



Department of Computing

SE-315: Cloud Computing

Lab 11: Serverless App Dev – Creating a Streaming Data Pipeline for a Real-Time Dashboard with Dataflow – Part A

CLO4: Display skills to effectively use cloud centric solutions such as serverless application development.

Date: 4.12.24



Lab 11: Serverless App Dev – Creating a Streaming Data Pipeline for a Real-Time Dashboard with Dataflow – Part A

Introduction:

Assume that you own a fleet of New York City taxi cabs and are looking to monitor how well your business is doing in real time. In this lab, you build a streaming data pipeline to capture taxi revenue, passenger count, ride status, and much more, and then visualize the results in a management dashboard.

Lab Objectives: In this lab, the students will learn how to:

- Create a Dataflow job from a template
- Stream a Dataflow pipeline into BigQuery
- Monitor a Dataflow pipeline in BigQuery
- Analyze results with SQL
- Visualize key metrics in Looker Studio

Lab Tasks

Go through the following link:

<https://www.cloudskillsboost.google/focuses/19077?parent=catalog>

which will take you to the 'Creating a Streaming Data Pipeline for a Real-Time Dashboard with Dataflow' page. The list of tasks is given below. Make sure to take screenshots of each task as you will need to add them in the solution section given below.



Setting up qwilabs account:

```
(qwilabs-gcp-04-8540b18fd6be) x + Editor
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to qwilabs-gcp-04-8540b18fd6be.
Use "gcloud config set project [PROJECT_ID]" to change to a different project.
student_00_30944efe8ad1@cloudshell:~ (qwilabs-gcp-04-8540b18fd6be) $ gcloud auth list
Credentialed Accounts

ACTIVE: *
ACCOUNT: student-00-30944efe8ad1@qwilabs.net

To set the active account, run:
$ gcloud config set account `ACCOUNT`

student_00_30944efe8ad1@cloudshell:~ (qwilabs-gcp-04-8540b18fd6be) $
```

Task 1. Create a BigQuery dataset

create the **taxirides** dataset:

```
$ gcloud config set account `ACCOUNT`

student_00_30944efe8ad1@cloudshell:~ (qwilabs-gcp-04-8540b18fd6be) $ bq --location=us-east4 mk taxirides
Dataset 'qwilabs-gcp-04-8540b18fd6be:taxirides' successfully created.
```

create the **taxirides.realtime** table

```
student_00_30944efe8ad1@cloudshell:~ (qwilabs-gcp-04-8540b18fd6be) $ bq --location=us-east4 mk \
--time_partitioning_field timestamp \
--schema ride_id:string,point_idx:integer,latitude:float,longitude:float,\
timestamp:timestamp,meter_reading:float,meter_increment:float,ride_status:string,\
passenger_count:integer -t taxirides.realtime
Table 'qwilabs-gcp-04-8540b18fd6be:taxirides.realtime' successfully created.
```



Task 2. Copy required lab artifacts

run the following commands to move files needed for the Dataflow job.

here, we copy the files from the google storage bucket.

```
student_00_30944efe8ad1@cloudshell:~ (qwiklabs-gcp-04-8540b18fd6be)$ gcloud storage cp gs://cloud-training/bdml/taxisrdata/schema.json gs://qwiklabs-gcp-04-8540b18fd6be-bucket/tmp/schema.json
gcloud storage cp gs://cloud-training/bdml/taxisrdata/transform.js gs://qwiklabs-gcp-04-8540b18fd6be-bucket/tmp/transform.js
gcloud storage cp gs://cloud-training/bdml/taxisrdata/rt_taxidata.csv gs://qwiklabs-gcp-04-8540b18fd6be-bucket/tmp/rt_taxidata.csv
Copying gs://cloud-training/bdml/taxisrdata/schema.json to gs://qwiklabs-gcp-04-8540b18fd6be-bucket/tmp/schema.json
Completed files 1/1 | 610.0B/610.0B
Copying gs://cloud-training/bdml/taxisrdata/transform.js to gs://qwiklabs-gcp-04-8540b18fd6be-bucket/tmp/transform.js
Completed files 1/1 | 435.0B/435.0B
Copying gs://cloud-training/bdml/taxisrdata/rt_taxidata.csv to gs://qwiklabs-gcp-04-8540b18fd6be-bucket/tmp/rt_taxidata.csv
Completed files 1/1 | 108.3kiB/108.3kiB
```

Task 3. Set up a Dataflow Pipeline

in this task, we set up a streaming data pipeline to read files from your Cloud Storage bucket and write data to BigQuery. ([Dataflow](#) is a serverless way to carry out data analysis.)

Restart the connection to the Dataflow API.

```
student_00_30944efe8ad1@cloudshell:~ (qwiklabs-gcp-04-8540b18fd6be)$ gcloud services disable dataflow.googleapis.com
gcloud services enable dataflow.googleapis.com
Operation "operations/acat.p17-71696123621-4c259990-eb36-427d-8f07-538c4118fcb2" finished successfully.
Operation "operations/acf.p2-71696123621-f0bb7b27-dfb6-4970-8384-7842a4d363be" finished successfully.
```



Create a new streaming pipeline:

setting up job name, endpoint and dataflow template

Job name *

streaming-taxi-pipeline

Must be unique among running jobs

Some locations have been restricted due to a policy set by your organization.

Regional endpoint *

us-east4 (Northern Virginia) ▼ ?

Choose a Dataflow regional endpoint to deploy worker instances and store job metadata. You can optionally deploy worker instances to any available Google Cloud region or zone by using the worker region or worker zone parameters. Job metadata is always stored in the Dataflow regional endpoint. [Learn more](#)

Dataflow template *

Cloud Storage Text to BigQuery (Stream) ▼ ?

setting up required parameters

Source

Cloud Storage Input File(s) *

☒ gs:// qwiklabs-gcp-04-8540b18fd6be-bucket/tmp/rt_taxidata.csv BROWSE

Path of the file pattern glob to read from. (Example: gs://your-bucket/path/*.csv)

Target

Cloud Storage location of your BigQuery schema file, described as a JSON *

☒ gs:// qwiklabs-gcp-04-8540b18fd6be-bucket/tmp/schema.json BROWSE

JSON file with BigQuery Schema description. JSON Example: { "BigQuery Schema": [{ "name": "location", "type": "STRING" }, { "name": "name", "type": "STRING" }, { "name": "age", "type": "STRING" }, { "name": "color", "type": "STRING" }, { "name": "coffee", "type": "STRING" }] }

BigQuery output table *

☒ qwiklabs-gcp-04-8540b18fd6be:taxirides.realtime BROWSE

Temporary directory for BigQuery loading process *

☒ gs:// qwiklabs-gcp-04-8540b18fd6be-bucket/tmp BROWSE

Temporary directory for BigQuery loading process (Example: gs://your-bucket/your-files/temp_dir)

Required Parameters



setting up optional parameters:

Optional Parameters ^

The dead-letter table name to output failed messages to BigQuery BROWSE

JavaScript UDF path in Cloud Storage ☒ gs:// qwiklabs-gcp-04-8540b18fd6be-bucket/tmp/ SELECT

CREATE UDF

The Cloud Storage URI of the .js file that defines the JavaScript user-defined function (UDF) to use. For example, 'gs://my-bucket/my-udfs/my_file.js'.

JavaScript UDF name

The name of the JavaScript user-defined function (UDF) to use. For example, if your JavaScript function code is `myTransform(inJson) { /*...do stuff...*/ }`, then the function name is `myTransform`. For sample JavaScript UDFs, see UDF Examples (<https://github.com/GoogleCloudPlatform/DataflowTemplates#udf-examples>).

JavaScript UDF auto-reload interval (minutes)

Define the interval that workers may check for JavaScript UDF changes to reload the files. Defaults to: 0.

Max workers

The maximum number of Google Compute Engine instances to be made available to your pipeline during execution, must be larger than 0

Number of workers

The initial number of Google Compute Engine instances to use, must be larger than 0



setting up machine type:

The machine type for Google Compute Engine instances used in your pipeline execution. e.g., n1-standard-1. [Learn more](#)

✓ General purpose

Compute optimized

Memory optimized

GPUs

Machine types for common workloads, optimized for cost and flexibility

Series

E2

CPU platform selection based on availability

Machine type

e2-medium (2 vCPU, 1 core, 4 GB memory)



vCPU

1-2 vCPU (1 shared core)

