National University of Sciences and Technology (NUST)
School of Electrical Engineering and Computer Science

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

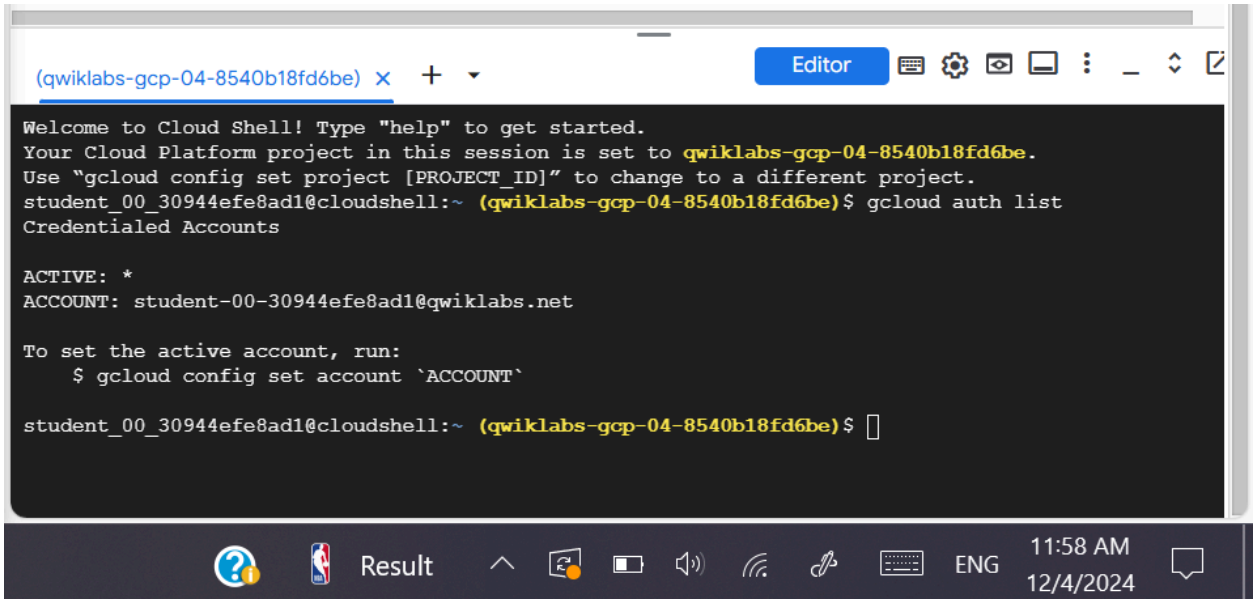
Lab Tasks

Go through the following link:

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

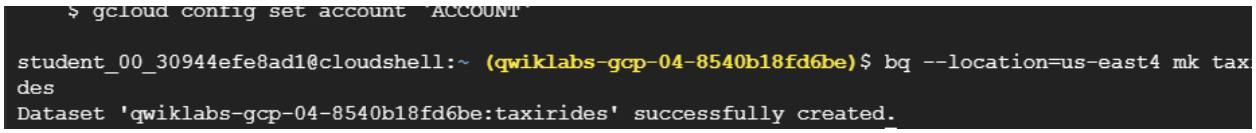
which will take you to the 'Creating a Streaming Data Pipeline for a Real-Time Dashboard with Dataflow' page. The list of tasks is given below. Make sure to take screenshots of each task as you will need to add them in the solution section given below.

Setting up qwilabs account:

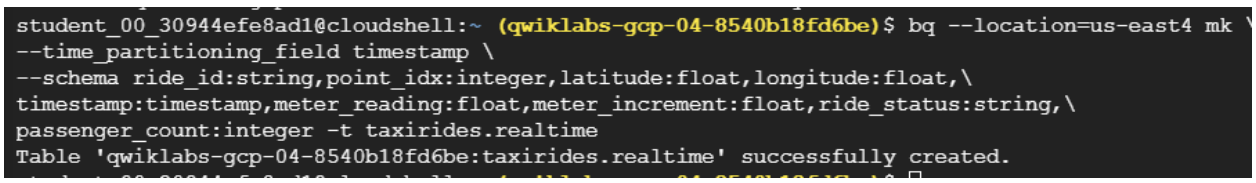


Task 1. Create a BigQuery dataset

create the **`taxirides`** dataset:



create the **`taxirides.realtime`** table



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-
ining/bdml/taxisrcdata/schema.json  gs://qwiklabs-gcp-04-8540b18fd6be-bucket/tmp/schema.json
gcloud storage cp gs://cloud-training/bdml/taxisrcdata/transform.js  gs://qwiklabs-gcp-04-8540b18f
e-bucket/tmp/transform.js
gcloud storage cp gs://cloud-training/bdml/taxisrcdata/rt_taxidata.csv  gs://qwiklabs-gcp-04-8540b
d6be-bucket/tmp/rt_taxidata.csv
Copying gs://cloud-training/bdml/taxisrcdata/schema.json to gs://qwiklabs-gcp-04-8540b18fd6be-buck
tmp/schema.json
  Completed files 1/1 | 610.0B/610.0B
Copying gs://cloud-training/bdml/taxisrcdata/transform.js to gs://qwiklabs-gcp-04-8540b18fd6be-buc
/tmp/transform.js
  Completed files 1/1 | 435.0B/435.0B
Copying gs://cloud-training/bdml/taxisrcdata/rt_taxidata.csv to gs://qwiklabs-gcp-04-8540b18fd6be-
ket/tmp/rt_taxidata.csv
  Completed files 1/1 | 108.3kiB/108.3kiB
```

Task 3. Set up a Dataflow Pipeline

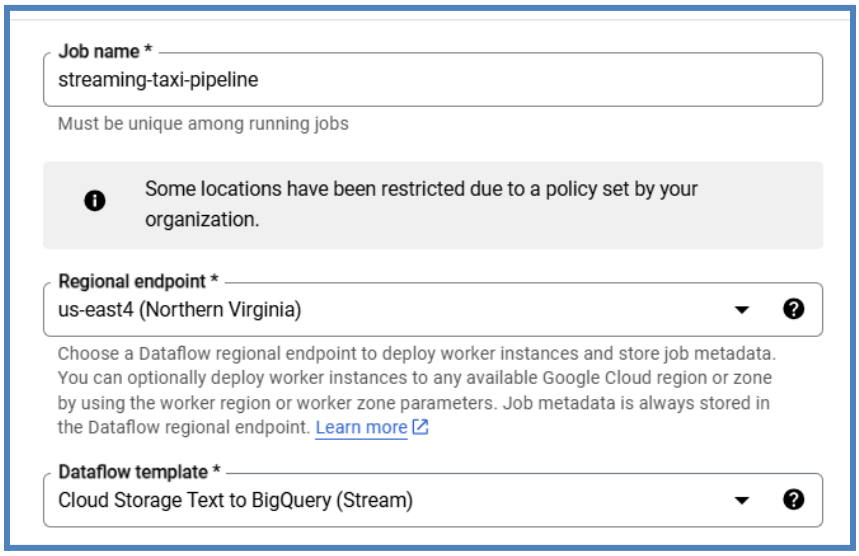
in this task, we set up a streaming data pipeline to read files from your Cloud Storage bucket and write data to BigQuery. (Dataflow is a serverless way to carry out data analysis.)

Restart the connection to the Dataflow API.

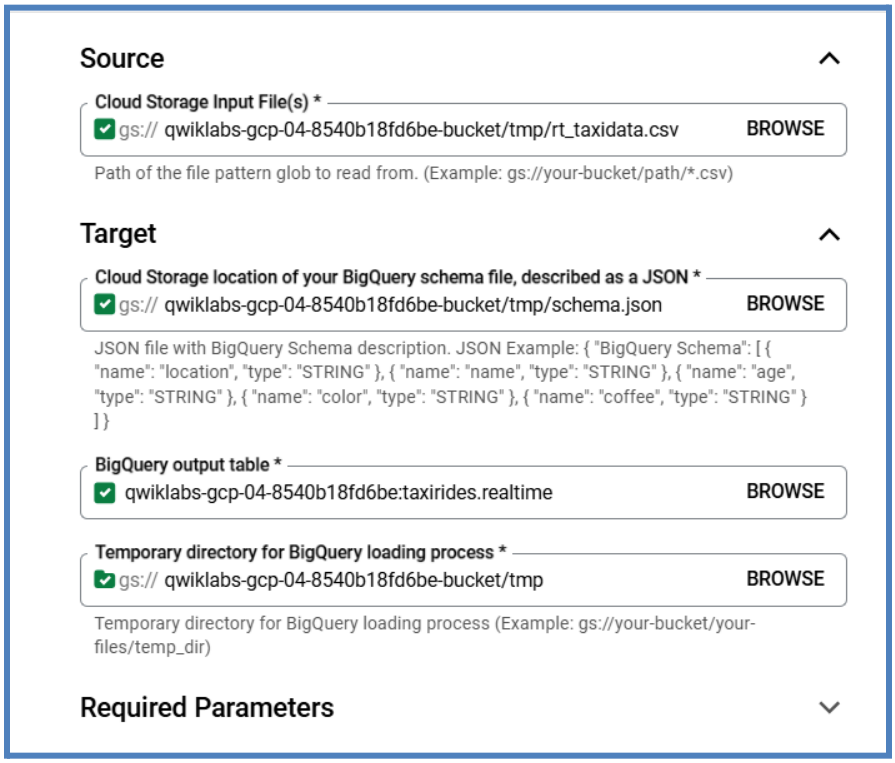
```
student_00_30944efe8ad1@cloudshell:~ (qwiklabs-gcp-04-8540b18fd6be)$ gcloud services disable dataf
.googleapis.com
gcloud services enable dataflow.googleapis.com
Operation "operations/acat.p17-71696123621-4c259990-eb36-427d-8f07-538c4118fcb2" finished successf
y.
Operation "operations/acf.p2-71696123621-f0bb7b27-dfb6-4970-8384-7842a4d363be"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

Create a new streaming pipeline:

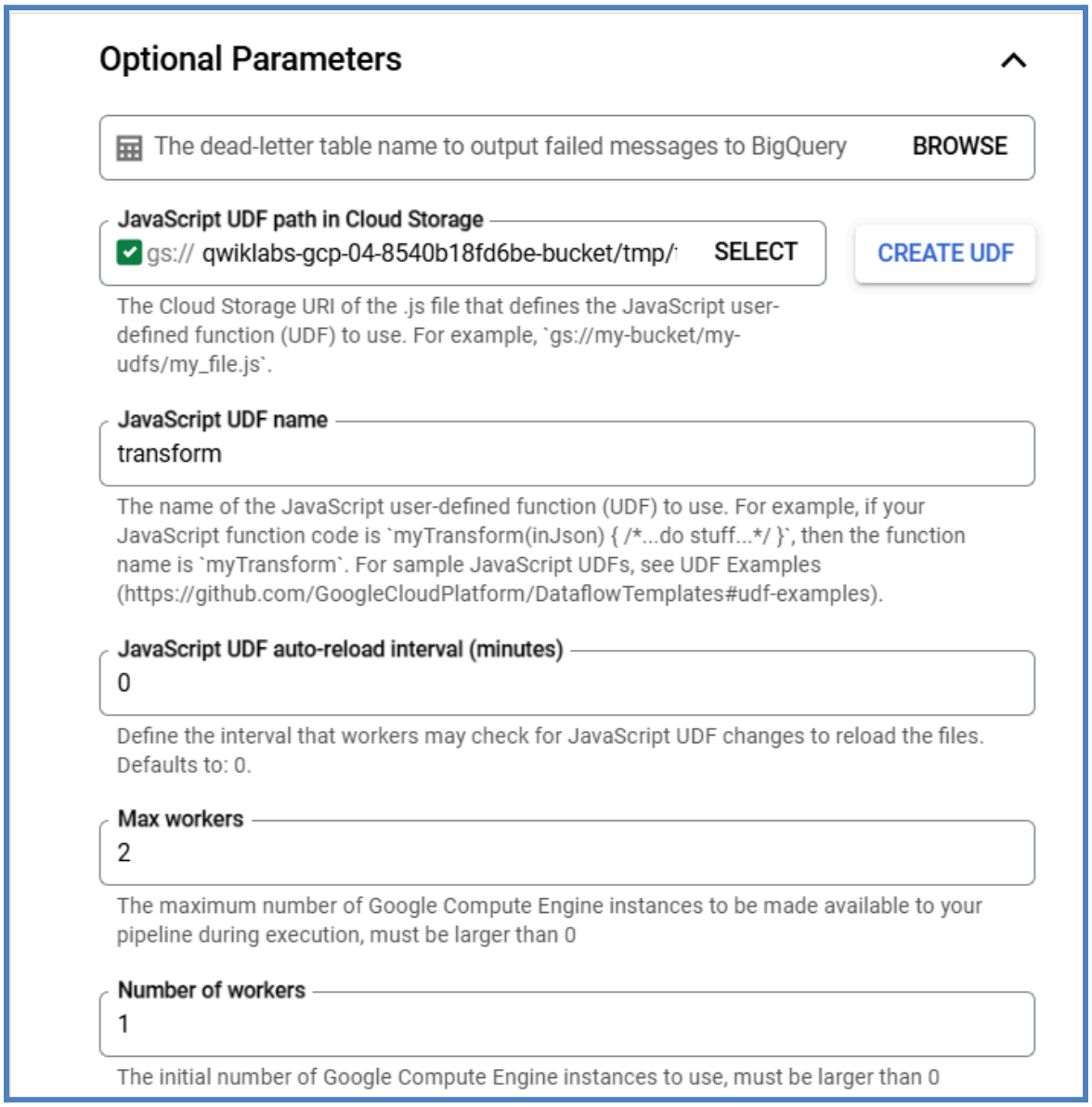
setting up job name, endpoint and dataflow template



