the Dataflow pipeline to read from PubSub and write into Cloud Bigtable.

```
cd ~/training-data-analyst/courses/streaming/process/sandiego

./run_oncloud.sh $DEVSHELL_PROJECT_ID $BUCKET CurrentConditions --
bigtable
```

Example successful run:

```
[INFO] ---------------------------------------------------------------
---------
[INFO] BUILD SUCCESS
[INFO] ---------------------------------------------------------------
---------
[INFO] Total time: 47.582 s
[INFO] Finished at: 2018-06-08T21:25:32+00:00
[INFO] Final Memory: 58M/213M
[INFO] ---------------------------------------------------------------
---------
```

Click Check my progress to verify the objective. Launch Dataflow Pipeline

Task 4: Explore the pipeline

1. Return to the browser tab for Console. On the **Navigation menu** (≡), click **Dataflow** and click on the new pipeline job. Confirm that the pipeline job is listed and verify that it is running without errors.
2. Find the **write:cbt** step in the pipeline graph, and click on the down arrow on the right to see the writer in action. Click on given writer. Review the **Bigtable Options** in the **Step summary**.

Task 5: Query Bigtable data

1. In the second **training-vm** SSH terminal, run the **quickstart.sh** script to launch the HBase shell.

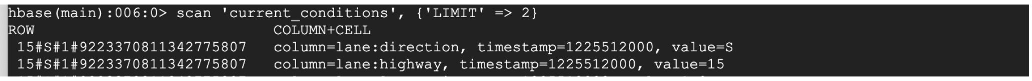
```
cd ~/training-data-
analyst/courses/streaming/process/sandiego/quickstart

./quickstart.sh
```

1. When the script completes, you will be in an HBase shell prompt that looks like this:

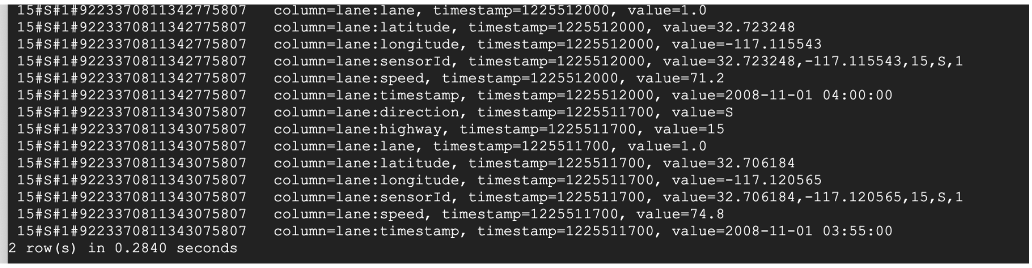
```
hbase(main):001:0>
```

1. At the HBase shell prompt, type the following query to retrieve 2 rows from your Bigtable table that was populated by the pipeline. It may take a few minutes for results to return via the HBase query. Please repeat the 'scan' command until you see a list of rows returned.

```
scan 'current_conditions', {'LIMIT' => 2}
```



```
hbase(main):006:0> scan 'current_conditions', {'LIMIT' => 2}
ROW                            COLUMN+CELL
 15#S#1#9223370811342775807    column=lane:direction, timestamp=1225512000, value=S
 15#S#1#9223370811342775807    column=lane:highway, timestamp=1225512000, value=15
```



```
 15#S#1#9223370811342775807   column=lane:lane, timestamp=1225512000, value=1.0
 15#S#1#9223370811342775807   column=lane:latitude, timestamp=1225512000, value=32.723248
 15#S#1#9223370811342775807   column=lane:longitude, timestamp=1225512000, value=-117.115543
 15#S#1#9223370811342775807   column=lane:sensorId, timestamp=1225512000, value=32.723248,-117.115543,15,S,1
 15#S#1#9223370811342775807   column=lane:speed, timestamp=1225512000, value=71.2
 15#S#1#9223370811342775807   column=lane:timestamp, timestamp=1225512000, value=2008-11-01 04:00:00
 15#S#1#9223370811343075807   column=lane:direction, timestamp=1225511700, value=S
 15#S#1#9223370811343075807   column=lane:highway, timestamp=1225511700, value=15
 15#S#1#9223370811343075807   column=lane:lane, timestamp=1225511700, value=1.0
 15#S#1#9223370811343075807   column=lane:latitude, timestamp=1225511700, value=32.706184
 15#S#1#9223370811343075807   column=lane:longitude, timestamp=1225511700, value=-117.120565
 15#S#1#9223370811343075807   column=lane:sensorId, timestamp=1225511700, value=32.706184,-117.120565,15,S,1
 15#S#1#9223370811343075807   column=lane:speed, timestamp=1225511700, value=74.8
 15#S#1#9223370811343075807   column=lane:timestamp, timestamp=1225511700, value=2008-11-01 03:55:00
2 row(s) in 0.2840 seconds
```

1. Review the output. Notice each row is broken into column, timestamp, value combinations.
2. Run another query. This time look only at the **lane: speed** column, limit to 10 rows, and specify **rowid patterns** for start and end rows to scan over.

```
scan 'current_conditions', {'LIMIT' => 10, STARTROW => '15#S#1',
ENDROW => '15#S#999', COLUMN => 'lane:speed'}
```

1. Review the output. Notice that you see 10 of the column, timestamp, value combinations, all of which correspond to Highway 15. Also notice that column is restricted to **lane: speed**.
2. Feel free to run other queries if you are familiar with the syntax. Once you're satisfied, `quit` to exit the shell.

```
quit
```

Cleanup

1. Run the script to delete your Bigtable instance. If prompted, press `Enter`.

```
cd ~/training-data-analyst/courses/streaming/process/sandiego
./delete_cbt.sh
```

1. On your Dataflow page in your Cloud Console, click on the pipeline job name and click the `Stop job` on the right panel.
2. Go back to the first Cloud Shell tab with the publisher and type `Ctrl+C` to stop it.
3. Go to the BigQuery console and delete the dataset demos.

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.