Erick Franco
Cs430p
3/6/25

09.1g.3..2
09.1g.4..4
09.1g.9..5
09.2g.3..6
09.2g.6..6
09.2g.8..7
09.2g.9..7
09.2g.10..7
09.2g.11..8
09.2g.12..10
09.3g.6..11
09.3g.8..11
09.4g.3..12
09.4g.4..12
09.4g.5..13
09.4g.6..13
09.4g.9..14
09.4g.12..16
09.4g.14..16
09.4g.15..17
09.4g.16..18

09.1g.3

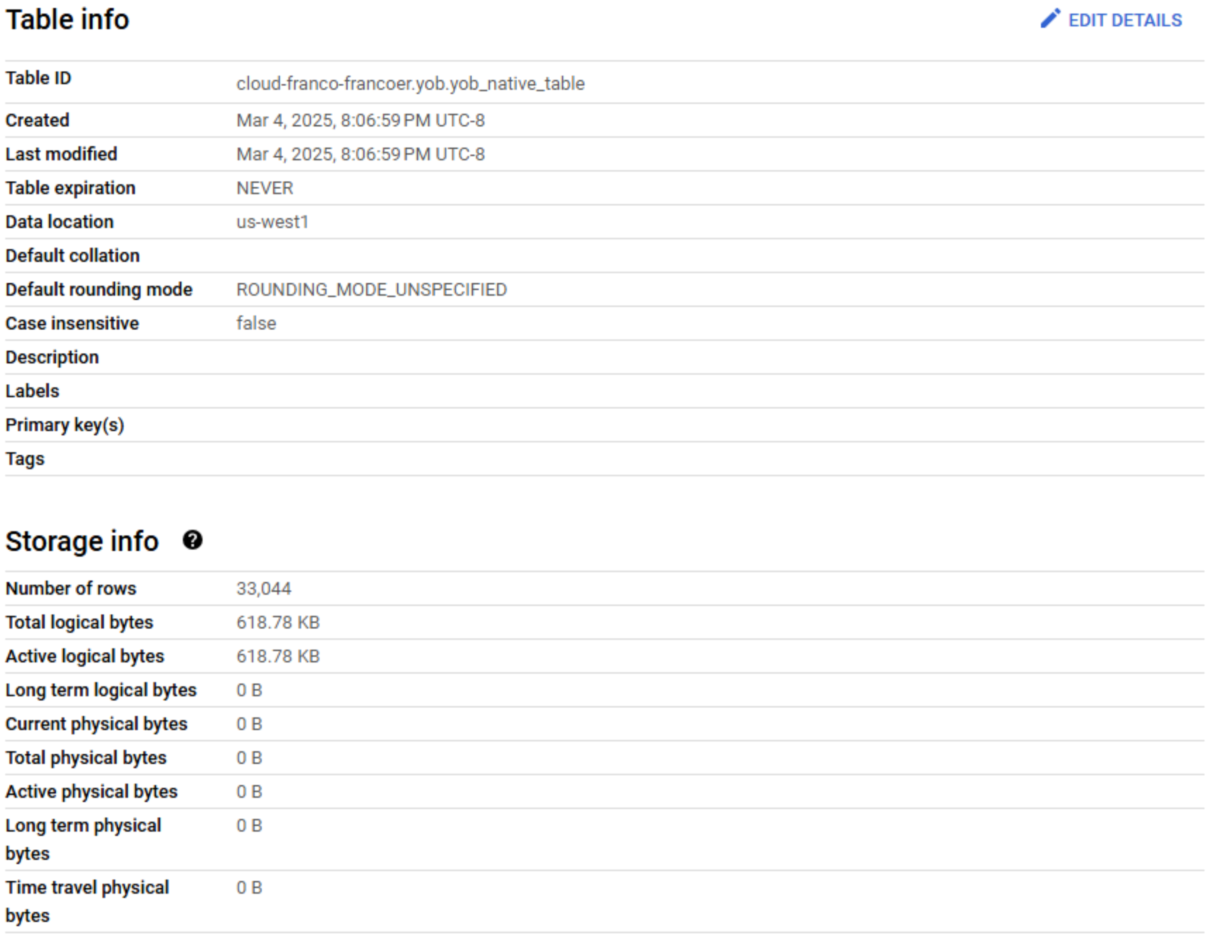


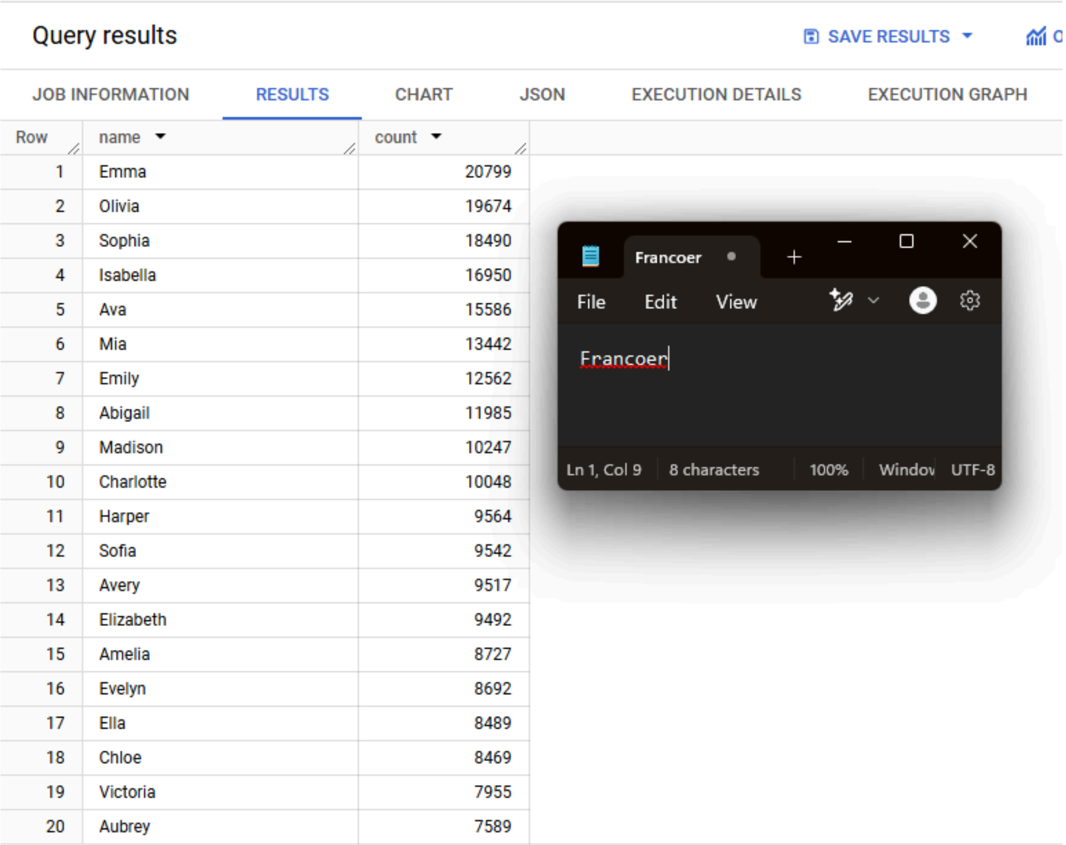
Table info

Table ID	cloud-franco-francoer.yob.yob_native_table
Created	Mar 4, 2025, 8:06:59 PM UTC-8
Last modified	Mar 4, 2025, 8:06:59 PM UTC-8
Table expiration	NEVER
Data location	us-west1
Default collation	
Default rounding mode	ROUNDING_MODE_UNSPECIFIED
Case insensitive	false
Description	
Labels	
Primary key(s)	
Tags	

Storage info

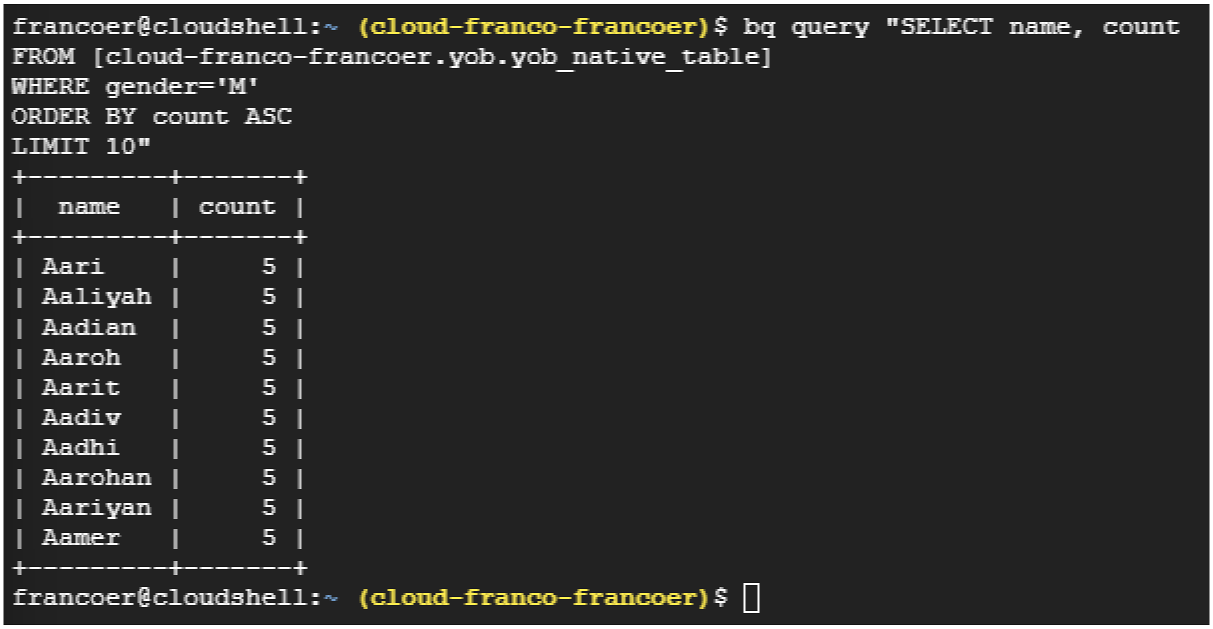
Number of rows	33,044
Total logical bytes	618.78 KB
Active logical bytes	618.78 KB
Long term logical bytes	0 B
Current physical bytes	0 B
Total physical bytes	0 B
Active physical bytes	0 B
Long term physical bytes	0 B
Time travel physical bytes	0 B

09.1g.4

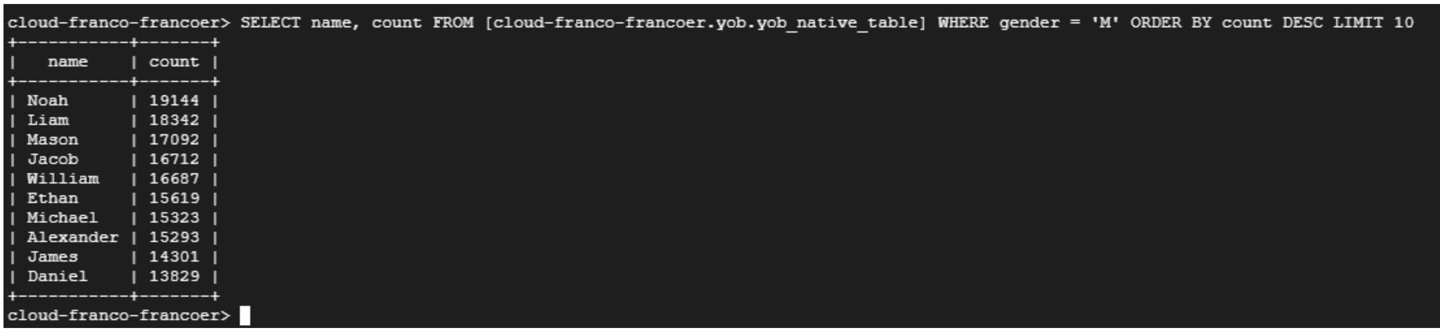


Row	name	count
1	Emma	20799
2	Olivia	19674
3	Sophia	18490
4	Isabella	16950
5	Ava	15586
6	Mia	13442
7	Emily	12562
8	Abigail	11985
9	Madison	10247
10	Charlotte	10048
11	Harper	9564
12	Sofia	9542
13	Avery	9517
14	Elizabeth	9492
15	Amelia	8727
16	Evelyn	8692
17	Ella	8489
18	Chloe	8469
19	Victoria	7955
20	Aubrey	7589

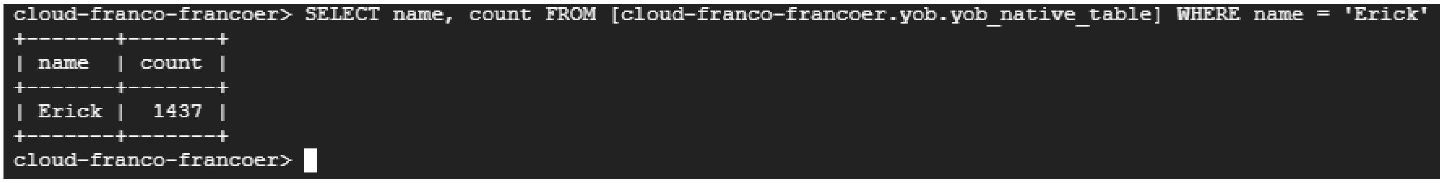
Francoer



```
francoer@cloudshell:~ (cloud-franco-francoer)$ bq query "SELECT name, count
FROM [cloud-franco-francoer.yob.yob_native_table]
WHERE gender='M'
ORDER BY count ASC
LIMIT 10"
+---------+-------+
|  name   | count |
+---------+-------+
| Aari    |     5 |
| Aaliyah |     5 |
| Aadian  |     5 |
| Aaroh   |     5 |
| Aarit   |     5 |
| Aadiv   |     5 |
| Aadhi   |     5 |
| Aarohan |     5 |
| Aariyan |     5 |
| Aamer   |     5 |
+---------+-------+
francoer@cloudshell:~ (cloud-franco-francoer)$
```

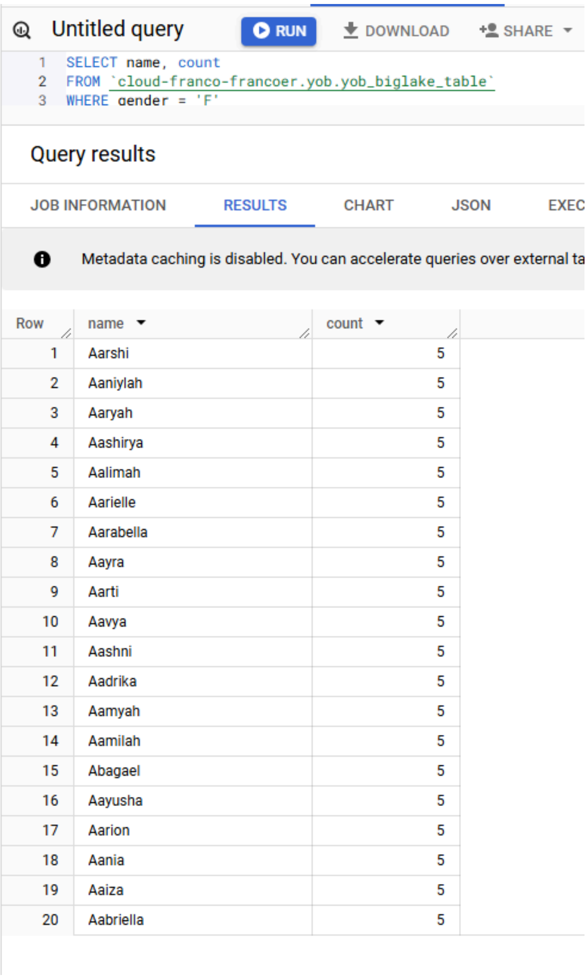


```
cloud-franco-francoer> SELECT name, count FROM [cloud-franco-francoer.yob.yob_native_table] WHERE gender = 'M' ORDER BY count DESC LIMIT 10
+-----------+-------+
|   name    | count |
+-----------+-------+
| Noah      | 19144 |
| Liam      | 18342 |
| Mason     | 17092 |
| Jacob     | 16712 |
| William   | 16687 |
| Ethan     | 15619 |
| Michael   | 15323 |
| Alexander | 15293 |
| James     | 14301 |
| Daniel    | 13829 |
+-----------+-------+
cloud-franco-francoer>
```



```
cloud-franco-francoer> SELECT name, count FROM [cloud-franco-francoer.yob.yob_native_table] WHERE name = 'Erick'
+-------+-------+
| name  | count |
+-------+-------+
| Erick |  1437 |
+-------+-------+
cloud-franco-francoer>
```

09.1g.9



```
SELECT name, count
FROM `cloud-franco-francoer.yob.yob_biglake_table`
WHERE gender = 'F'
```

