Streaming Data Processing - Lab 4: Streaming Data Pipelines into Bigtable v1.3

## 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

## 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** ( 7a91d354499ac9f1.png), 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 software in the background. Verify that setup is complete by checking that the following directory exists. If it does not exist, wait a few minutes and try again.

ls /training

Wait until setup is complete before proceeding. You can verify the installation of maven with **mvn -version** and the JDK with **java -version**.

### **Copy files**

1. A repository has been downloaded to the VM. Copy the repository to your home directory.

cp -r /training/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

## Task 2: Simulate traffic sensor data into Pub/Sub

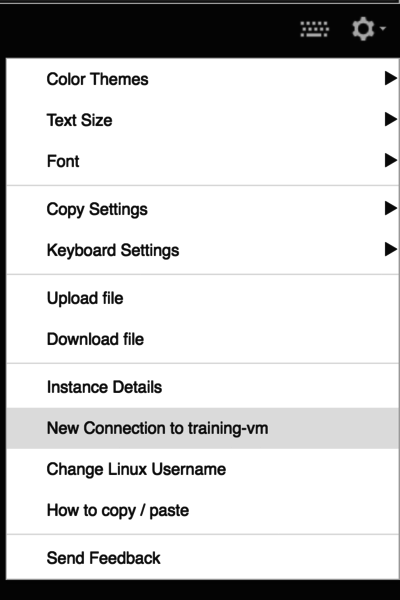
1. In the **training-vm** SSH terminal, start the sensor simulator. The script reads sample data from a csv file and publishes it to Pub/Sub.

/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 ( 9649d58acf1c4e06.png) 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

## 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] ------------------------------------------------------------------------

## Task 4: Explore the pipeline

1. Return to the browser tab for Console. On the **Navigation menu** ( 7a91d354499ac9f1.png), 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. If the script runs successfully, you would be in an HBase shell prompt that looks something 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.

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

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.