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

Department of Computing

SE-315: Cloud Computing

Lab 12: Serverless App Dev – Creating a Streaming Data Pipeline for a Real-Time Dashboard with Dataflow – Part B

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

Date: 11.12.24

Lab 12: Serverless App Dev – Creating a Streaming Data Pipeline for a Real-Time Dashboard with Dataflow – Part B

Introduction:

This lab task is an extension of previous lab task where you owned a fleet of New York City taxi cabs and you were live monitoring your business. In this lab, you have to build a similar streaming data pipeline but for the citywide payroll data to explore it and 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

Recapture your knowledge of previous lab by going through the following link:

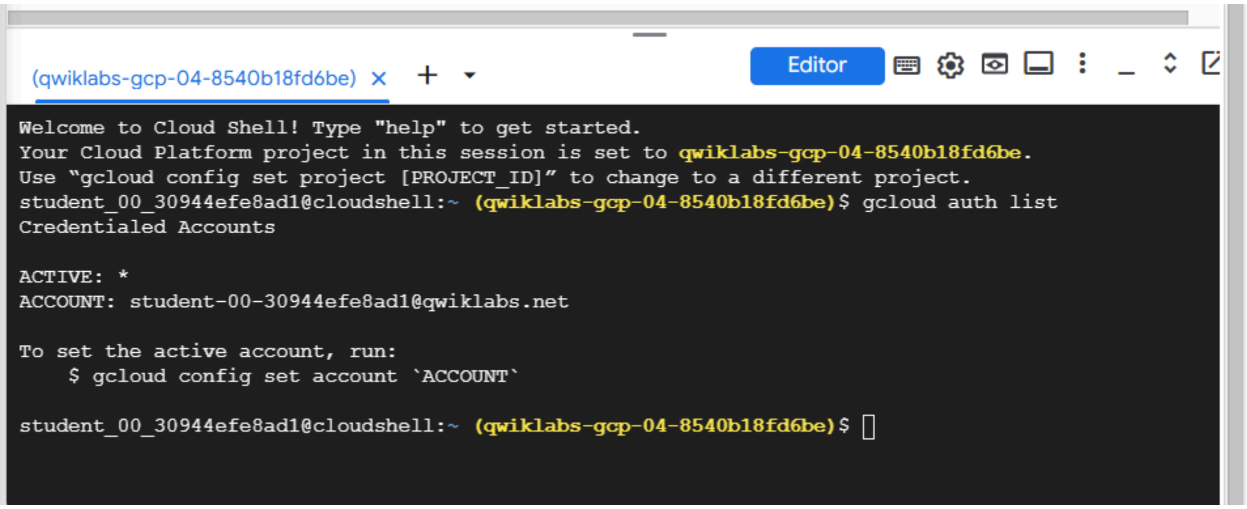
https://www.cloudskillsboost.google/focuses/19077?parent=catalog

Citywide Payroll Data download link:
https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e

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.

(Hint: You can take help from previous lab tasks as the list of tasks are same, however, the dataset is changed.)

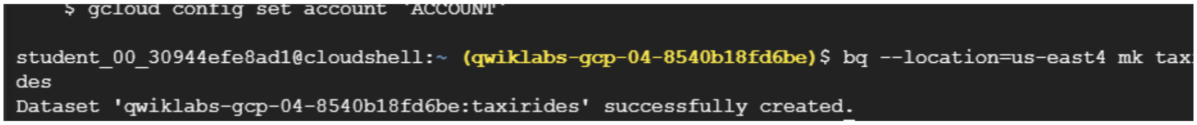
Setting up qwilabs account:



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

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

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

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

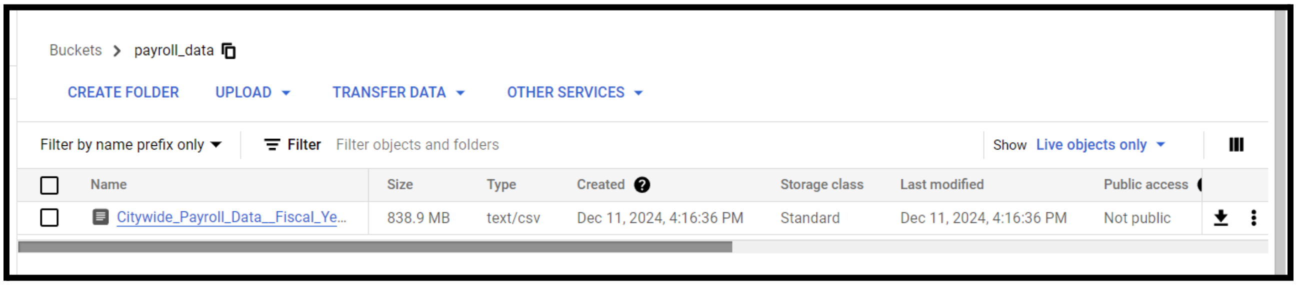
Task 1. Create a BigQuery dataset

create the **Citywide Payroll** dataset:



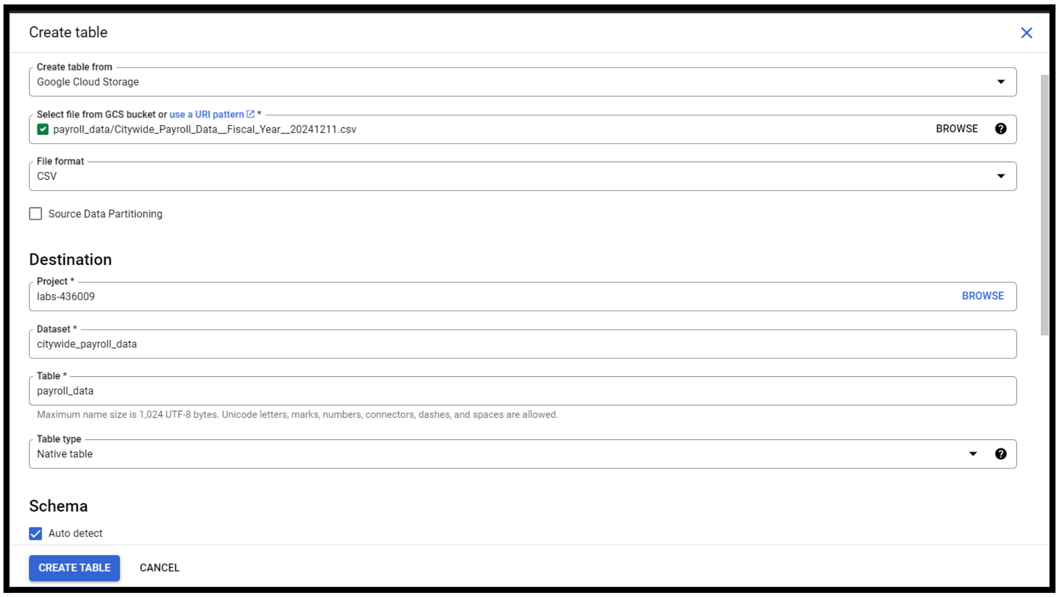
```
    $ gcloud config set account `ACCOUNT`

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

Uploading the data on a GCloud Bucket:



	Name	Size	Type	Created	Storage class	Last modified	Public access
□	Citywide_Payroll_Data__Fiscal_Ye...	838.9 MB	text/csv	Dec 11, 2024, 4:16:36 PM	Standard	Dec 11, 2024, 4:16:36 PM	Not public

creating the payroll_data table:



Create table

Create table from: Google Cloud Storage

Select file from GCS bucket or use a URI pattern *: ☑ payroll_data/Citywide_Payroll_Data__Fiscal_Year__20241211.csv

File format: CSV

□ Source Data Partitioning

Destination

Project *: labs-436009

Dataset *: citywide_payroll_data

Table *: payroll_data

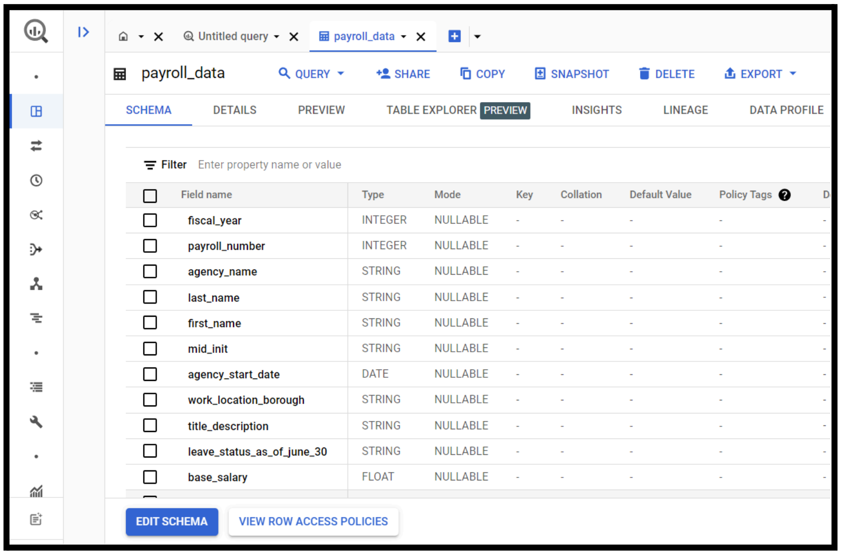
Maximum name size is 1,024 UTF-8 bytes. Unicode letters, marks, numbers, connectors, dashes, and spaces are allowed.

Table type: Native table

Schema

☑ Auto detect

attributes within payroll_data table:



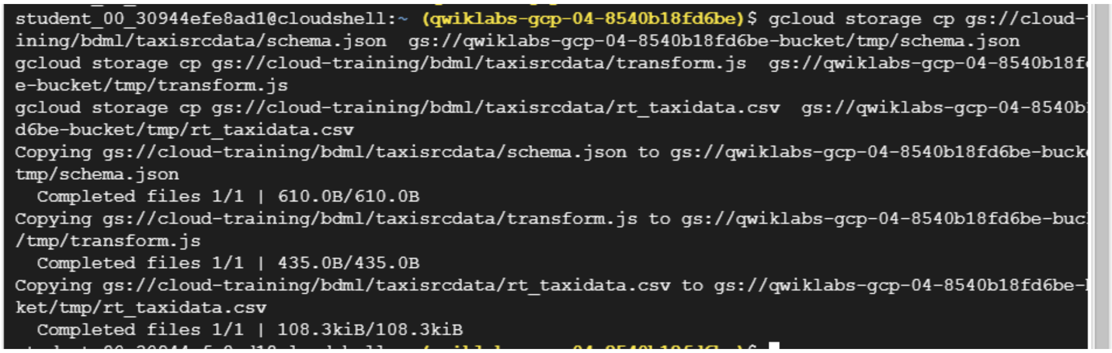
payroll_data

□	Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	D
□	fiscal_year	INTEGER	NULLABLE	-	-	-	-	-
□	payroll_number	INTEGER	NULLABLE	-	-	-	-	-
□	agency_name	STRING	NULLABLE	-	-	-	-	-
□	last_name	STRING	NULLABLE	-	-	-	-	-
□	first_name	STRING	NULLABLE	-	-	-	-	-
□	mid_init	STRING	NULLABLE	-	-	-	-	-
□	agency_start_date	DATE	NULLABLE	-	-	-	-	-
□	work_location_borough	STRING	NULLABLE	-	-	-	-	-
□	title_description	STRING	NULLABLE	-	-	-	-	-
□	leave_status_as_of_june_30	STRING	NULLABLE	-	-	-	-	-
□	base_salary	FLOAT	NULLABLE	-	-	-	-	-

Task 2. Copy required lab artifacts

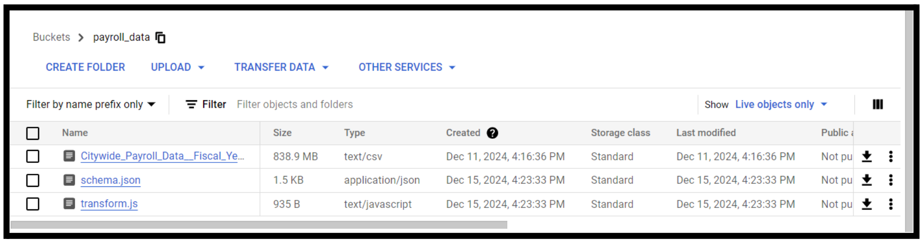
run the following commands to move files needed for the Dataflow job.

here, we copy the files from the google storage bucket using the **gcloud storage cp** command



```
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
```

resulting bucket:

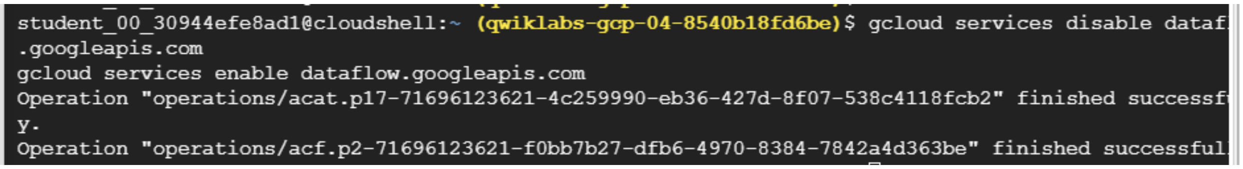


□	Name	Size	Type	Created	Storage class	Last modified	Public a
□	Citywide_Payroll_Data__Fiscal_Ye...	838.9 MB	text/csv	Dec 11, 2024, 4:16:36 PM	Standard	Dec 11, 2024, 4:16:36 PM	Not pu
□	schema.json	1.5 KB	application/json	Dec 15, 2024, 4:23:33 PM	Standard	Dec 15, 2024, 4:23:33 PM	Not pu
□	transform.js	935 B	text/javascript	Dec 15, 2024, 4:23:33 PM	Standard	Dec 15, 2024, 4:23:33 PM	Not pu

Task 3. Set up a Dataflow Pipeline

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

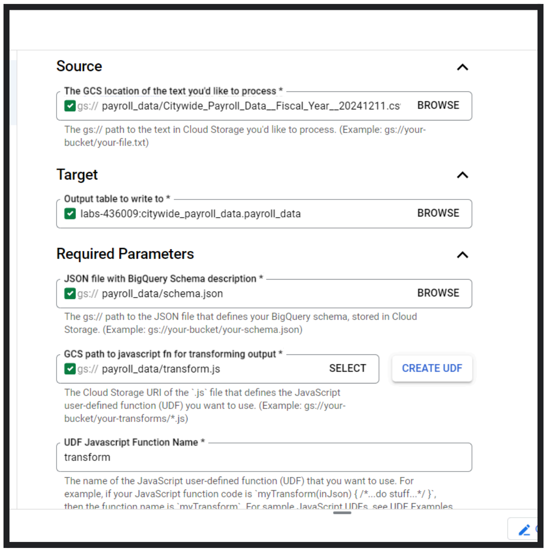
Restart the connection to the Dataflow API.



```
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:



Source

The GCS location of the text you'd like to process *
☑ gs:// payroll_data/Citywide_Payroll_Data__Fiscal_Year__20241211.cs BROWSE
The gs:// path to the text in Cloud Storage you'd like to process. (Example: gs://your-bucket/your-file.txt)

Target

Output table to write to *
☑ labs-436009:citywide_payroll_data.payroll_data BROWSE

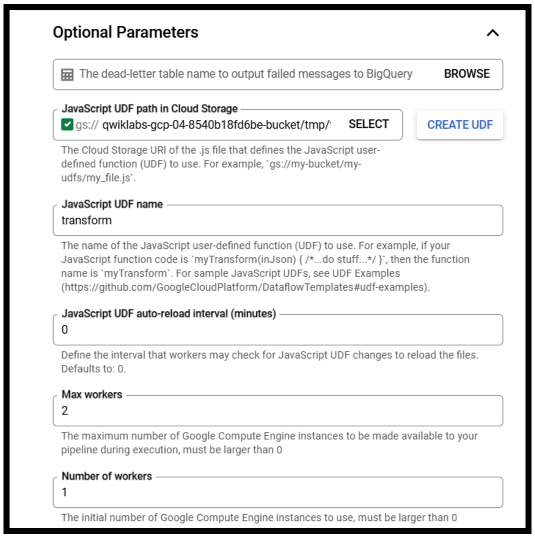
Required Parameters

JSON file with BigQuery Schema description *
☑ gs:// payroll_data/schema.json BROWSE
The gs:// path to the JSON file that defines your BigQuery schema, stored in Cloud Storage. (Example: gs://your-bucket/your-schema.json)

GCS path to javascript fn for transforming output *
☑ gs:// payroll_data/transform.js SELECT CREATE UDF
The Cloud Storage URI of the `.js` file that defines the JavaScript user-defined function (UDF) you want to use. (Example: gs://your-bucket/your-transforms/*.js)

UDF Javascript Function Name *
transform
The name of the JavaScript user-defined function (UDF) that you want 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

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`.

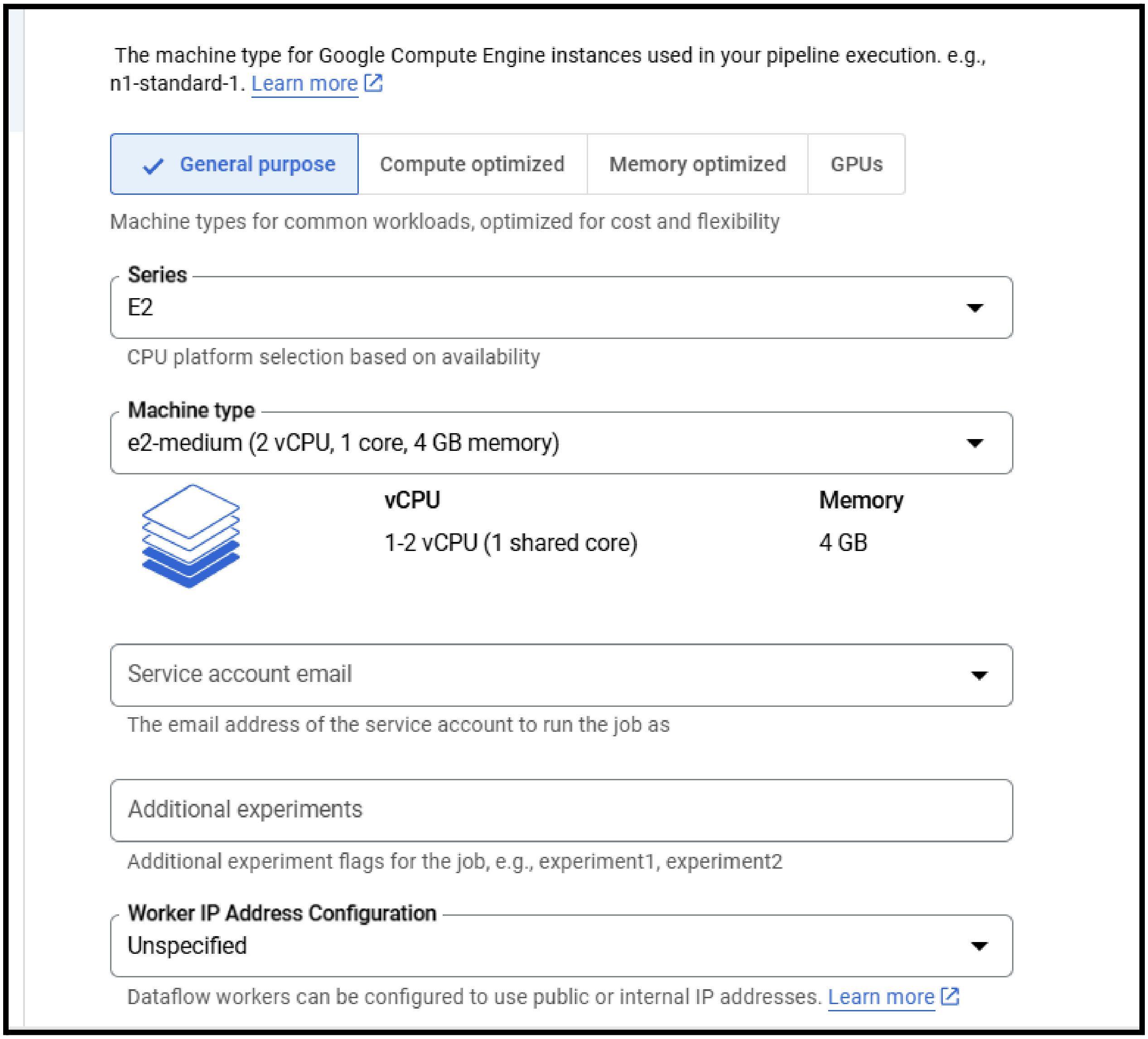
JavaScript UDF name
transform
The name of the JavaScript user-defined function (UDF) to use. For example, if your JavaScript function code is `myTransform(inJson) { /*...do stuff...*/ }`, then the function name is `myTransform`. For sample JavaScript UDFs, see UDF Examples (https://github.com/GoogleCloudPlatform/DataflowTemplates#udf-examples).

JavaScript UDF auto-reload interval (minutes)
0
Define the interval that workers may check for JavaScript UDF changes to reload the files. Defaults to: 0.

Max workers
2
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
1
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

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	Memory
1-2 vCPU (1 shared core)	4 GB

Service account email

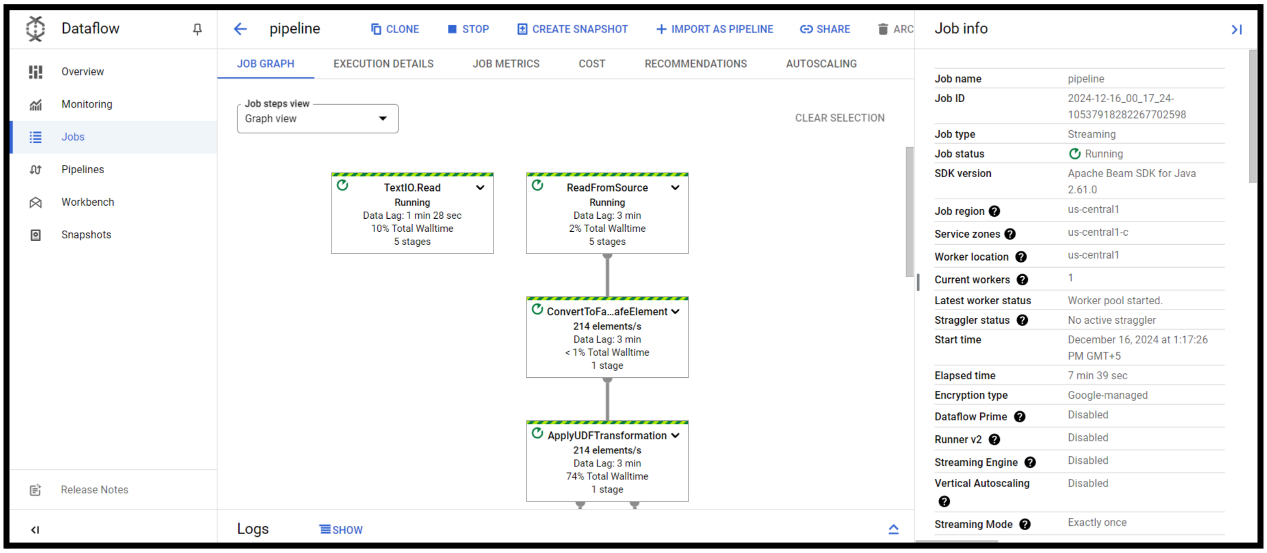
The email address of the service account to run the job as

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 finally, our `payroll` pipeline job is created:



Job steps view: Graph view

TextIO.Read
Running
Data Lag: 1 min 28 sec
10% Total Walltime
5 stages

ReadFromSource
Running
Data Lag: 3 min
2% Total Walltime
5 stages

ConvertToFa...afeElement
214 elements/s
Data Lag: 3 min
< 1% Total Walltime
1 stage

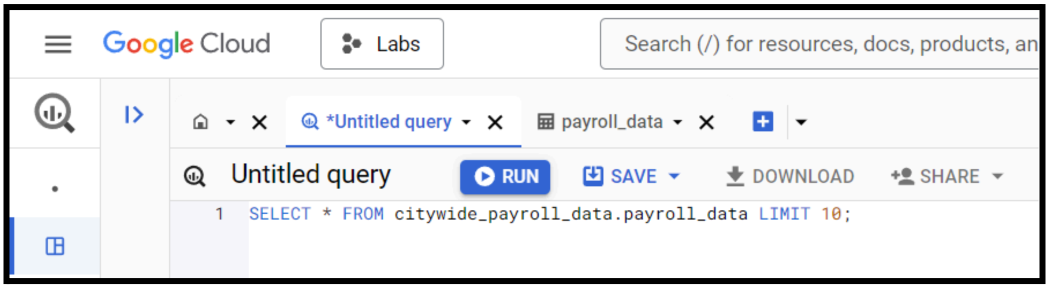
ApplyUDFTransformation
214 elements/s
Data Lag: 3 min
74% Total Walltime
1 stage

Job info

Job name	pipeline
Job ID	2024-12-16_00_17_24-10537918282267702598
Job type	Streaming
Job status	Running
SDK version	Apache Beam SDK for Java 2.61.0
Job region	us-central1
Service zones	us-central1-c
Worker location	us-central1
Current workers	1
Latest worker status	Worker pool started.
Straggler status	No active straggler
Start time	December 16, 2024 at 1:17:26 PM GMT+5
Elapsed time	7 min 39 sec
Encryption type	Google-managed
Dataflow Prime	Disabled
Runner v2	Disabled
Streaming Engine	Disabled
Vertical Autoscaling	Disabled
Streaming Mode	Exactly once

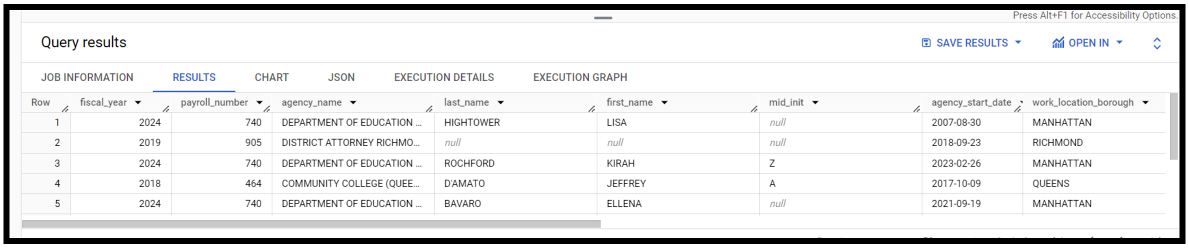
Task 4. Analyze the taxi data using BigQuery

I ran the select query to display top 10 rows;



```
1   SELECT * FROM citywide_payroll_data.payroll_data LIMIT 10;
```

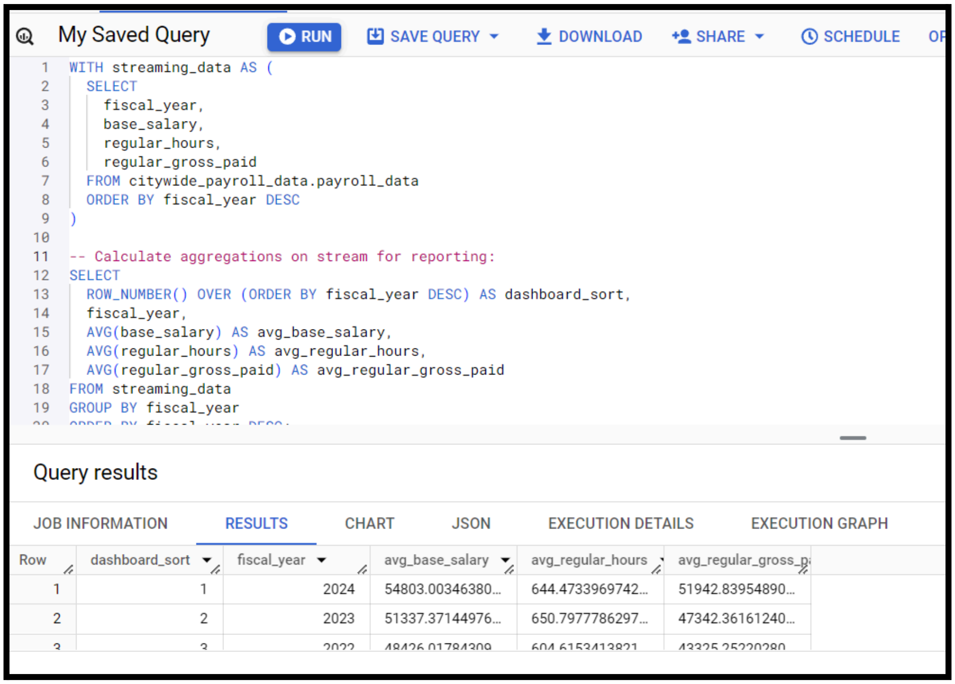
top 10 rows in payroll_data table:



Query results

Row	fiscal_year	payroll_number	agency_name	last_name	first_name	mid_init	agency_start_date	work_location_borough
1	2024	740	DEPARTMENT OF EDUCATION ...	HIGHTOWER	LISA	null	2007-08-30	MANHATTAN
2	2019	905	DISTRICT ATTORNEY RICHMO...	null	null	null	2018-09-23	RICHMOND
3	2024	740	DEPARTMENT OF EDUCATION ...	ROCHFORD	KIRAH	Z	2023-02-26	MANHATTAN
4	2018	464	COMMUNITY COLLEGE (QUEE...	D'AMATO	JEFFREY	A	2017-10-09	QUEENS
5	2024	740	DEPARTMENT OF EDUCATION ...	BAVARO	ELLENA	null	2021-09-19	MANHATTAN

Task 5. Perform aggregations on the stream for reporting

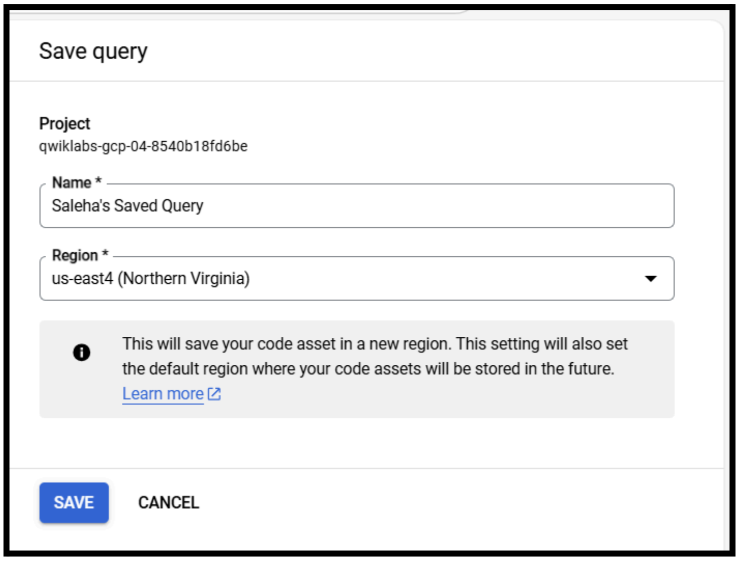


```
WITH streaming_data AS (
  SELECT
    fiscal_year,
    base_salary,
    regular_hours,
    regular_gross_paid
  FROM citywide_payroll_data.payroll_data
  ORDER BY fiscal_year DESC
)

-- Calculate aggregations on stream for reporting:
SELECT
  ROW_NUMBER() OVER (ORDER BY fiscal_year DESC) AS dashboard_sort,
  fiscal_year,
  AVG(base_salary) AS avg_base_salary,
  AVG(regular_hours) AS avg_regular_hours,
  AVG(regular_gross_paid) AS avg_regular_gross_paid
FROM streaming_data
GROUP BY fiscal_year
ORDER BY fiscal_year DESC;
```

Query results

Row	dashboard_sort	fiscal_year	avg_base_salary	avg_regular_hours	avg_regular_gross_pa
1	1	2024	54803.00346380...	644.4733969742...	51942.83954890...
2	2	2023	51337.37144976...	650.7977786297...	47342.36161240...
3	3	2022	48426.01784309	604.6153413821	43325.25220280

Saving the query:


Save query

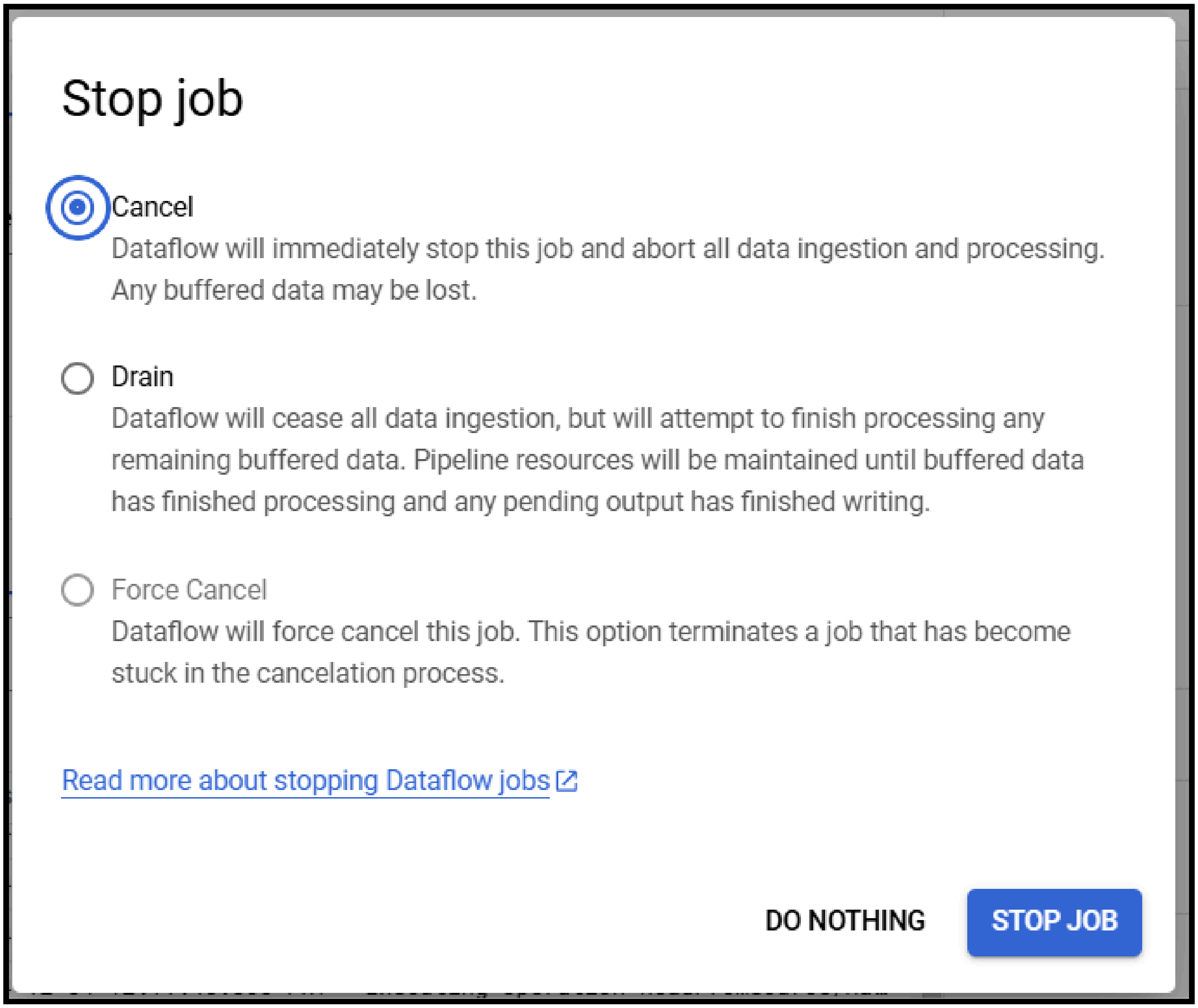
Project
qwiklabs-gcp-04-8540b18fd6be

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

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.

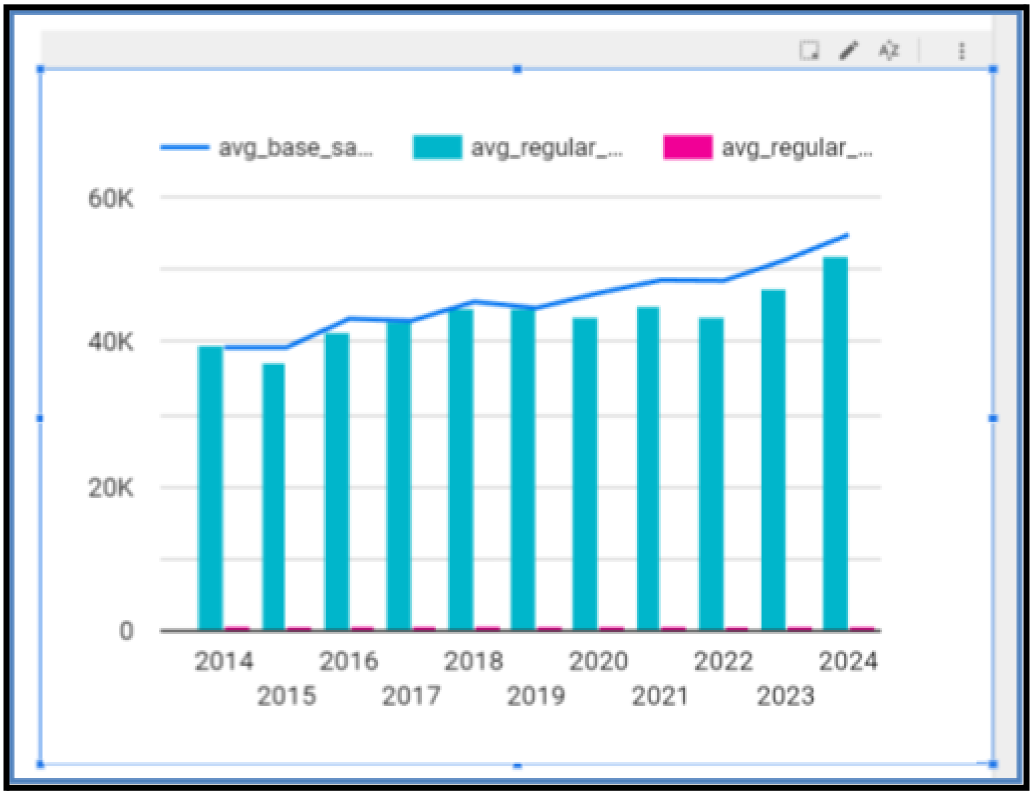
○ Drain
Dataflow will cease all data ingestion, but will attempt to finish processing any remaining buffered data. Pipeline resources will be maintained until buffered data has finished processing and any pending output has finished writing.

○ 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

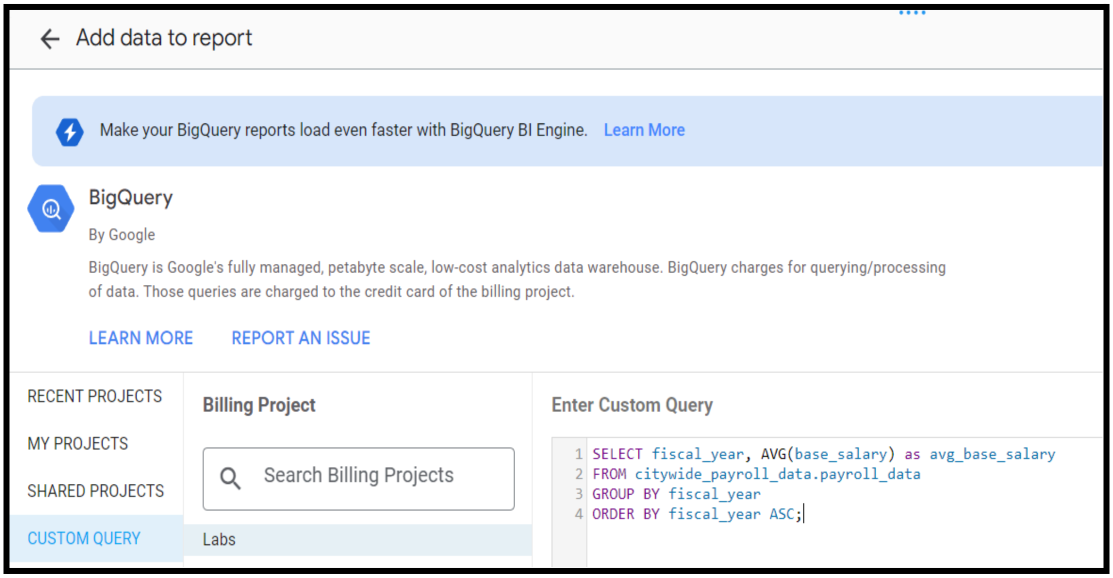
Task 7. Create a real-time dashboard

running the saved query and creating the chart displaying salaries:



Task 8. Create a time series dashboard

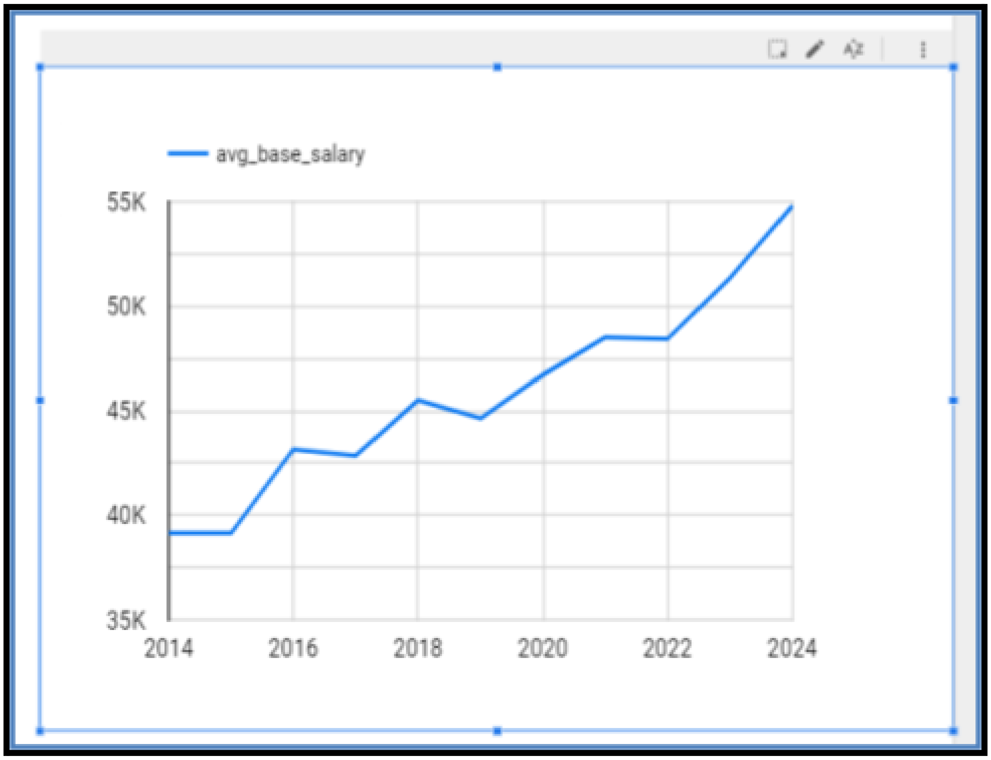
adding the custom query to display average salary from payroll_data:



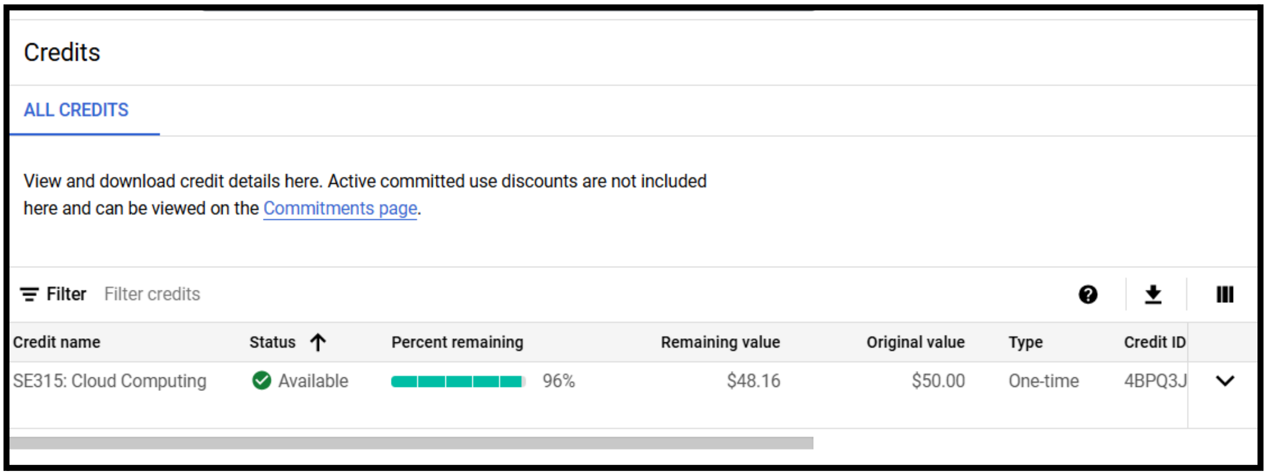
BigQuery is Google's fully managed, petabyte scale, low-cost analytics data warehouse. BigQuery charges for querying/processing of data. Those queries are charged to the credit card of the billing project.

```
SELECT fiscal_year, AVG(base_salary) as avg_base_salary
FROM citywide_payroll_data.payroll_data
GROUP BY fiscal_year
ORDER BY fiscal_year ASC;
```

Create a time series chart:



Remaining credits:



Credits

View and download credit details here. Active committed use discounts are not included here and can be viewed on the Commitments page.

Credit name	Status ↑	Percent remaining	Remaining value	Original value	Type	Credit ID
SE315: Cloud Computing	✓ Available	96%	$48.16	$50.00	One-time	4BPQ3J