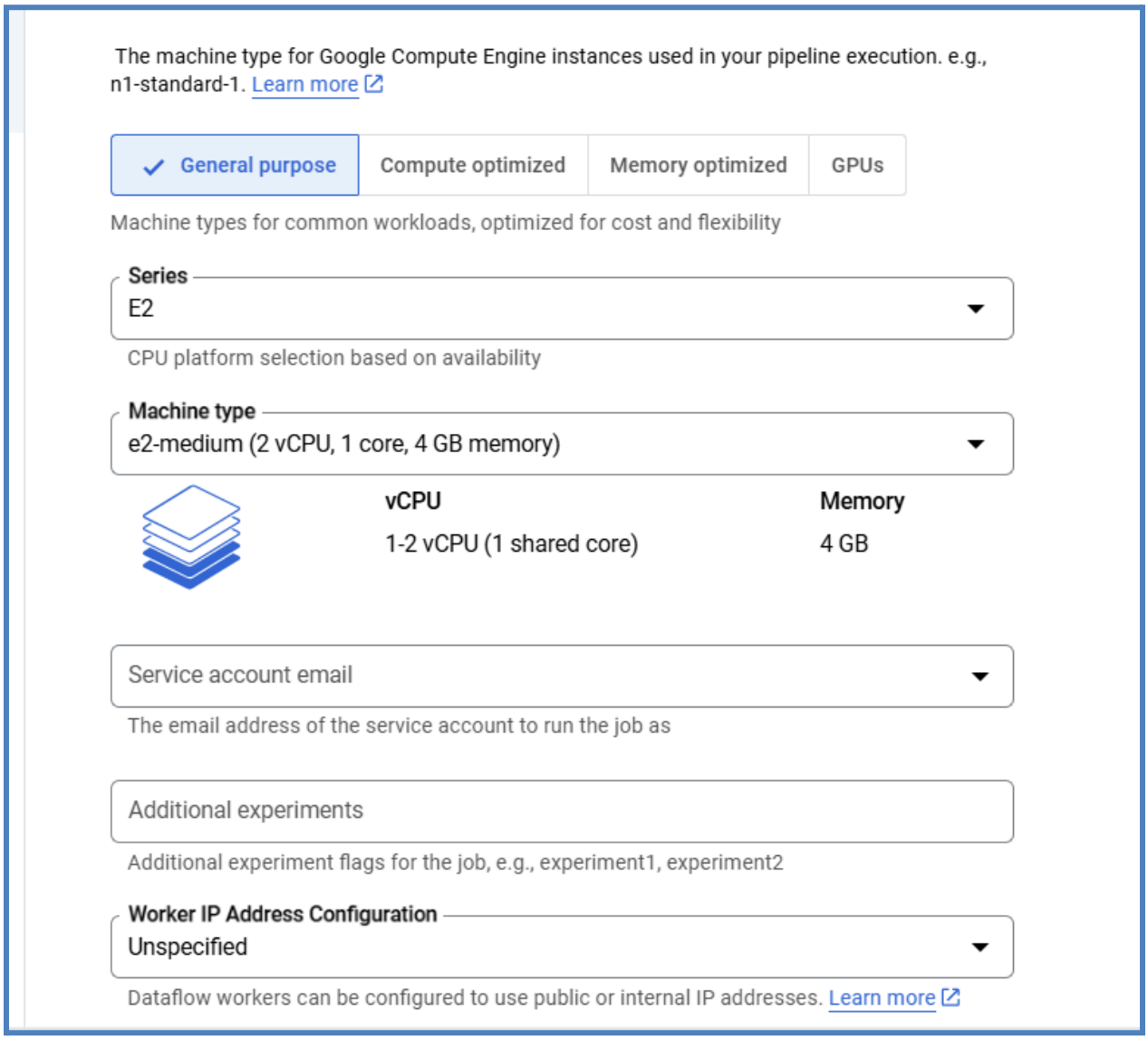
setting up required parameters



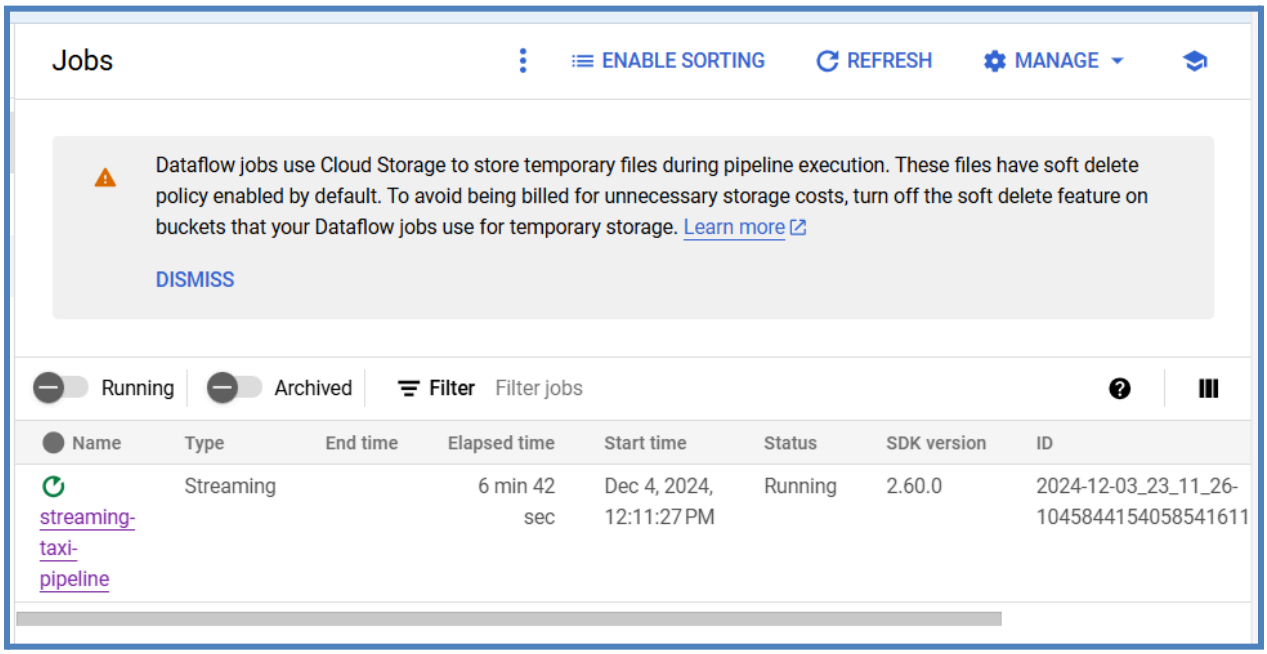
setting up optional parameters:

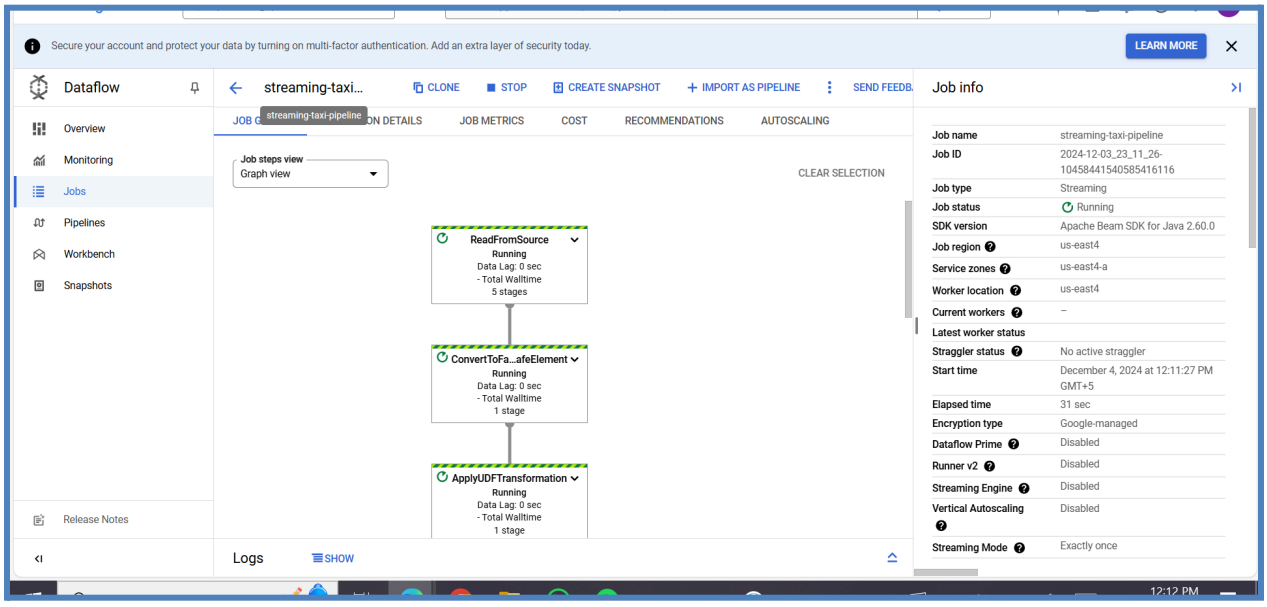


setting up machine type:



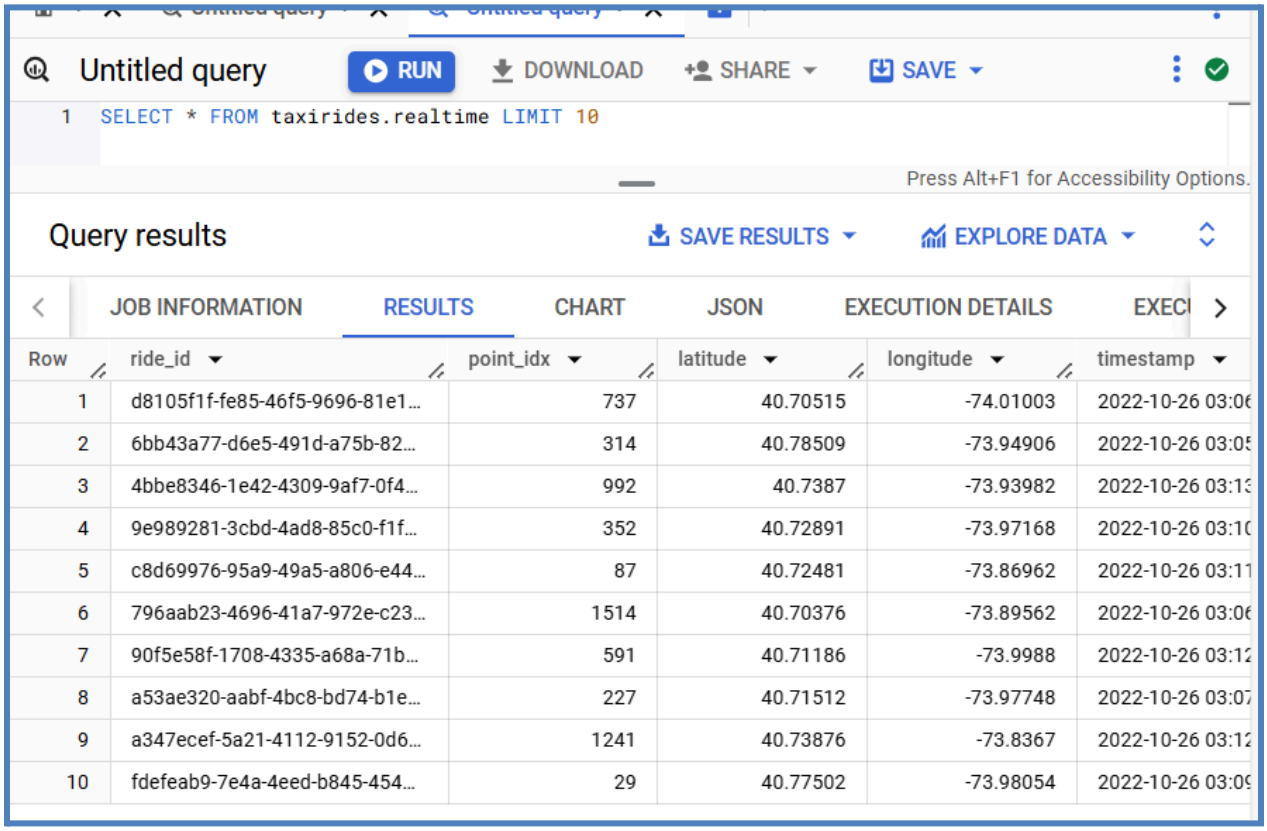
finally, our `streaming-taxi-pipeline` job is created:





Task 4. Analyze the taxi data using BigQuery

I ran the select query to dispaly top 10 rows;



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

```
WITH streaming_data AS (

SELECT
  timestamp,
  TIMESTAMP_TRUNC(timestamp, HOUR, 'UTC') AS hour,
  TIMESTAMP_TRUNC(timestamp, MINUTE, 'UTC') AS minute,
  TIMESTAMP_TRUNC(timestamp, SECOND, 'UTC') AS second,
  ride_id,
  latitude,
  longitude,
  meter_reading,
  ride_status,
  passenger_count
FROM
```

Press Alt+F1 for Accessibility Options.

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

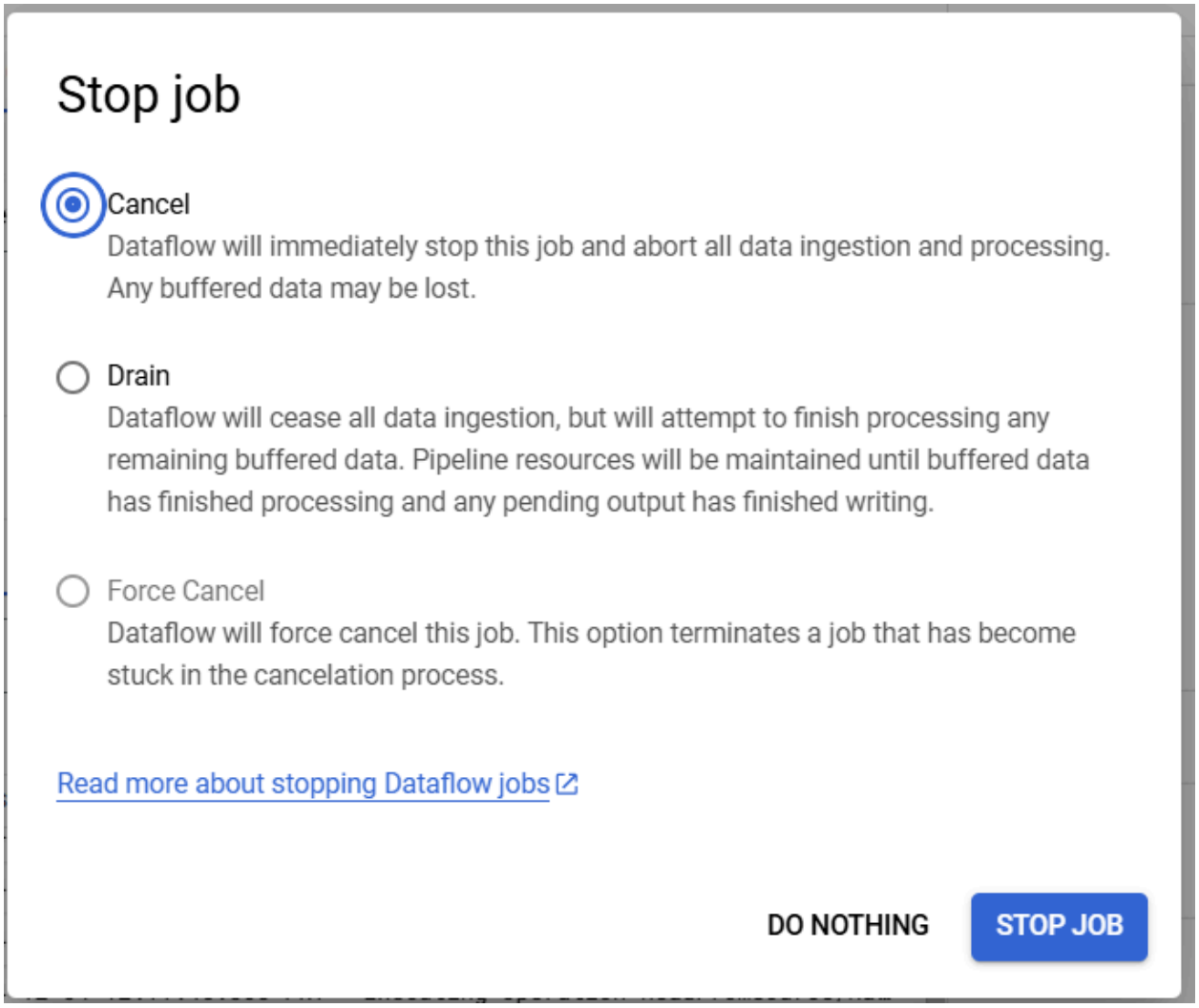
Name *
Saleha's Saved Query

Region *
us-east4 (Northern Virginia)

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

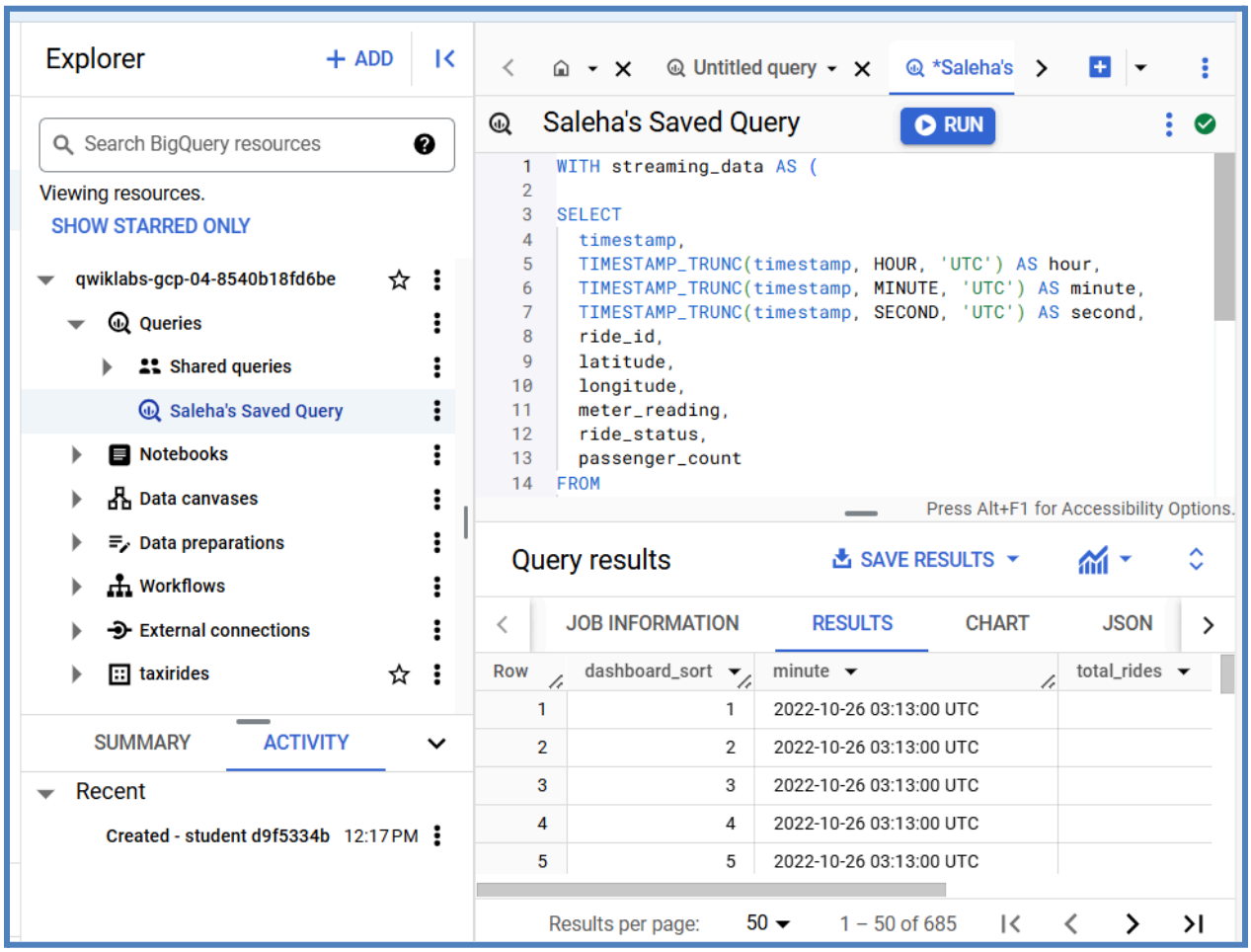
SAVE CANCEL

Task 6. Stop the Dataflow Job

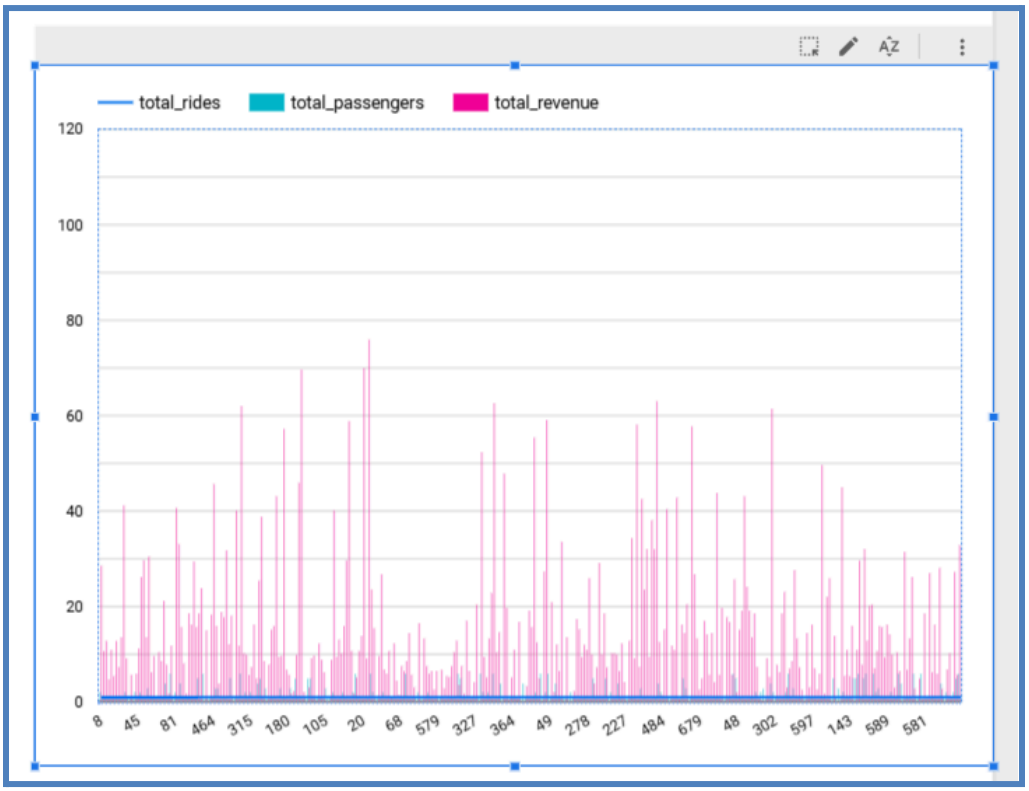


Task 7. Create a real-time dashboard

running the saved query:

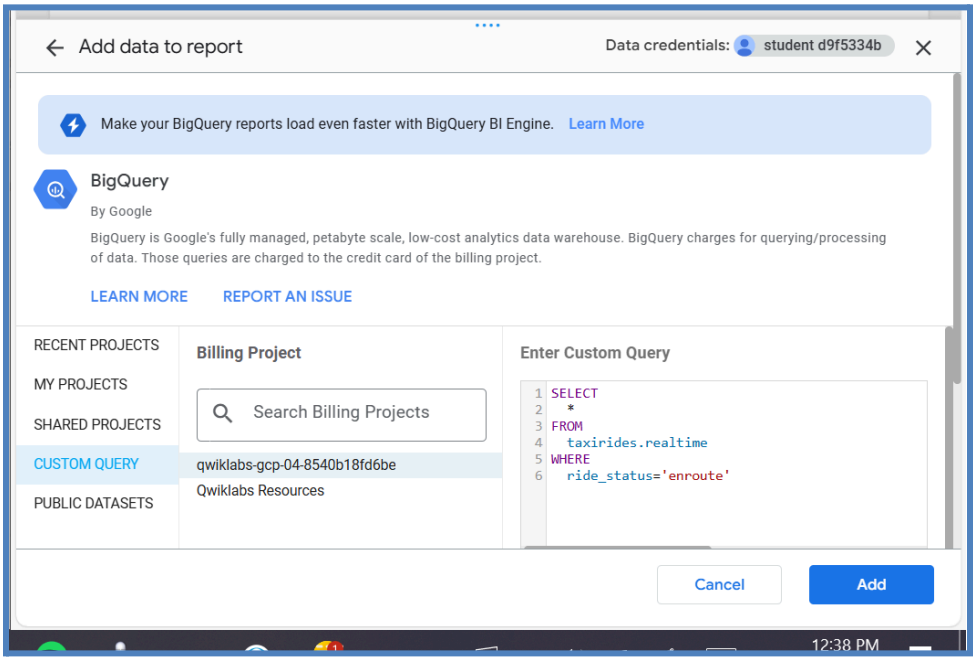


Created chart:



Task 8. Create a time series dashboard

adding the custom query:



Create a time series chart: