# Lab : Streaming Data Pipelines into Bigtable

# Preparation

- ✓ **Open the SSH terminal and connect to the training VM**
  **Compute Engine** > **VM instances**> **training-vm** > **Connect**.
- ✓ in VM terminal write code ls /training
- ✓ **Download Code Repository**
  git clone https://github.com/GoogleCloudPlatform/training-data-analyst
- ✓ in VM terminal **Set environment variable**
  **source /training/project_env.sh**
- ✓ **Prepare HBase quickstart files**

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

# Simulate traffic sensor data into Pub/Sub

- ✓ in VM terminal run sensor magic code
  **/training/sensor_magic.sh**
- ✓ upper right corner of the **training-vm** SSH terminal **New Connection to training-vm**
- ✓ **Set environment variables in second VM terminal**
  **source /training/project_env.sh**

# Launch Dataflow Pipeline

- ✓ In the second **training-vm**
  cd ~/training-data-analyst/courses/streaming/process/sandiego
  nano run_oncloud.sh
  **CTRL+X** # interrupt
- ✓ Run the following script to **create the Bigtable instance**
  **cd ~/training-data-analyst/courses/streaming/process/sandiego**
  **./create_cbt.sh**
- ✓ 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**

# Explore the pipeline , Query Bigtable data

- ✓ **Navigation** >> **Dataflow >> graph >> write:cbt** step, Review the **Bigtable Options** in **Step summary**
- ✓ second **training-vm**

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

  ./quickstart.sh
- ✓ script completes you are in **HBase shell prompt**

  scan 'current_conditions', {'LIMIT' => 2} # query to retrieve 2 rows from your Bigtable

  **"**each row is broken into column, timestamp, value combinations**"**
- ✓ 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'}
- ✓ **Exit shell**

  **quit**

# Cleanup (release sources )

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