Memory

4 GB

Service account email

The email address of the service account to run the job as

Additional experiments

Additional experiment flags for the job, e.g., experiment1, experiment2

Worker IP Address Configuration

Unspecified

Dataflow workers can be configured to use public or internal IP addresses. [Learn more](#)



National University of Sciences and Technology (NUST)

School of Electrical Engineering and Computer Science

finally, our streaming-taxi-pipeline job is created:

Jobs

⋮ [ENABLE SORTING](#) [REFRESH](#) [MANAGE](#)

Dataflow jobs use Cloud Storage to store temporary files during pipeline execution. These files have soft delete policy enabled by default. To avoid being billed for unnecessary storage costs, turn off the soft delete feature on buckets that your Dataflow jobs use for temporary storage. [Learn more](#)

[DISMISS](#)

☐ Running ☐ Archived

[Filter](#) Filter jobs

Name	Type	End time	Elapsed time	Start time	Status	SDK version	ID
streaming-taxi-pipeline	Streaming		6 min 42 sec	Dec 4, 2024, 12:11:27 PM	Running	2.60.0	2024-12-03_23_11_26-1045844154058541611

Dataflow

Overview

Monitoring

Jobs

Pipelines

Workbench

Snapshots

Release Notes

streaming-taxi...

[CLONE](#) [STOP](#) [CREATE SNAPSHOT](#) [IMPORT AS PIPELINE](#) [SEND FEEDBACK](#)

JOB DETAILS

JOB METRICS

COST

RECOMMENDATIONS

AUTOSCALING

Job steps view

Graph view

Logs

[SHOW](#)

Job info

Job name

streaming-taxi-pipeline

Job ID

2024-12-03_23_11_26-1045844154058541611

Job type

Streaming

Job status

SDK version

Apache Beam SDK for Java 2.60.0

Job region

us-east4

Service zones

us-east4-a

Worker location

us-east4

Current workers

1

Latest worker status

Straggler status

No active straggler

Start time

December 4, 2024 at 12:11:27 PM GMT+5

Elapsed time

31 sec

Encryption type

Google-managed

Dataflow Prime

Disabled

Runner v2

Disabled

Streaming Engine

Disabled

Vertical Autoscaling

Disabled

Streaming Mode

Exactly once



Task 4. Analyze the taxi data using BigQuery

I ran the select query to display top 10 rows;

The screenshot shows the Google BigQuery interface. At the top, there's a query editor with the text: `1 SELECT * FROM taxirides.realtime LIMIT 10`. Below the editor, there are buttons for **RUN**, **DOWNLOAD**, **SHARE**, and **SAVE**. The **Query results** section is active, showing a table with 10 rows. The table has columns: **Row**, **ride_id**, **point_idx**, **latitude**, **longitude**, and **timestamp**. The data represents taxi trips from October 26, 2022.

Row	ride_id	point_idx	latitude	longitude	timestamp
1	d8105f1f-fe85-46f5-9696-81e1...	737	40.70515	-74.01003	2022-10-26 03:06
2	6bb43a77-d6e5-491d-a75b-82...	314	40.78509	-73.94906	2022-10-26 03:05
3	4bbe8346-1e42-4309-9af7-0f4...	992	40.7387	-73.93982	2022-10-26 03:13
4	9e989281-3cbd-4ad8-85c0-f1f...	352	40.72891	-73.97168	2022-10-26 03:10
5	c8d69976-95a9-49a5-a806-e44...	87	40.72481	-73.86962	2022-10-26 03:11
6	796aab23-4696-41a7-972e-c23...	1514	40.70376	-73.89562	2022-10-26 03:06
7	90f5e58f-1708-4335-a68a-71b...	591	40.71186	-73.9988	2022-10-26 03:12
8	a53ae320-aabf-4bc8-bd74-b1e...	227	40.71512	-73.97748	2022-10-26 03:07
9	a347ecef-5a21-4112-9152-0d6...	1241	40.73876	-73.8367	2022-10-26 03:12
10	fdefeab9-7e4a-4eed-b845-454...	29	40.77502	-73.98054	2022-10-26 03:09



Task 5. Perform aggregations on the stream for reporting

```
1 WITH streaming_data AS (  
2  
3 SELECT  
4     timestamp,  
5     TIMESTAMP_TRUNC(timestamp, HOUR, 'UTC') AS hour,  
6     TIMESTAMP_TRUNC(timestamp, MINUTE, 'UTC') AS minute,  
7     TIMESTAMP_TRUNC(timestamp, SECOND, 'UTC') AS second,  
8     ride_id,  
9     latitude,  
10    longitude,  
11    meter_reading,  
12    ride_status,  
13    passenger_count  
14 FROM
```

Query results [SAVE RESULTS](#) [EXPLORE DATA](#)

JOB INFORMATION **RESULTS** CHART JSON EXECUTION DETAILS EXEC

Row	dashboard_sort	minute	total_rides	total_revenue	total_passeng
1	1	2022-10-26 03:13:00 UTC	2	0.0	
2	2	2022-10-26 03:13:00 UTC	1	26.307207	
3	3	2022-10-26 03:13:00 UTC	2	23.3	
4	4	2022-10-26 03:13:00 UTC	1	8.120775	
5	5	2022-10-26 03:13:00 UTC	3	18.599999	

Results per page: 50 1 – 50 of 685

Saving the query:

Save query

Project
qwiklabs-gcp-04-8540b18fd6be

Name *
Saleha's Saved Query

Region *
us-east4 (Northern Virginia)

Info This will save your code asset in a new region. This setting will also set the default region where your code assets will be stored in the future. [Learn more](#)

SAVE **CANCEL**



Task 6. Stop the Dataflow Job

Stop job



Cancel

Dataflow will immediately stop this job and abort all data ingestion and processing. Any buffered data may be lost.



Drain

Dataflow will cease all data ingestion, but will attempt to finish processing any remaining buffered data. Pipeline resources will be maintained until buffered data has finished processing and any pending output has finished writing.



Force Cancel

Dataflow will force cancel this job. This option terminates a job that has become stuck in the cancelation process.

[Read more about stopping Dataflow jobs](#)

DO NOTHING

STOP JOB



Task 7. Create a real-time dashboard

running the saved query:

The screenshot shows the Google Cloud BigQuery interface. On the left is the Explorer panel with a search bar and a list of resources. The main panel displays a saved query named 'Saleha's Saved Query' with a 'RUN' button. Below the query editor, the 'Query results' section is active, showing a table with 5 rows of data. The table has columns for 'Row', 'dashboard_sort', 'minute', and 'total_rides'. The data shows timestamps for 2022-10-26 at 03:13:00 UTC.

Explorer

Search BigQuery resources

Viewing resources.

[SHOW STARRED ONLY](#)

- qwiklabs-gcp-04-8540b18fd6be
 - Queries
 - Shared queries
 - Saleha's Saved Query**
 - Notebooks
 - Data canvases
 - Data preparations
 - Workflows
 - External connections
 - taxirides

Saleha's Saved Query [RUN](#)

```
1 WITH streaming_data AS (  
2  
3 SELECT  
4     timestamp,  
5     TIMESTAMP_TRUNC(timestamp, HOUR, 'UTC') AS hour,  
6     TIMESTAMP_TRUNC(timestamp, MINUTE, 'UTC') AS minute,  
7     TIMESTAMP_TRUNC(timestamp, SECOND, 'UTC') AS second,  
8     ride_id,  
9     latitude,  
10    longitude,  
11    meter_reading,  
12    ride_status,  
13    passenger_count  
14 FROM
```

Press Alt+F1 for Accessibility Options.

Query results [SAVE RESULTS](#)

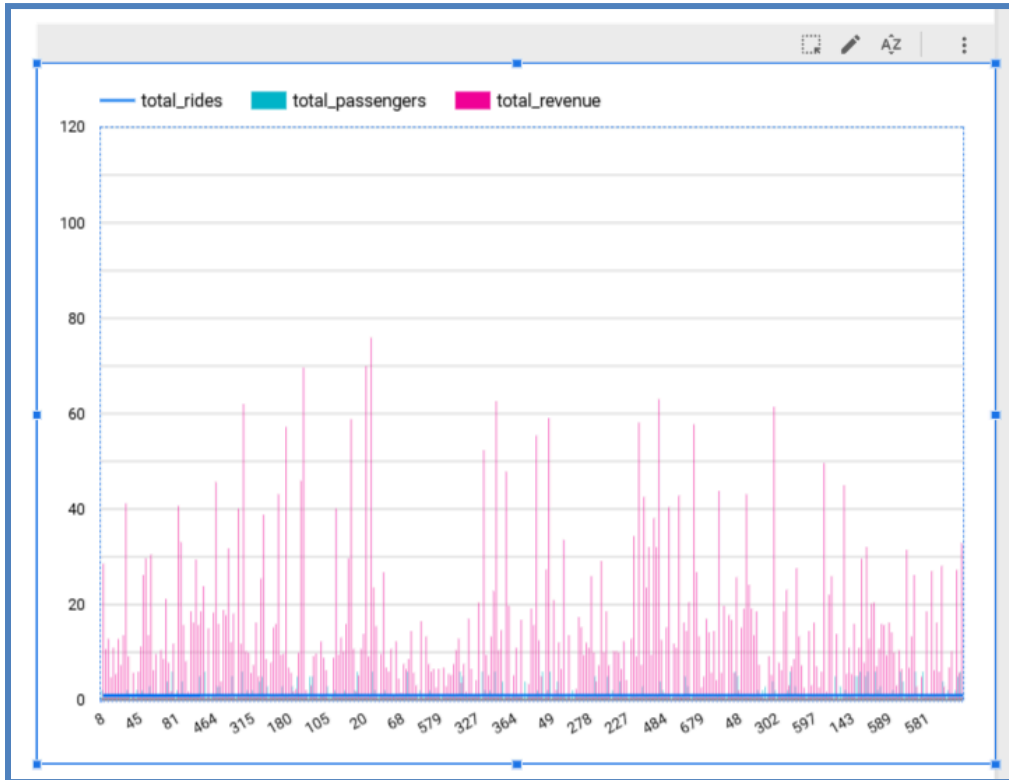
[JOB INFORMATION](#) [RESULTS](#) [CHART](#) [JSON](#)

Row	dashboard_sort	minute	total_rides
1	1	2022-10-26 03:13:00 UTC	
2	2	2022-10-26 03:13:00 UTC	
3	3	2022-10-26 03:13:00 UTC	
4	4	2022-10-26 03:13:00 UTC	
5	5	2022-10-26 03:13:00 UTC	

Results per page: 50 1 - 50 of 685



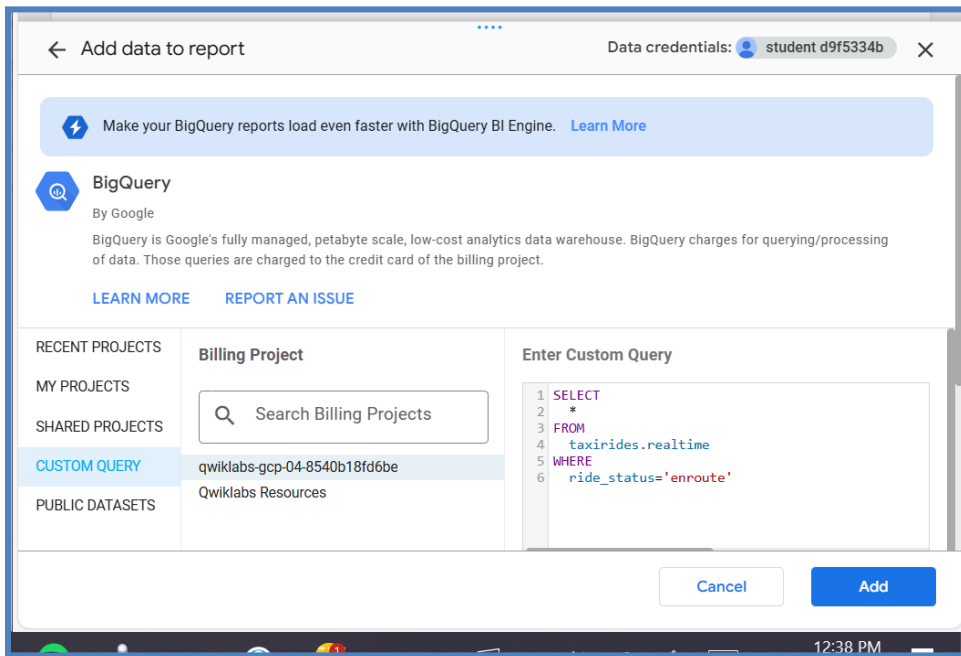
Created chart:





Task 8. Create a time series dashboard

adding the custom query:



Create a time series chart:

