Lab Week 9 — marvleon

9.1g: BigQuery, BigLake

9.1.3 Create dataset

yob_native_table QUERY SHARE

SCHEMA DETAILS PREVIEW LINEAGE DA

Table info

Table ID	cloud-leon-marvleon.yob.yob_native_table
Created	Nov 27, 2023, 12:29:26 PM UTC-8
Last modified	Nov 27, 2023, 12:29:26 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)	

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
Total physical bytes	0 B
Active physical bytes	0 B
Long term physical bytes	0 B
Time travel physical bytes	0 B

9.1.4 Query Data

Untitled

RUN SAVE DOWNLOAD SHARE

```
SELECT name, count
FROM `cloud-leon-marvleon.yob.yob_native_table`
WHERE gender='F'
ORDER BY count DESC
LIMIT 20
```

Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON

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

```
marvleon@cloudshell:~/code/9.1 (cloud-leon-marvleon)$ bq query "SELECT
 name, count FROM [cloud-leon-marvleon.yob.yob_native_table] WHERE gen
der='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 |
+---------+-------+
```

```
marvleon@cloudshell:~/code/9.1 (cloud-leon-marvleon)$ bq shell
Welcome to BigQuery! (Type help for more information.)
Cannot read termcap database;
using dumb terminal settings.
cloud-leon-marvleon> SELECT name, count FROM [cloud-leon-marvleon.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-leon-marvleon>
```

Not super popular my name!

```
cloud-leon-marvleon> SELECT count FROM [cloud-leon-marvleon.yob.yob_na
tive_table] WHERE name='Marvin'
+-------+
| count |
+-------+
|   567 |
+-------+
cloud-leon-marvleon>
```

9.1.9 Query Data

us-west1.biglake | *Untitled 3

Untitled 3 RUN SAVE DOWNLOAD

```
SELECT name, count
FROM `cloud-leon-marvleon.yob.yob_biglake_table`
WHERE gender='F'
ORDER BY count ASC
LIMIT 20
```

Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON

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

9.2g Jupyter Notebooks

9.2.3 BigQuery query

How much less data does this query process compared to the size of the table?
Was: 21.94gb
Now: 3.05gb
Difference: 18.89gb less
How many twins were born during this time range?
375,362
How much lighter on average are they compared to single babies?
2.17lbs

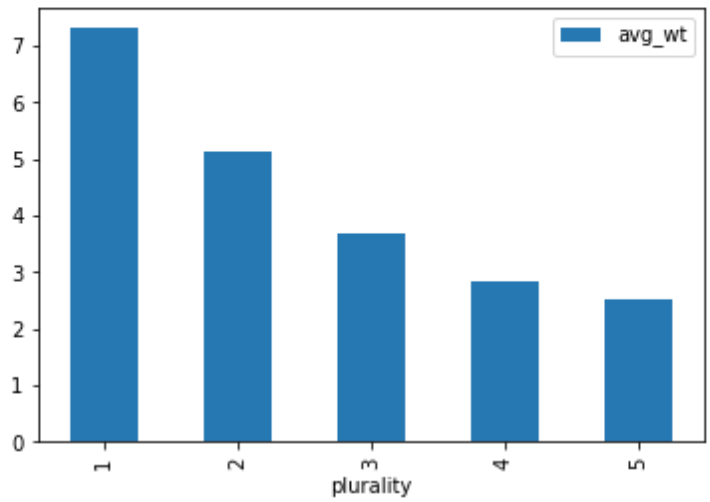
9.2.6 Run queries

Which two features are the strongest predictors for a newborn baby's weight?

Plurality and Gestation time

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

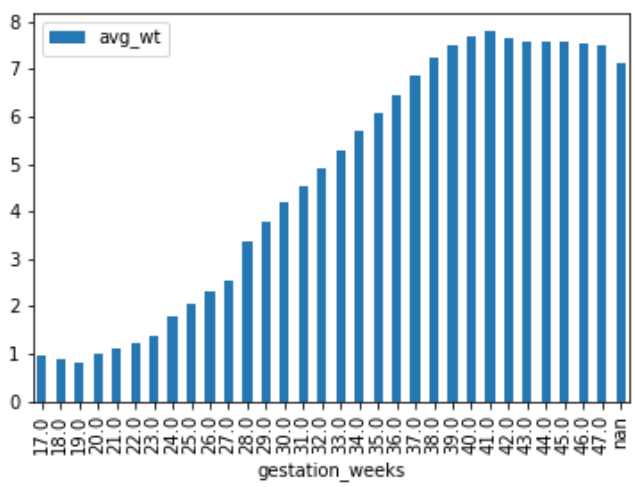
```
[10]: <matplotlib.axes._subplots.AxesSubplot at 0x7f90bcb3ded0>
```



marvleon

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

```
[12]: <matplotlib.axes._subplots.AxesSubplot at 0x7f90bc5176d0>
```



marvleon

9.2.8 Mobility

What day saw the largest spike in trips to grocery and pharmacy stores?

03-13-2020

On the day the stay-at-home order took effect, what was the total impact on workplace trips?

-49% from the baseline

9.2.9 Airport traffic

Which three airports were impacted the most in April 2020?

Newark Liberty International
Daniel K. Inouye International
Chicago O'Hare International

Which three airports were impacted the most in August 2020?

Newark Liberty International
Charlotte Douglas International
Dallas/Fort Worth International

9.2.10 Mortality

What table and columns identify the place name, the starting date, and the number of excess deaths from COVID-19?

table: excess_deaths
column: placename
column: start_date
column: excess_death

What table and columns identify the date, county, and deaths from COVID-19?

table: us_counties
column: date
column: county
column: deaths

What table and columns identify the date, state, and confirmed cases of COVID-19?

table: us_states
column: date
column: state_name
column: confirmed_cases

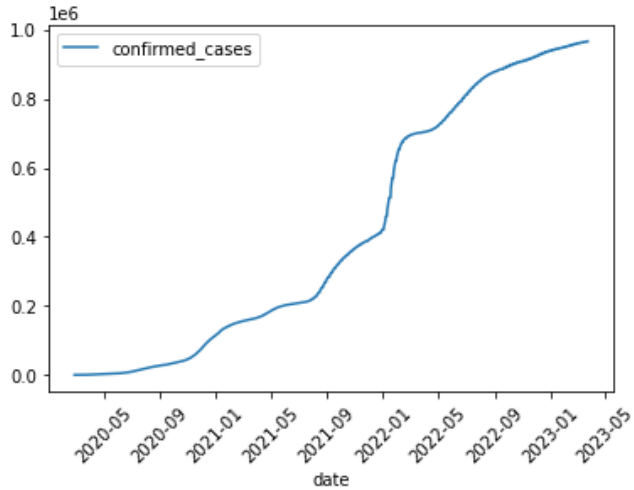
What table and columns identify a county code and the percentage of its residents that report they always wear masks?

table: mask_use_by_county
column: county_fibs_code
column: always

9.2.11 Run example queries

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

```
[4]: <matplotlib.axes._subplots.AxesSubplot at 0x7f9d2690a9d0>
```



marvleon

```
[6]: 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
"""
```

```
[7]: df = bigquery.Client().query(query_string).to_dataframe()
df.head(10)
```

[7]:

	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

marvleon

```
df = bigquery.Client().query(query_string).to_dataframe()
df.head(5)
```

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

marvleon

9.2.12 Write queries

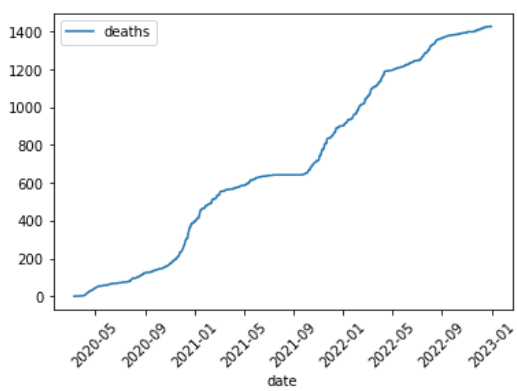
[16]:
```
query_string = """
SELECT date, deaths
FROM `bigquery-public-data.covid19_nyt.us_counties`
WHERE county = 'Multnomah' AND state_name = 'Oregon'
ORDER BY date ASC
"""
df = bigquery.Client().query(query_string).to_dataframe()
df.head()
```

[16]:

	date	deaths
0	2020-03-10	0
1	2020-03-11	0
2	2020-03-12	0
3	2020-03-13	0
4	2020-03-14	1

[17]:
```
df.plot(x='date', y='deaths', kind='line', rot=45)
```

[17]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9d256f5f10>
```



marvleon

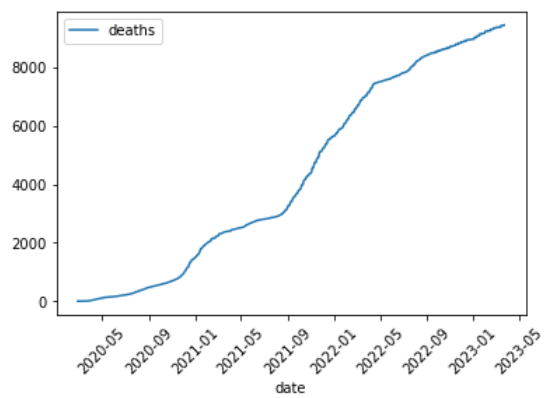
```
[22]: query_string = """
SELECT date, deaths
FROM `bigquery-public-data.covid19_nyt.us_states`
WHERE state_name = 'Oregon'
ORDER BY date ASC
"""
df = bigquery.Client().query(query_string).to_dataframe()
df.head()
```

[22]:

	date	deaths
0	2020-02-28	0
1	2020-02-29	0
2	2020-03-01	0
3	2020-03-02	0
4	2020-03-03	0

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

```
[23]: <matplotlib.axes._subplots.AxesSubplot at 0x7f9d256405d0>
```



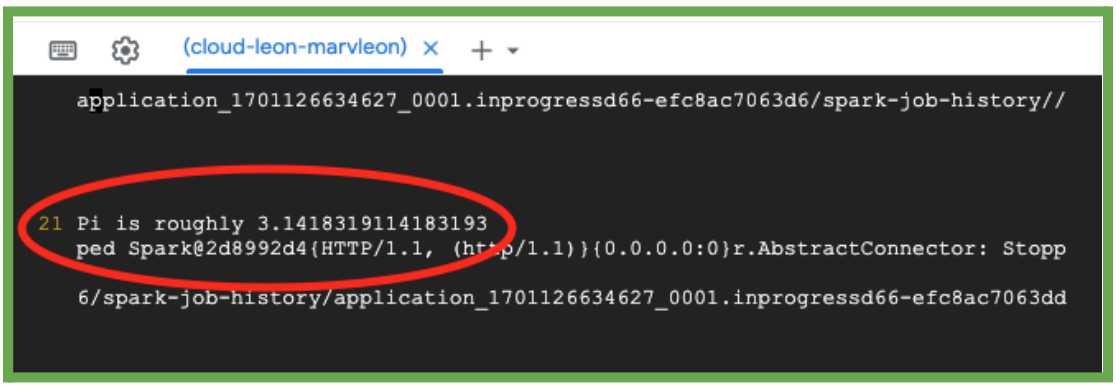
marvleon

9.3g Dataproc

9.3.6 Run computation

How long did the job take to execute?

1 minute 25 seconds



9.3.8 Run computation

How long did the job take to execute? How much faster did it take?
15 seconds, it was 00:01:19 faster

```
e8943b6ba310247ccb0ba14/driveroutput
ected potential high latency for operation op_creat
Pi is roughly 3.141588191415882
23/11/27 23:55:49 INFO com.google.cloud.hadoop.fs.c
```

marvleon

9.4g Dataflow

9.4.3 Beam code

Answer the following questions for your lab notebook.

Where is the input taken from by default?
../javahelp/src/main/java/com/google/cloud/training/dataanalyst/javahelp/
Where does the output go by default?
/tmp/output
What operation does the 'PackageUse()' transform implement?
mapping operation
What operation does the TotalUse operation implement?
summation operation
Which operations correspond to a "Map"?
GetImports and PackageUse
Which operation corresponds to a "Shuffle-Reduce"?
TotalUse
Which operation corresponds to a "Reduce"?
Top_5

9.4.4 Run pipeline locally

```
[('org', 45), ('org.apache', 44), ('org.apache.beam', 44), ('org.apache.beam.sdk', 43), ('org.apache.beam.
sdk.transforms', 16)]
(env) marvleon@cloudshell:~/code/training-data-analyst/courses/machine_learning/deepdive/04_features/dataf
```

This output corresponds to a Java package and the number of times it was used or imported across all analyzed Java files. ‘org’ was imported 45 times, ‘org.apache.beam’ was imported 44 times, etc. Also, the counts show many times these specific parts of the package hierarchy were referenced (‘org.apache.beam’ is more specific than just ‘org.apache’).

9.4.5 Dataflow Lab #2 (Word count)

What are the names of the stages in the pipeline? Describe what each stage does.

Read stage, Split stage, PairWIthOne stage, GroupAndSUm stage, Format stage, Write stage. Read Stage: It uses the ReadFromText function to read the input text file into a PCollection. This collection will contain lines of text from the file specified in the input argument. Split Stage: This stage processes the lines of text to extract words. It uses a custom DoFn (Dataflow Operation Function) called WordExtractingDoFn. This function applies a regular expression to each line to find all word-like sequences, which are then returned as an iterator. PairWithOne Stage: In the 'PairWithOne' stage, each word from the previous stage is mapped (word, 1) using the *beam.Map* transform. This stage prepares each word to be counted by associating it with the number 1. GroupAndSum Stage: This stage combines all the tuples with the same word (as the key) and sums their associated values. This is achieved using the *beam.CombinePerKey(sum)* transform, which effectively counts the occurrences of each word. Format Stage: The 'Format' stage formats the word counts into a readable string format. This is done by mapping each key-value pair (word, count) to a string using *beam.MapTuple* along with a custom formatting function, which outputs each word followed by its count. Write Stage:Writes the output of the pipeline to a specified output file. This is done using the WriteToText transform. The output file location is specified by the output argument.

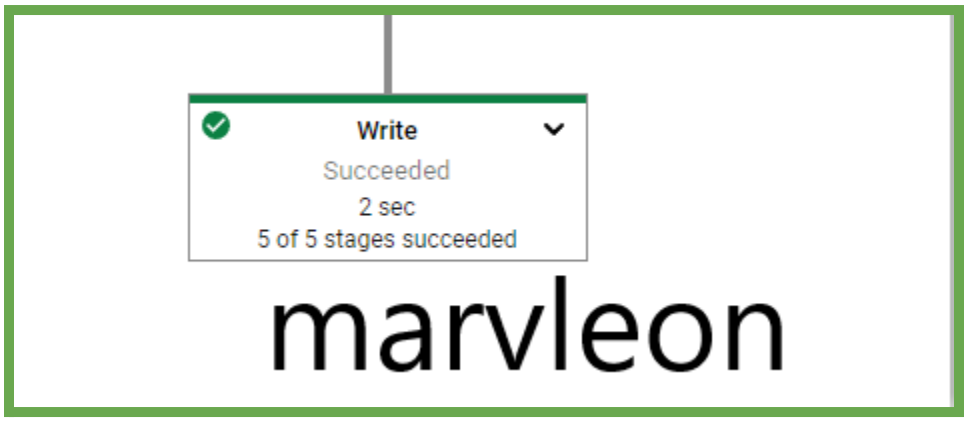
9.4.6 Run code locally

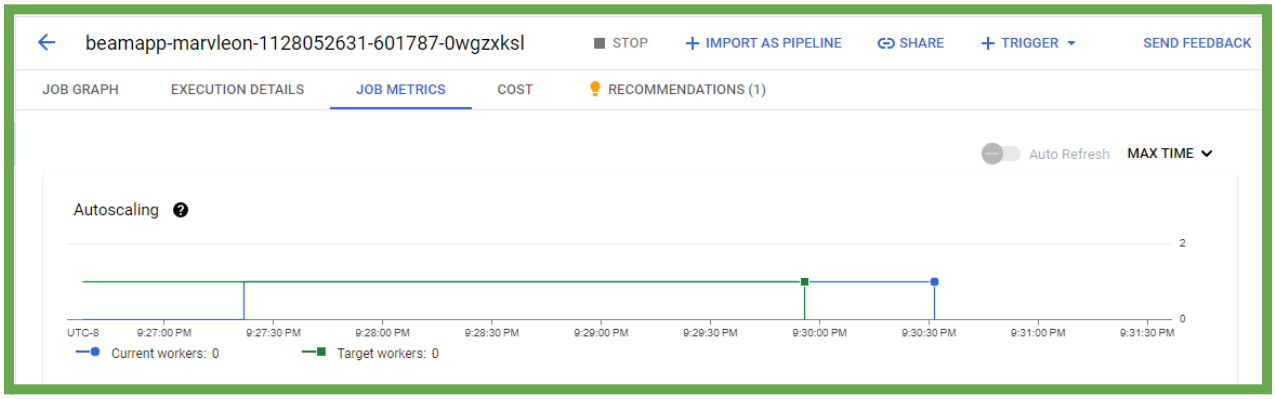
```
(env) marvleon@cloudshell:~/code/training-data-analyst/courses/machine_learning/d
eepdive/04_features/dataflow/python (cloud-leon-marvleon)$ wc -l outputs-00000-of
-00001
4784 outputs-00000-of-00001
```

```
(env) marvleon@cloudshell:~/code/training-data-analyst/courses/machine_learning/deepdive/04
_features/dataflow/python (cloud-leon-marvleon)$ sort -k2,2nr outputs-00000-of-00001 | head
 -n 3
the: 786
I: 622
and: 594
```

```
(env) marvleon@cloudshell:~/code/training-data-analyst/courses/machine_learning/deepdive/04
_features/dataflow/python (cloud-leon-marvleon)$ sort -k2,2nr outputs-00000-of-00001 | head
 -n 3
the: 908
and: 738
i: 622
```

9.4.9 Run code using Dataflow runner





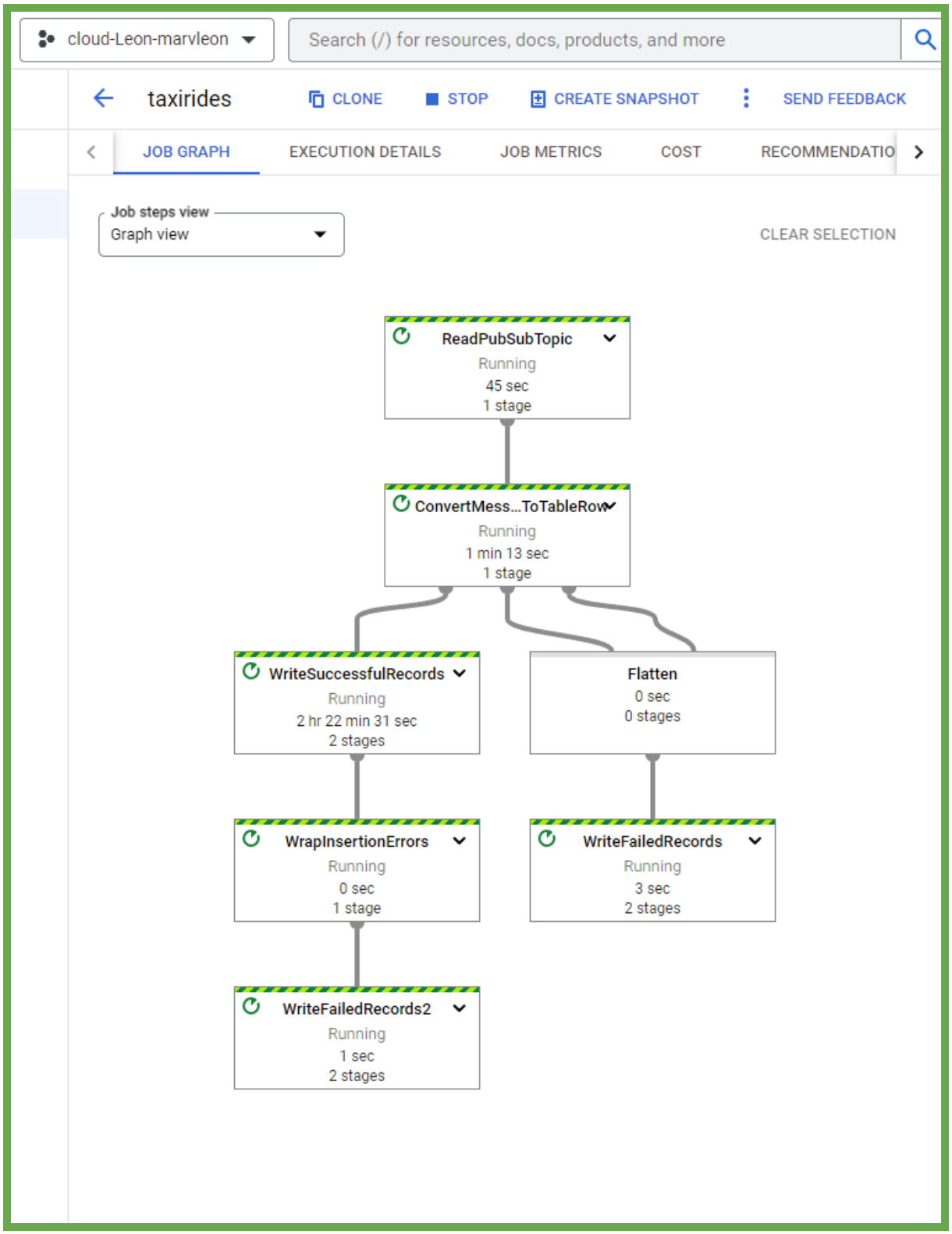
How many files has the final write stage in the pipeline created?

3 files

9.4.12 View raw data from PubSub

```
marvleon@cloudshell:~ (cloud-leon-marvleon)$ gcloud pubsub subscriptions pull taxisub --auto-ac
k
DATA: {"ride_id":"ce0efb94-4543-4b95-a50c-0e67456d46fe","point_idx":2274,"latitude":40.7374,"lo
ngitude":-73.93331,"timestamp":"2023-11-28T00:39:35.86264-05:00","meter_reading":46.316616,"met
er_increment":0.020367905,"ride_status":"enroute","passenger_count":1}
MESSAGE_ID: 9720117840575201
ORDERING_KEY:
ATTRIBUTES: ts=2023-11-28T00:39:35.86264-05:00
DELIVERY_ATTEMPT:
ACK_STATUS: SUCCESS
```

9.4.14 Run Dataflow job from template



9.4.15 Query data in BigQuery

Row	ride_id	point_idx	latitude	longitude	timestamp	meter_reading	meter_incremen	ride_status	passenger_coun
1	be09344e-96e4-4979-8543-e83...	38	40.7473	-73.98779	2023-11-28 05:43:18.880920 U...	1.9429865	0.051131222	enroute	1

realtime

cloud-leon-marvleon.taxirides

Last modified	Nov 27, 2023, 9:42:17 PM UTC-8
Data location	US
Description	
Labels	
Table type	table

bytes	
Time travel physical bytes	0 B

Streaming buffer statistics

Estimated size	353.51 MB
Estimated rows	2,233,857
Earliest entry time	Nov 27, 2023, 9:45:17 PM UTC-8

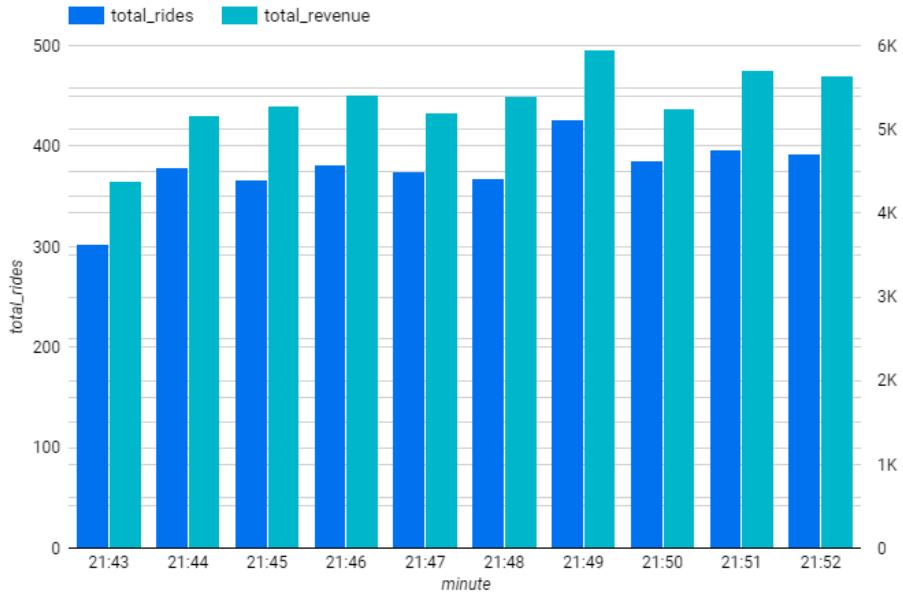
Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	minute	total_rides	total_passengers	total_revenue
1	21:43	303	525	4378.1999925
2	21:44	379	623	5171.0100014
3	21:45	367	599	5272.9400039
4	21:46	382	622	5413.129998800...
5	21:47	374	633	5198.6899873
6	21:48	368	591	5399.130007000...
7	21:49	427	713	5945.060004200...
8	21:50	386	654	5245.259992900...
9	21:51	396	634	5710.519981200...
10	21:52	393	648	5644.280006900...
11	21:53	394	626	5654.839994200...
12	21:54	412	689	5932.469984300...
13	21:55	390	699	5433.239988900...
14	21:56	429	670	5658.1600028
15	21:57	253	411	4064.809983500...

marvleon

9.4.16 Data visualization



marvleon