Query results

Metadata caching is disabled. You can accelerate queries over external ta

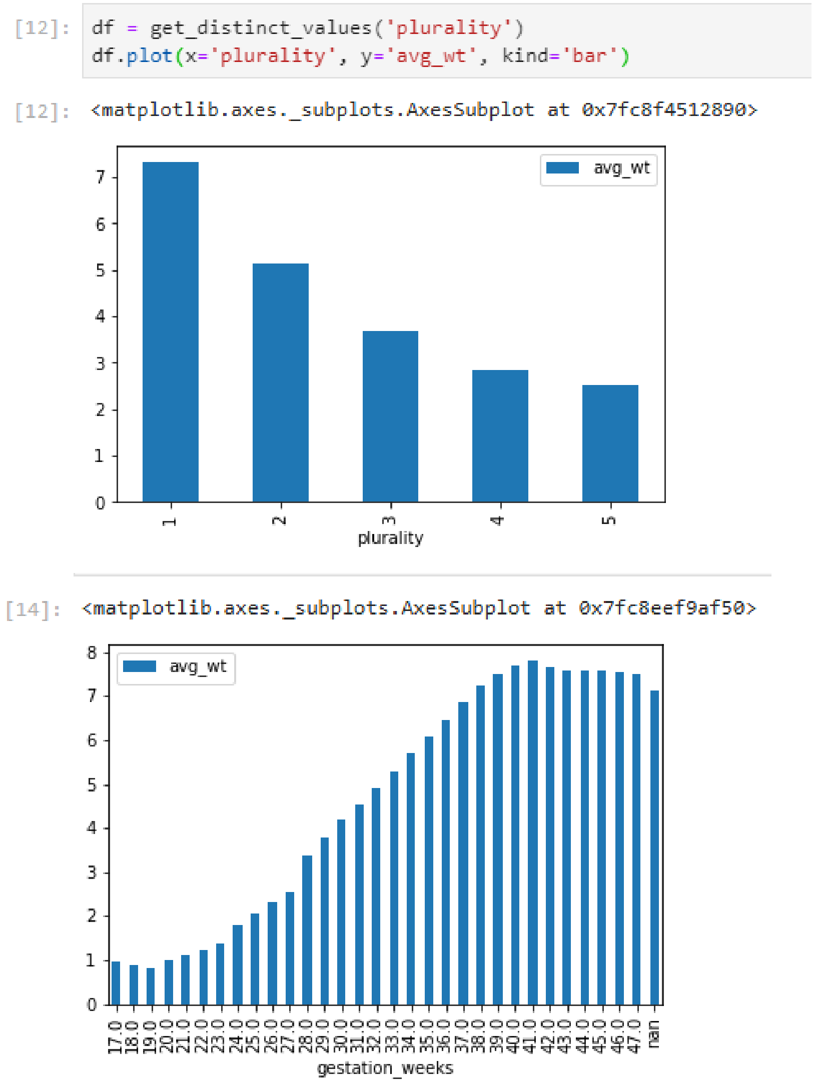
Row	name	count
1	Aarshi	5
2	Aaniylah	5
3	Aaryah	5
4	Aashirya	5
5	Aalimah	5
6	Aarielle	5
7	Aarabella	5
8	Aayra	5
9	Aarti	5
10	Aavya	5
11	Aashni	5
12	Aadrika	5
13	Aamyah	5
14	Aamilah	5
15	Abagael	5
16	Aayusha	5
17	Aarion	5
18	Aania	5
19	Aaiza	5
20	Aabriella	5

09.2g.3

- How much less data does this query process compared to the size of the table?
  - This one is 18gb less than the whole table, the original was around 21 gb and this query is about 3gb
- How many twins were born during this time range?
  - 375362
- How much lighter on average are they compared to single babies?
  - 1.9 lbs lighter

09.2g.6

Plurality and Gestation weeks are the most important indicators of new born weight



```
[12]: df = get_distinct_values('plurality')
      df.plot(x='plurality', y='avg_wt', kind='bar')
```

```
[12]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc8f4512890>
```

```
[14]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc8eef9af50>
```

09.2g.8

- What day saw the largest spike in trips to grocery and pharmacy stores?
  - 12-13 had the largest posititive spike, and 29th had the largest negative spike
- On the day the stay-at-home order took effect (3/23/2020), what was the total impact on workplace trips?
  - -49%

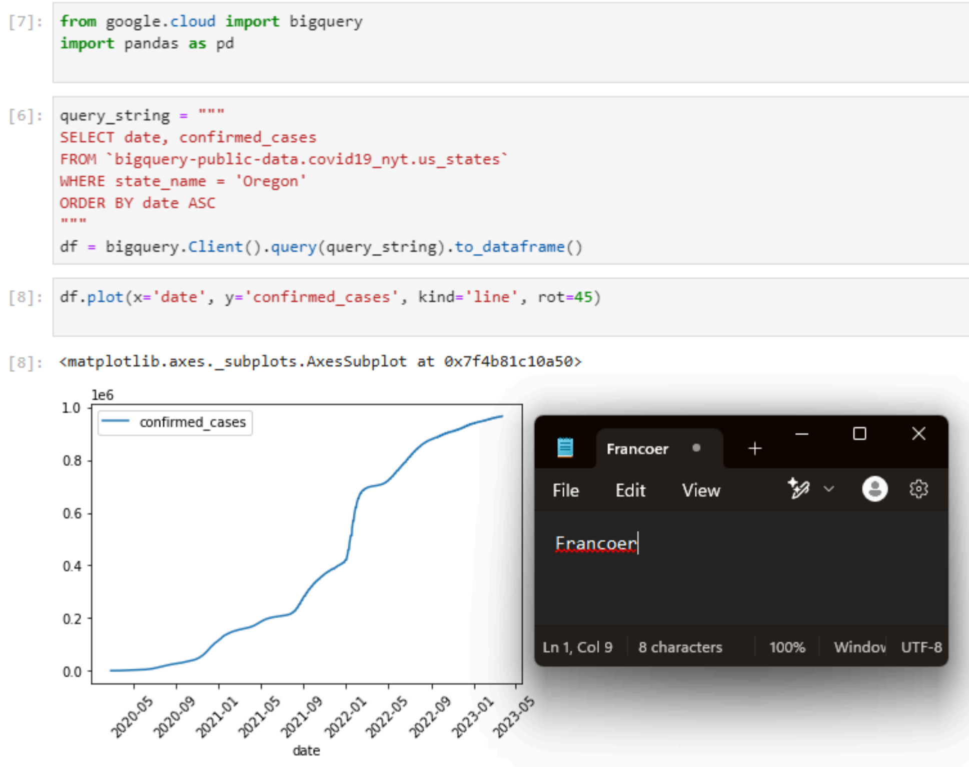
09.2g.9

- Which three airports were impacted the most in April 2020 (the month when lockdowns became widespread)?
  - Detroit Metropolitan Wayne County
  - McCarran International
  - San Francisco International
- Run the query again using the month of August 2020. Which three airports were impacted the most?
  - McCarran International
  - Detroit Metropolitan Wayne County
  - San Francisco International

09.2g.10

- What table and columns identify the place name, the starting date, and the number of excess deaths from COVID-19?
  - table : excess_deaths
    - start_date
    - excess_deaths
- What table and columns identify the date, county, and deaths from COVID-19?
  - Table: us_counties
    - date
    - deaths
- What table and columns identify the date, state, and confirmed cases of COVID-19?
  - Table: us_states
    - date
    - confirmed_cases
- What table and columns identify a county code and the percentage of its residents that report they always wear masks?
  - Table: mask_use_by_county
    - county_fips_code
    - always

09.2g.11



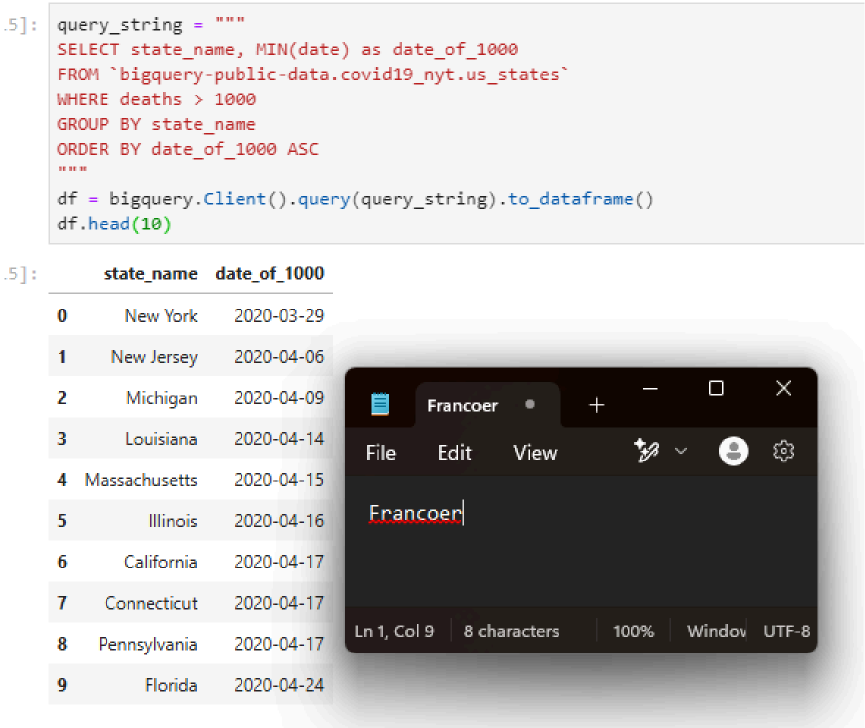
```
[7]: from google.cloud import bigquery
     import pandas as pd
```

```
[6]: query_string = """
     SELECT date, confirmed_cases
     FROM `bigquery-public-data.covid19_nyt.us_states`
     WHERE state_name = 'Oregon'
     ORDER BY date ASC
     """
     df = bigquery.Client().query(query_string).to_dataframe()
```

```
[8]: df.plot(x='date', y='confirmed_cases', kind='line', rot=45)
```

```
[8]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4b81c10a50>
```

Francoer

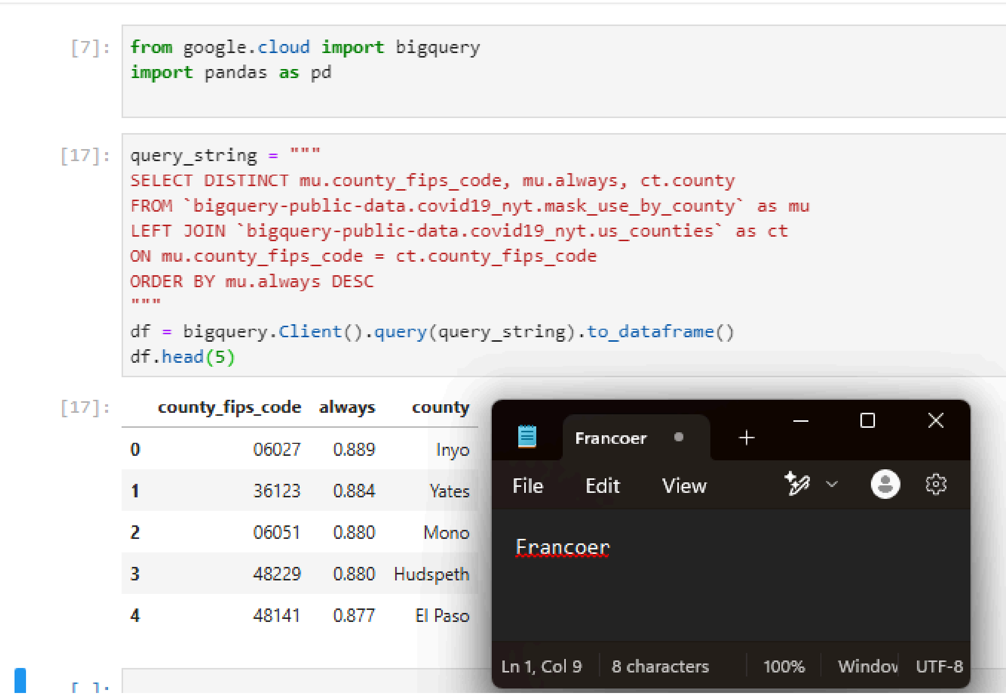


```
.5]: query_string = """
     SELECT state_name, MIN(date) as date_of_1000
     FROM `bigquery-public-data.covid19_nyt.us_states`
     WHERE deaths > 1000
     GROUP BY state_name
     ORDER BY date_of_1000 ASC
     """
     df = bigquery.Client().query(query_string).to_dataframe()
     df.head(10)
```

.5]:

	state_name	date_of_1000
0	New York	2020-03-29
1	New Jersey	2020-04-06
2	Michigan	2020-04-09
3	Louisiana	2020-04-14
4	Massachusetts	2020-04-15
5	Illinois	2020-04-16
6	California	2020-04-17
7	Connecticut	2020-04-17
8	Pennsylvania	2020-04-17
9	Florida	2020-04-24

Francoer



[7]:

```
from google.cloud import bigquery
import pandas as pd
```

[17]:

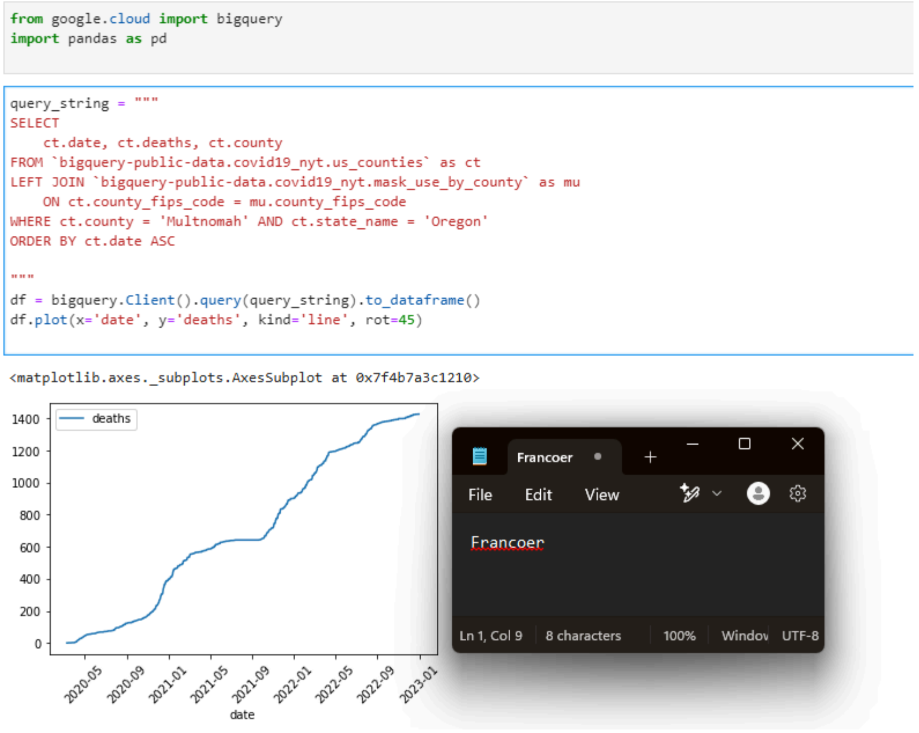
```
query_string = """
SELECT DISTINCT mu.county_fips_code, mu.always, ct.county
FROM `bigquery-public-data.covid19_nyt.mask_use_by_county` as mu
LEFT JOIN `bigquery-public-data.covid19_nyt.us_counties` as ct
ON mu.county_fips_code = ct.county_fips_code
ORDER BY mu.always DESC
"""
df = bigquery.Client().query(query_string).to_dataframe()
df.head(5)
```

[17]:

	county_fips_code	always	county
0	06027	0.889	Inyo
1	36123	0.884	Yates
2	06051	0.880	Mono
3	48229	0.880	Hudspeth
4	48141	0.877	El Paso

Francoer

09.2g.12

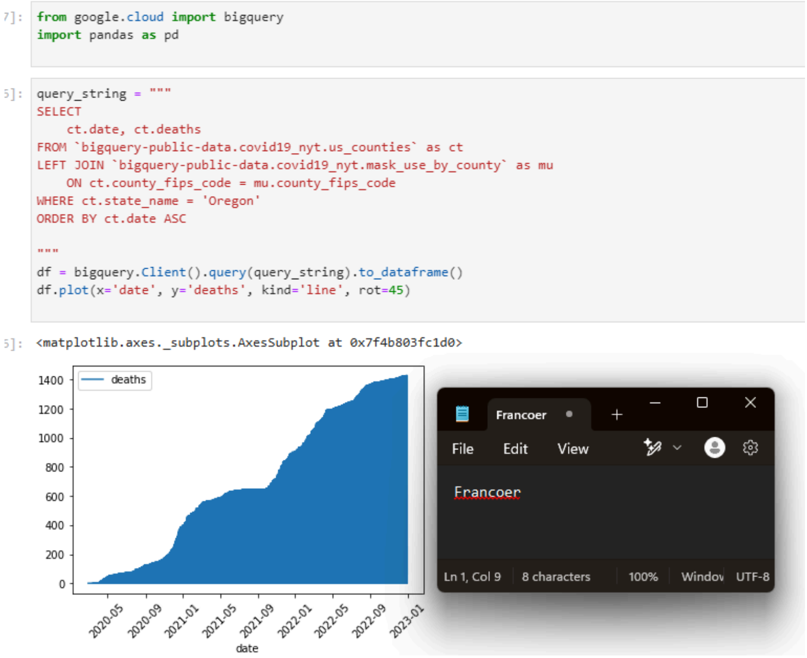


```
from google.cloud import bigquery
import pandas as pd
```

```
query_string = """
SELECT
    ct.date, ct.deaths, ct.county
FROM `bigquery-public-data.covid19_nyt.us_counties` as ct
LEFT JOIN `bigquery-public-data.covid19_nyt.mask_use_by_county` as mu
    ON ct.county_fips_code = mu.county_fips_code
WHERE ct.county = 'Multnomah' AND ct.state_name = 'Oregon'
ORDER BY ct.date ASC

"""
df = bigquery.Client().query(query_string).to_dataframe()
df.plot(x='date', y='deaths', kind='line', rot=45)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4b7a3c1210>
```

Francoer



```
7]: from google.cloud import bigquery
    import pandas as pd
```

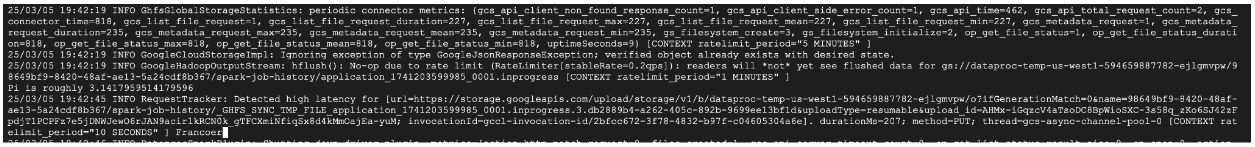
```
5]: query_string = """
    SELECT
        ct.date, ct.deaths
    FROM `bigquery-public-data.covid19_nyt.us_counties` as ct
    LEFT JOIN `bigquery-public-data.covid19_nyt.mask_use_by_county` as mu
        ON ct.county_fips_code = mu.county_fips_code
    WHERE ct.state_name = 'Oregon'
    ORDER BY ct.date ASC

    """
    df = bigquery.Client().query(query_string).to_dataframe()
    df.plot(x='date', y='deaths', kind='line', rot=45)
```

```
5]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4b803fc1d0>
```

Francoer

09.3g.6

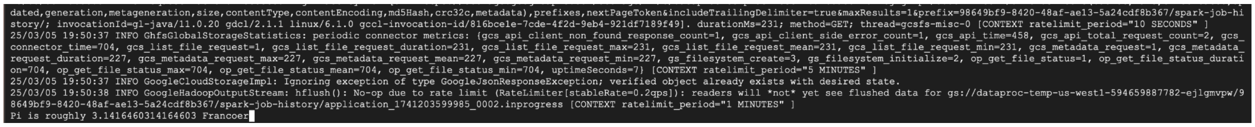
- **How long did the job take to execute?**
  - ○ **55 seconds**
- **Examine output.txt and show the estimate of π calculated**



```
25/03/05 19:42:19 INFO GhfsGlobalStorageStatistics: periodic connector metrics: {gcs_api_client_non_found_response_count=1, gcs_api_client_side_error_count=1, gcs_api_time=462, gcs_api_total_request_count=2, gcs_
connector_time=818, gcs_list_file_request=1, gcs_list_file_request_duration=227, gcs_list_file_request_max=227, gcs_list_file_request_mean=227, gcs_list_file_request_min=227, gcs_metadata_request=1, gcs_metadata_
request_duration=235, gcs_metadata_request_max=235, gcs_metadata_request_mean=235, gcs_metadata_request_min=235, gs_filesystem_create=3, gs_filesystem_initialize=2, op_get_file_status=1, op_get_file_status_durati
on=818, op_get_file_status_max=818, op_get_file_status_mean=818, op_get_file_status_min=818, uptimeSeconds=9} [CONTEXT ratelimit_period="5 MINUTES" ]
25/03/05 19:42:19 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
25/03/05 19:42:19 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for gs://dataproc-temp-us-west1-594659887782-ejlgmvpw/9
8649bf9-8420-48af-ae13-5a24cdf8b367/spark-job-history/application_1741203599985_0001.inprogress [CONTEXT ratelimit_period="1 MINUTES" ]
Pi is roughly 3.1417959514179596
25/03/05 19:42:45 INFO RequestTracker: Detected high latency for [url=https://storage.googleapis.com/upload/storage/v1/b/dataproc-temp-us-west1-594659887782-ejlgmvpw/o?ifGenerationMatch=0&name=98649bf9-8420-48af-
ae13-5a24cdf8b367/spark-job-history/_GHFS_SYNC_TMP_FILE_application_1741203599985_0001.inprogress.3.db2889b4-a262-405c-892b-9699ee13bf1d&uploadType=resumable&upload_id=AHMx-iGqzcV4aTsoDC8BpWioSXC-3s58q_zKo6SJ42zF
pdjT1PCPFz7e5jDNWJewO6rJAN9acirlkRCN0k_gTFCXmiNfiqSx8d4kMmOajEa-yuM; invocationId=gccl-invocation-id/2bfcc672-3f78-4832-b97f-c04605304a6e]. durationMs=207; method=PUT; thread=gcs-async-channel-pool-0 [CONTEXT rat
elimit_period="10 SECONDS" ] Francoer
```

09.3g.8

- **How long did the job take to execute? How much faster did it take?**
  - ○ **22 seconds, it was 33 seconds faster more than twice as fast**
- **Examine output2.txt and show the estimate of π calculated**



```
dated,generation,metageneration,size,contentType,contentEncoding,md5Hash,crc32c,metadata),prefixes,nextPageToken&includeTrailingDelimiter=true&maxResults=1&prefix=98649bf9-8420-48af-ae13-5a24cdf8b367/spark-job-hi
story/; invocationId=gl-java/11.0.20 gdcl/2.1.1 linux/6.1.0 gccl-invocation-id/816bce1e-7cde-4f2d-9eb4-921df7189f49]. durationMs=231; method=GET; thread=gcsfs-misc-0 [CONTEXT ratelimit_period="10 SECONDS" ]
25/03/05 19:50:37 INFO GhfsGlobalStorageStatistics: periodic connector metrics: {gcs_api_client_non_found_response_count=1, gcs_api_client_side_error_count=1, gcs_api_time=458, gcs_api_total_request_count=2, gcs_
connector_time=704, gcs_list_file_request=1, gcs_list_file_request_duration=231, gcs_list_file_request_max=231, gcs_list_file_request_mean=231, gcs_list_file_request_min=231, gcs_metadata_request=1, gcs_metadata_
request_duration=227, gcs_metadata_request_max=227, gcs_metadata_request_mean=227, gcs_metadata_request_min=227, gs_filesystem_create=3, gs_filesystem_initialize=2, op_get_file_status=1, op_get_file_status_durati
on=704, op_get_file_status_max=704, op_get_file_status_mean=704, op_get_file_status_min=704, uptimeSeconds=7} [CONTEXT ratelimit_period="5 MINUTES" ]
25/03/05 19:50:37 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
25/03/05 19:50:38 INFO GoogleHadoopOutputStream: hflush(): No-op due to rate limit (RateLimiter[stableRate=0.2qps]): readers will *not* yet see flushed data for gs://dataproc-temp-us-west1-594659887782-ejlgmvpw/9
8649bf9-8420-48af-ae13-5a24cdf8b367/spark-job-history/application_1741203599985_0002.inprogress [CONTEXT ratelimit_period="1 MINUTES" ]
Pi is roughly 3.1416460314164603 Francoer
```

09.4g.3

- Where is the input taken from by default?

  -
    ```
        input = '{0}*.java'.format(options.input)
    ```
- Where does the output go by default?

  -
    ```
      output_prefix = options.output_prefix
    ```
- Examine both the getPackages() function and the splitPackageName() function. What operation does the 'PackageUse()' transform implement?

  -
    ```
    | 'PackageUse' >> beam.FlatMap(lambda line: packageUse(line,
    keyword))
    ```
  - Which parses the file/library given into individual words
- Look up Beam's CombinePerKey. What operation does the TotalUse operation implement?

  -
    ```
    | 'TotalUse' >> beam.CombinePerKey(sum)
    ```

- Which operations correspond to a "Map"?

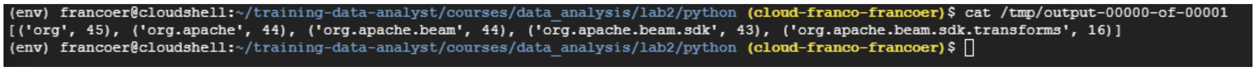
  -
    ```
          | 'GetImports' >> beam.FlatMap(lambda line:
    startsWith(line, keyword))
    ```

  -
    ```
          | 'PackageUse' >> beam.FlatMap(lambda line:
    packageUse(line, keyword))
    ```
- Which operation corresponds to a "Shuffle-Reduce"?

  -
    ```
          | 'TotalUse' >> beam.CombinePerKey(sum)
    ```
- Which operation corresponds to a "Reduce"?

  ```
  | 'TotalUse' >> beam.CombinePerKey(sum)
  | 'Top_5' >> beam.transforms.combiners.Top.Of(5, key=lambda kv:
  kv[1])
  ```

09.4g.4



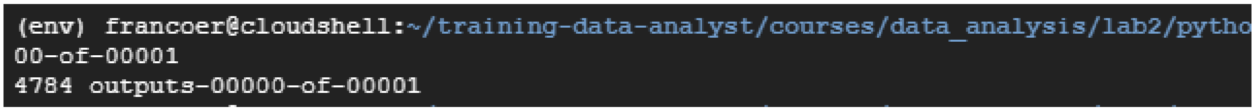
```
(env) francoer@cloudshell:~/training-data-analyst/courses/data_analysis/lab2/python (cloud-franco-francoer)$ cat /tmp/output-00000-of-00001
[('org', 45), ('org.apache', 44), ('org.apache.beam', 44), ('org.apache.beam.sdk', 43), ('org.apache.beam.sdk.transforms', 16)]
(env) francoer@cloudshell:~/training-data-analyst/courses/data_analysis/lab2/python (cloud-franco-francoer)$ 
```

- Explain what the data in this output file corresponds to based on your understanding of the program.
  - **It found the package given and parsed it into separate parts and counts the frequency of the files**
  - **Org.apache.beam.sdk.transforms**

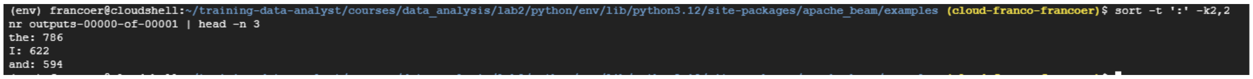
09.4g.5

- What are the names of the stages in the pipeline?
  - Read
  - Split
  - PairWithOne
  - GroupAndSum
  - Format
  - Write
- Describe what each stage does.
  - `Read the text file[pattern] into a PCollection`
  - Split makes each text into its individual words
  - Pairwithone will take each word and associate it with a value
  - groupandSum will group the values with the words and their count
  - `Format the counts into a PCollection of strings.`
  - `Write the output using a "Write" transform that has side effects.`

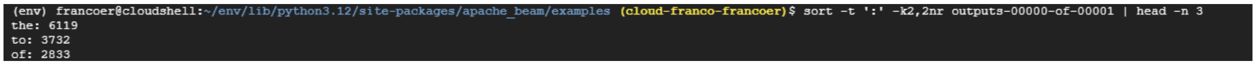
09.4g.6



```
(env) francoer@cloudshell:~/training-data-analyst/courses/data_analysis/lab2/pytho
00-of-00001
4784 outputs-00000-of-00001
```



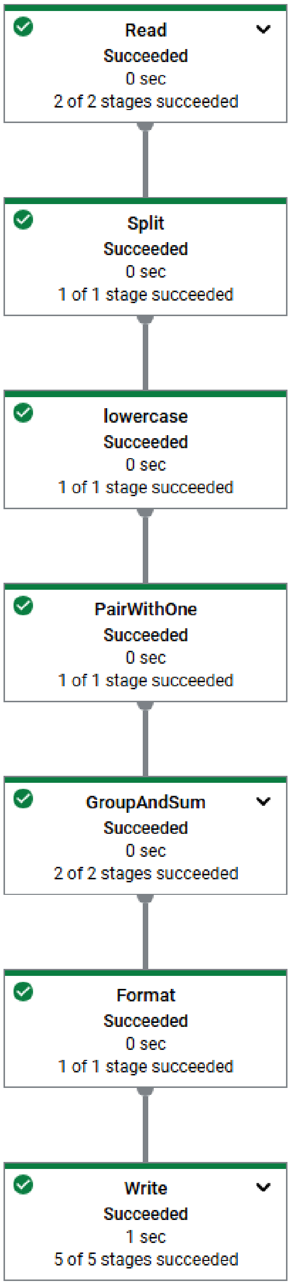
```
(env) francoer@cloudshell:~/training-data-analyst/courses/data_analysis/lab2/python/env/lib/python3.12/site-packages/apache_beam/examples (cloud-franco-francoer)$ sort -t ':' -k2,2
nr outputs-00000-of-00001 | head -n 3
the: 786
I: 622
and: 594
```



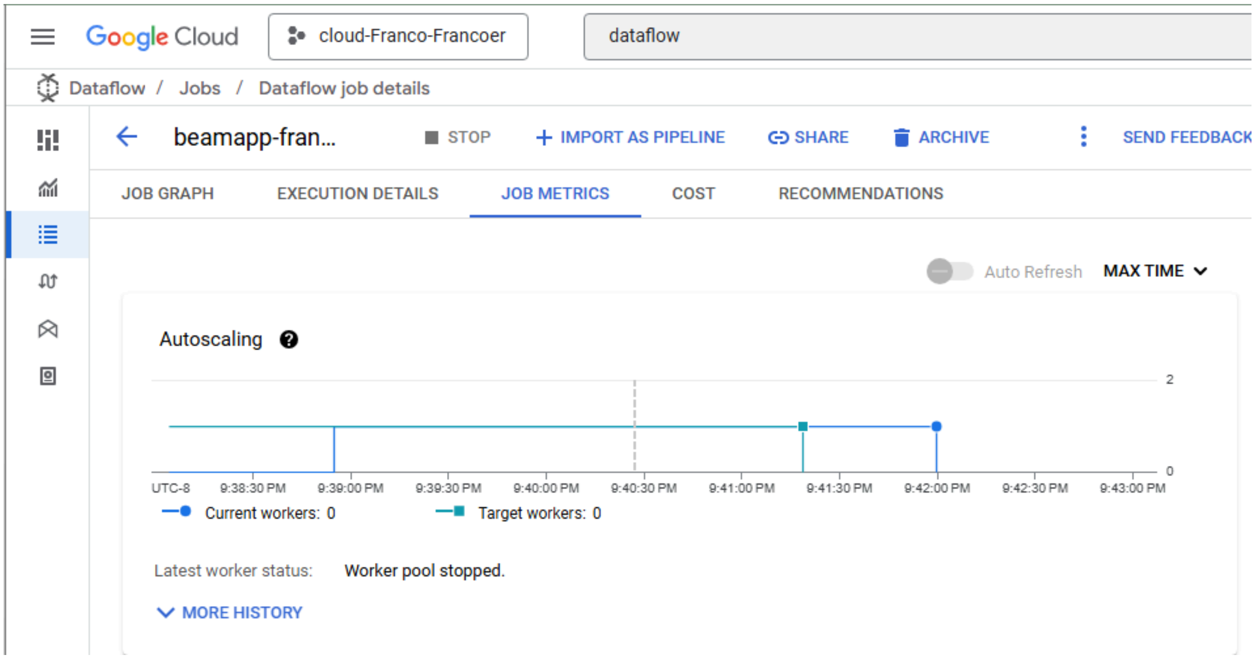
```
(env) francoer@cloudshell:~/env/lib/python3.12/site-packages/apache_beam/examples (cloud-franco-francoer)$ sort -t ':' -k2,2nr outputs-00000-of-00001 | head -n 3
the: 6119
to: 3732
of: 2833
```

09.4g.9

- The part of the job graph that has taken the longest time to complete.
    - The writing

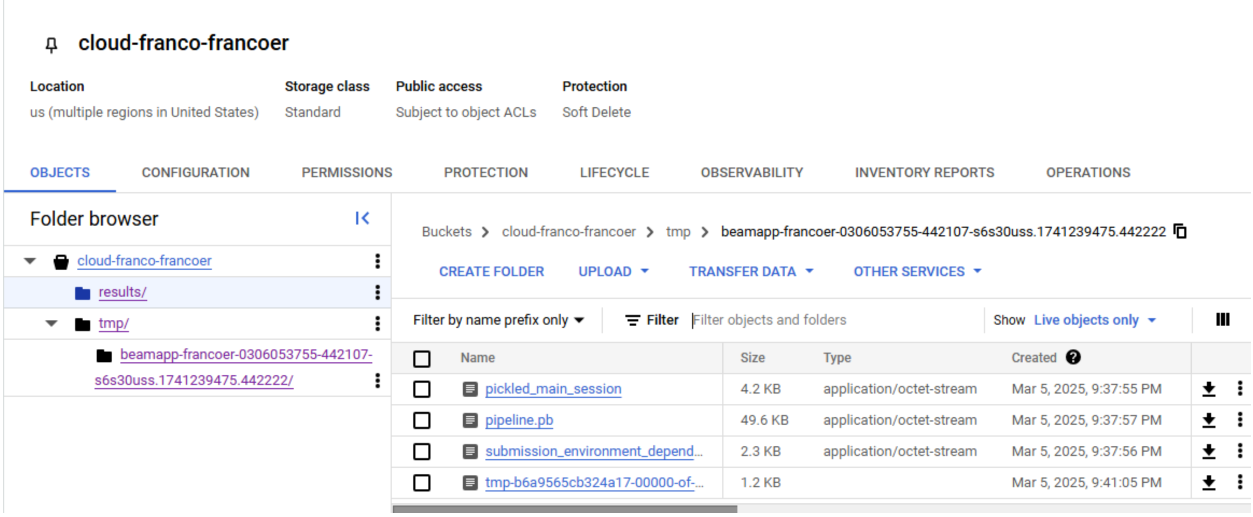


- The autoscaling graph showing when the worker was created and stopped.



Latest worker status: Worker pool stopped.

- Examine the output directory in Cloud Storage. How many files has the final write stage in the pipeline created?
  5 if we include the output file in the results directory



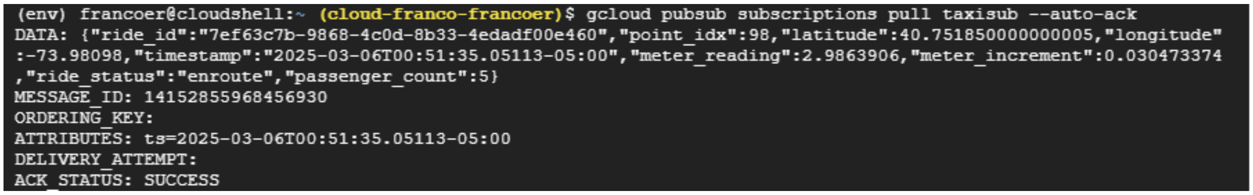
cloud-franco-francoer

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Subject to object ACLs	Soft Delete

Buckets > cloud-franco-francoer > tmp > beamapp-francoer-0306053755-442107-s6s30uss.1741239475.442222

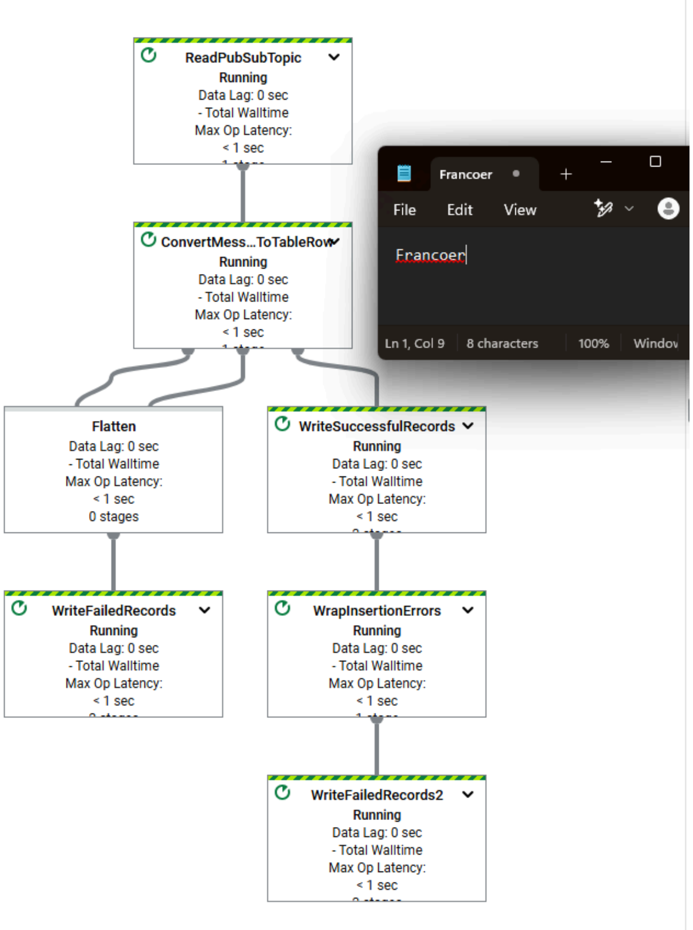
☐	Name	Size	Type	Created
☐	pickled_main_session	4.2 KB	application/octet-stream	Mar 5, 2025, 9:37:55 PM
☐	pipeline.pb	49.6 KB	application/octet-stream	Mar 5, 2025, 9:37:57 PM
☐	submission_environment_depend...	2.3 KB	application/octet-stream	Mar 5, 2025, 9:37:56 PM
☐	tmp-b6a9565cb324a17-00000-of-...	1.2 KB		Mar 5, 2025, 9:41:05 PM

09.4g.12



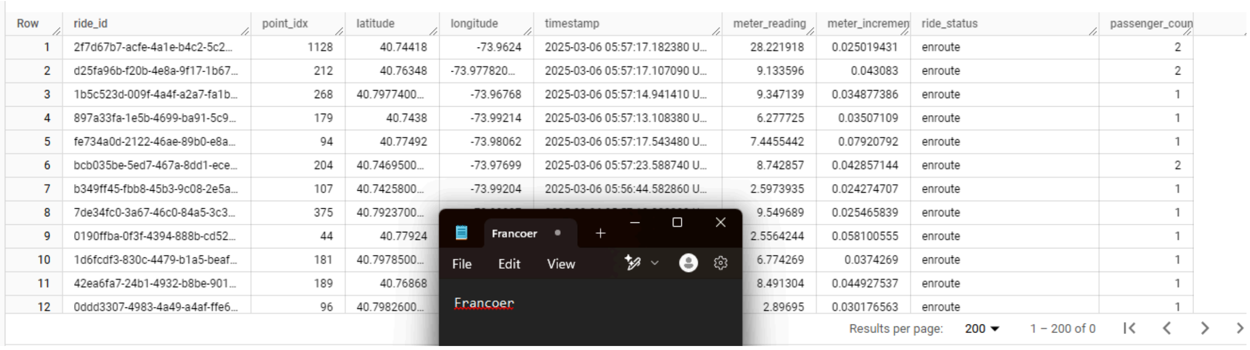
```
(env) francoer@cloudshell:~ (cloud-franco-francoer)$ gcloud pubsub subscriptions pull taxisub --auto-ack
DATA: {"ride_id":"7ef63c7b-9868-4c0d-8b33-4edadf00e460","point_idx":98,"latitude":40.751850000000005,"longitude"
:-73.98098,"timestamp":"2025-03-06T00:51:35.05113-05:00","meter_reading":2.9863906,"meter_increment":0.030473374
,"ride_status":"enroute","passenger_count":5}
MESSAGE_ID: 14152855968456930
ORDERING_KEY:
ATTRIBUTES: ts=2025-03-06T00:51:35.05113-05:00
DELIVERY_ATTEMPT:
ACK_STATUS: SUCCESS
```

09.4g.14



Francoer

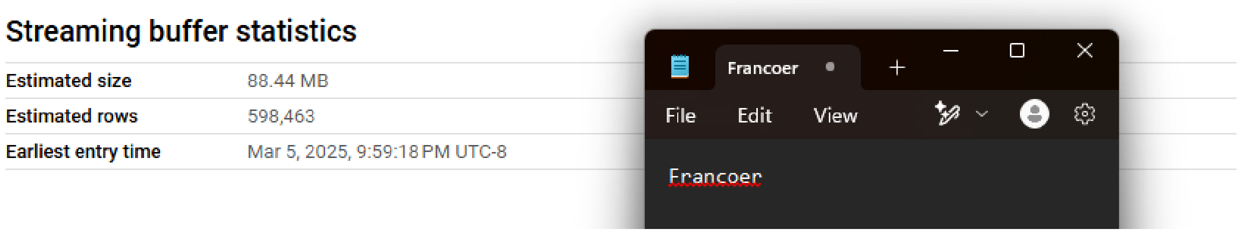
09.4g.15



Row	ride_id	point_idx	latitude	longitude	timestamp	meter_reading	meter_incremen	ride_status	passenger_coun
1	2f7d67b7-acfe-4a1e-b4c2-5c2...	1128	40.74418	-73.9624	2025-03-06 05:57:17.182380 U...	28.221918	0.025019431	enroute	2
2	d25fa96b-f20b-4e8a-9f17-1b67...	212	40.76348	-73.977820...	2025-03-06 05:57:17.107090 U...	9.133596	0.043083	enroute	2
3	1b5c523d-009f-4a4f-a2a7-fa1b...	268	40.7977400...	-73.96768	2025-03-06 05:57:14.941410 U...	9.347139	0.034877386	enroute	1
4	897a33fa-1e5b-4699-ba91-5c9...	179	40.7438	-73.99214	2025-03-06 05:57:13.108380 U...	6.277725	0.03507109	enroute	1
5	fe734a0d-2122-46ae-89b0-e8a...	94	40.77492	-73.98062	2025-03-06 05:57:17.543480 U...	7.4455442	0.07920792	enroute	1
6	bcb035be-5ed7-467a-8dd1-ece...	204	40.7469500...	-73.97699	2025-03-06 05:57:23.588740 U...	8.742857	0.042857144	enroute	2
7	b349ff45-fbb8-45b3-9c08-2e5a...	107	40.7425800...	-73.99204	2025-03-06 05:56:44.582860 U...	2.5973935	0.024274707	enroute	1
8	7de34fc0-3a67-46c0-84a5-3c3...	375	40.7923700...			9.549689	0.025465839	enroute	1
9	0190ffba-0f3f-4394-888b-cd52...	44	40.77924			2.5564244	0.058100555	enroute	1
10	1d6fcdf3-830c-4479-b1a5-beaf...	181	40.7978500...			6.774269	0.0374269	enroute	1
11	42ea6fa7-24b1-4932-b8be-901...	189	40.76868			8.491304	0.044927537	enroute	1
12	0ddd3307-4983-4a49-a4af-ffe6...	96	40.7982600...			2.89695	0.030176563	enroute	1

Francoer

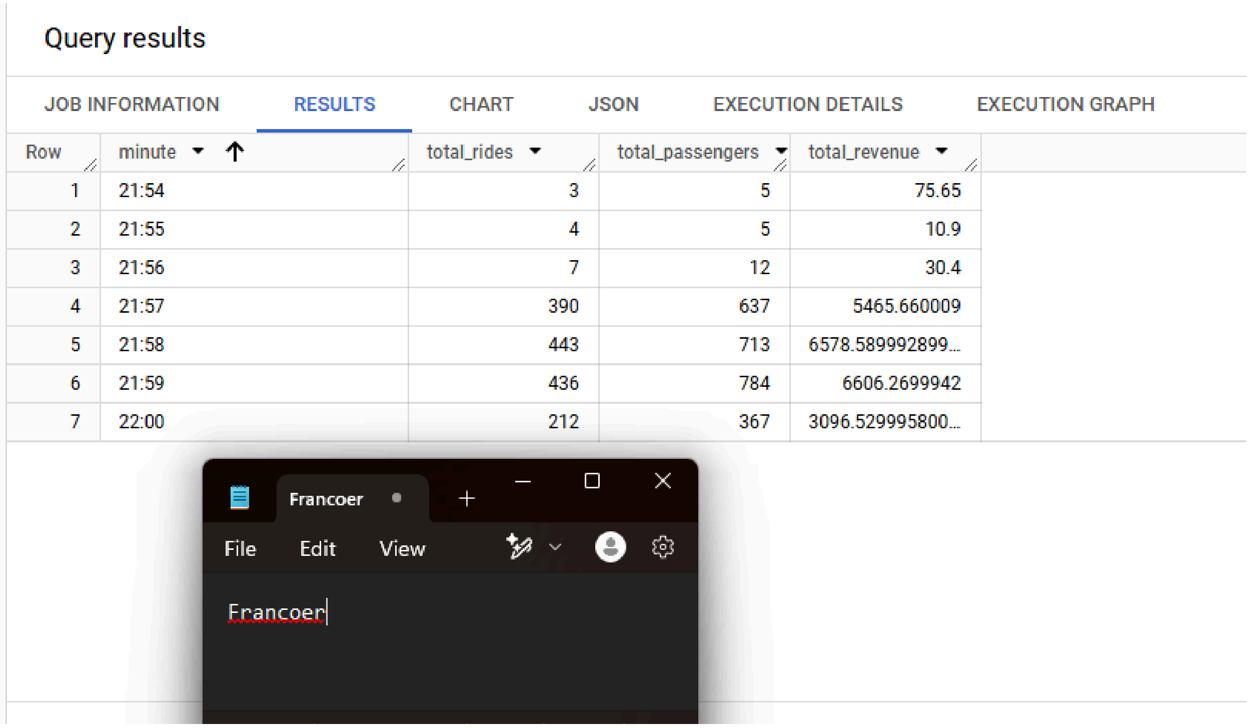
Results per page: 200 1 – 200 of 0



Streaming buffer statistics

Estimated size	88.44 MB
Estimated rows	598,463
Earliest entry time	Mar 5, 2025, 9:59:18 PM UTC-8

Francoer

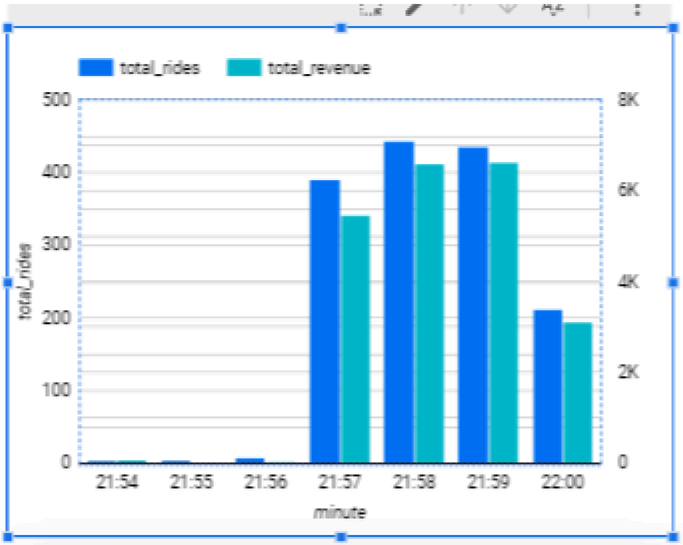


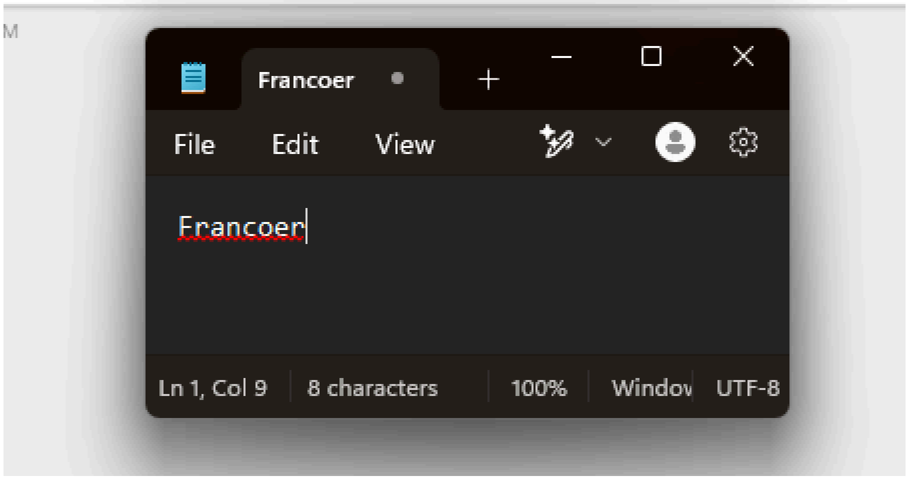
Query results

Row	minute	total_rides	total_passengers	total_revenue
1	21:54	3	5	75.65
2	21:55	4	5	10.9
3	21:56	7	12	30.4
4	21:57	390	637	5465.660009
5	21:58	443	713	6578.589992899...
6	21:59	436	784	6606.2699942
7	22:00	212	367	3096.529995800...

Francoer

09.4g.16





Francoer