Streaming Data Processing: Streaming Data Pipelines into Bigtable

## 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.

At the time of this writing, streaming pipelines are not available in the DataFlow Python SDK. So the streaming labs are written in Java.

## Objectives

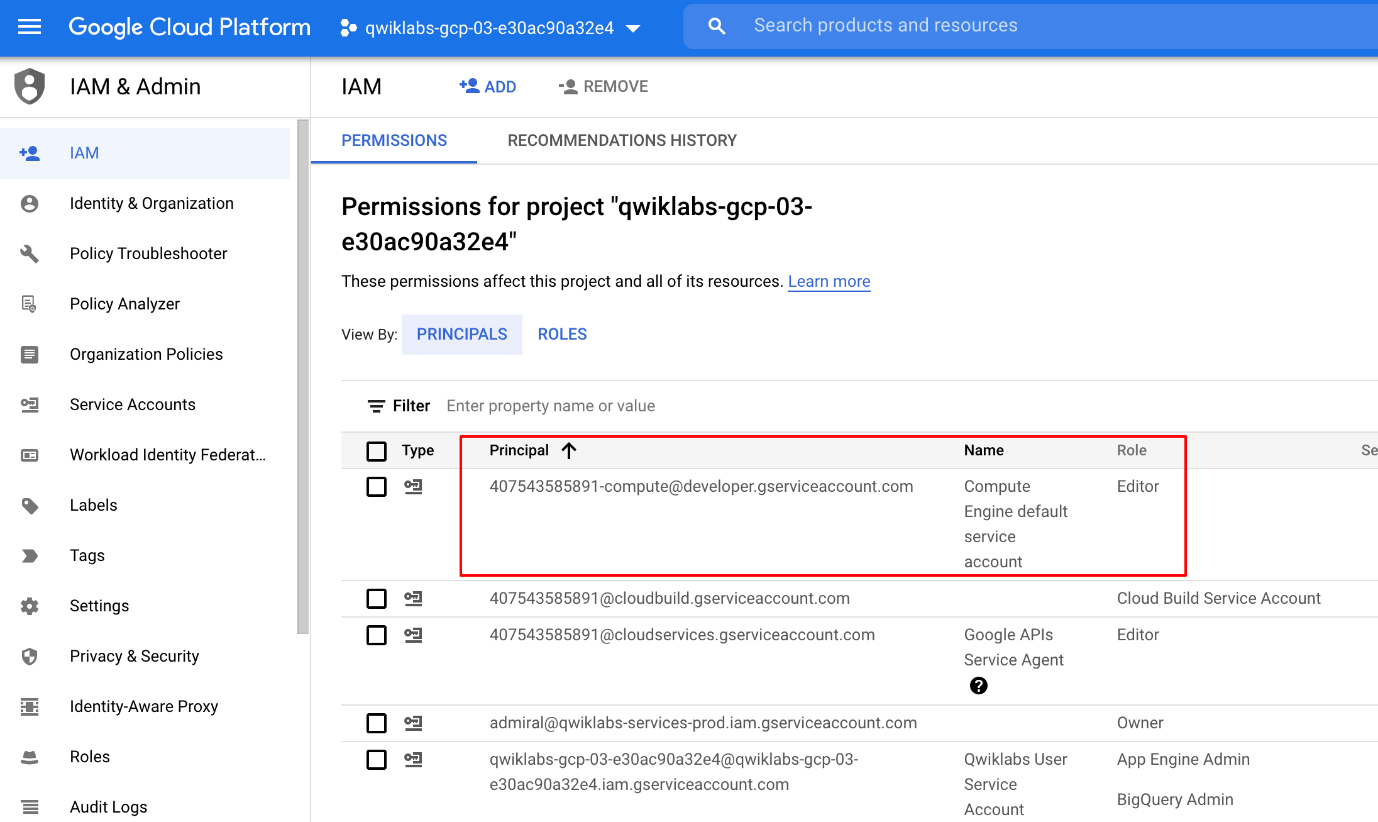
In this lab, you will perform the following tasks:

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

### **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: 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. Then click **Connect**.

In this lab, you will enter CLI commands on the **training-vm**.

### **Verify initialization is complete**

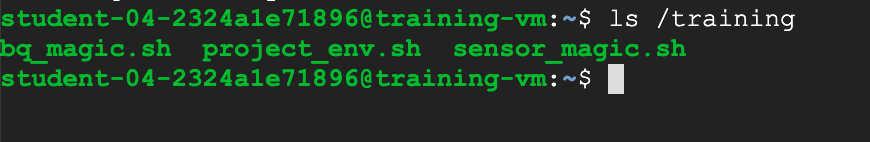
1. The **training-vm** is installing some software in the background. Verify that setup is complete by checking the contents of the new directory.

ls /training

Copied!

content\_copy

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.



### **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

Copied!

content\_copy

### **Set environment variables**

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

source /training/project\_env.sh

Copied!

content\_copy

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

Copied!

content\_copy

Click Check my progress to verify the objective.

Copy sample files to the training\_vm home directory

Check my progress

## 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

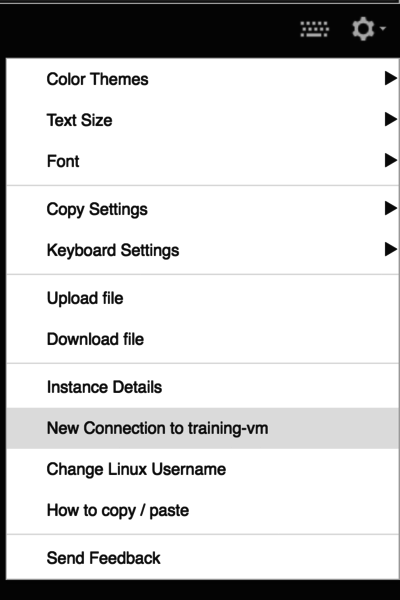
Copied!

content\_copy

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.



The new terminal session will not have the required environment variables. Complete the next step to set these variables.

1. In the new **training-vm** SSH terminal, enter the following:

source /training/project\_env.sh

Copied!

content\_copy

Click Check my progress to verify the objective.

Simulate traffic sensor data into Pub/Sub

Check my progress

## 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

Copied!

content\_copy

What does the script do?

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.

1. Press **CTRL+X** to exit.
2. Run the following script to create the Bigtable instance.

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

./create\_cbt.sh

Copied!

content\_copy

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

Copied!

content\_copy

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

Check my progress

## 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 the 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

Copied!

content\_copy

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

hbase(main):001:0>

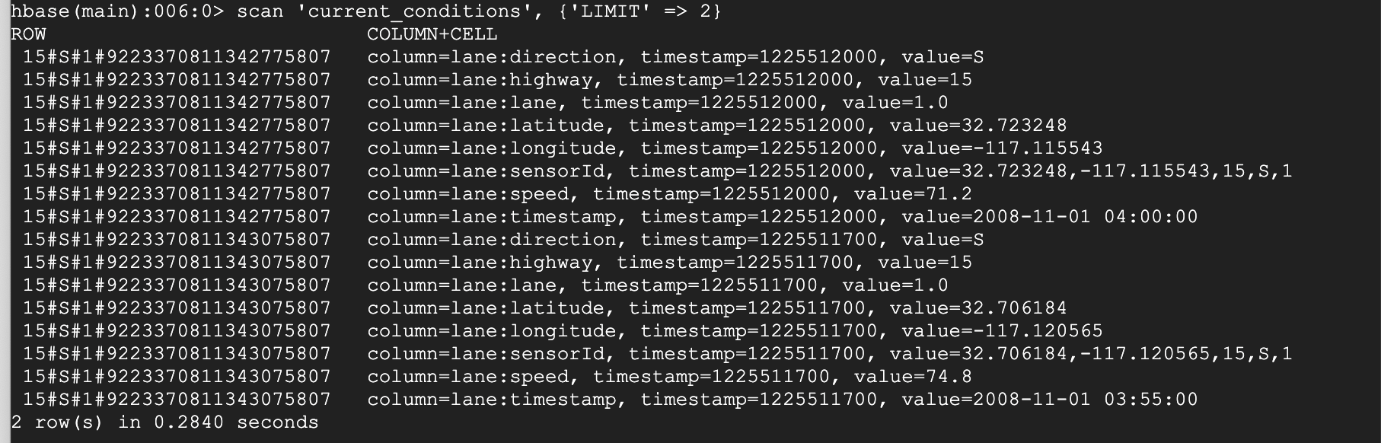
1. At the HBase shell prompt, type the following query to retrieve 2 rows from your Bigtable table that was populated by the pipeline. It may take a few minutes for results to return via the HBase query.

Repeat the 'scan' command until you see a list of rows returned.

scan 'current\_conditions', {'LIMIT' => 2}

Copied!

content\_copy



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'}

Copied!

content\_copy

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, enter quit to exit the shell.

quit

Copied!

content\_copy

## Cleanup

1. In the second **training-vm** SSH terminal, run the following script to delete your Bigtable instance.

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

./delete\_cbt.sh

Copied!

content\_copy

If prompted to confirm, enter Y.

1. On your Dataflow page in your Cloud Console, click on the pipeline job name.
2. Click **Stop** on the top menu bar. Select **Cancel**, and then click **Stop Job**.
3. Go back to the first SSH terminal with the publisher, and enter Ctrl+C to stop it.
4. In the BigQuery console, click on the three dots next to the **demos** dataset, and click **Delete**.
5. Type **delete** and then click **Delete**.

## End your lab