

# CS 530 INTERNET WEB AND CLOUD SYSTEMS

Name : Varsha Karinje

PSU ID: 925923534

<b>BigQuery, Notebooks Lab #1 (Ingesting data)</b>	<b>3</b>
<b>Examine dataset</b>	<b>3</b>
<b>Create dataset</b>	<b>4</b>
<b>Query data</b>	<b>5</b>
<b>BigQuery, Notebooks Lab #2 (Natality)</b>	<b>7</b>
<b>Launch notebook</b>	<b>8</b>
<b>BigQuery query</b>	<b>8</b>
<b>Jupyter notebook query</b>	<b>9</b>
<b>Exploring the dataset</b>	<b>10</b>
<b>Run queries</b>	<b>10</b>
<b>BigQuery, Notebooks Lab #3 (COVID-19 Mobility)</b>	<b>14</b>
<b>BigQuery, Notebooks Lab #4 (COVID-19 NYT)</b>	<b>17</b>
<b>Run example queries</b>	<b>19</b>
<b>Write queries</b>	<b>21</b>
<b>Clean up</b>	<b>24</b>
<b>Dataproc setup</b>	<b>25</b>
<b>Create Compute Engine cluster</b>	<b>25</b>
<b>Run computation</b>	<b>26</b>
<b>Scale cluster</b>	<b>27</b>
<b>Run computation again</b>	<b>30</b>
<b>Clean up</b>	<b>32</b>
<b>Setup</b>	<b>32</b>
<b>Beam code</b>	<b>33</b>

<b>Run pipeline locally</b>	<b>35</b>
<b>14. Dataflow Lab #2 (Word count)</b>	<b>35</b>
<b>Run code locally</b>	<b>36</b>
<b>Setup for Cloud Dataflow</b>	<b>39</b>
<b>Service account setup</b>	<b>39</b>
<b>Run code using Dataflow runner</b>	<b>39</b>
<b>Clean up</b>	<b>41</b>

## 09.2g: BigQuery, JupyterLab

### BigQuery, Notebooks Lab #1 (Ingesting data)

```

Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to cloud-cs-530-karinje-vkarinje.
Use "gcloud config set project [PROJECT_ID]" to change to a different project.
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud services list --available --filter=bigquery
NAME: alphafold-db.endpoints.bigquery-public-data.cloud.goog
TITLE: AlphaFold Protein Structure Database

NAME: analysis-ready-cloud-optimised-arco-ara5.endpoints.bigquery-public-data.cloud.goog
TITLE: Analysis-Ready, Cloud Optimised (ARCO) ERA5

NAME: annotation-bigquery-public-data.cloudpartnerservices.goog
TITLE: BigQuery Public Data Human Variant Annotation Datasets

NAME: bigquery.googleapis.com
TITLE: BigQuery API

NAME: bigqueryconnection.googleapis.com
TITLE: BigQuery Connection API

NAME: bigquerydatapolicy.googleapis.com
TITLE: BigQuery Data Policy API

NAME: bigquerydatatransfer.googleapis.com
TITLE: BigQuery Data Transfer API

NAME: bigquerymigration.googleapis.com
TITLE: BigQuery Migration API

NAME: bigqueryreservation.googleapis.com
TITLE: BigQuery Reservation API

NAME: bigquerystorage.googleapis.com
TITLE: BigQuery Storage API

NAME: bigtable-on-premises-2364.cloudpartnerservices.goog
TITLE: Google Teradata to BigQuery Data Migration (Beta)

NAME: cannabis-genome-bigquery-public-data.cloudpartnerservices.goog
TITLE: BigQuery Public Data 1000 Cannabis Genome Project

NAME: cloudtrace.googleapis.com
TITLE: Cloud Trace API

NAME: country-codes-bigquery-public-data.cloudpartnerservices.goog

vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud services enable bigqueryconnection.googleapis.com
Operation "operations/acat.p2-791612085972-ec2ebed1-9ec2-4c83-91a7-116c1c0841ca" finished successfully.
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud services enable bigquery.googleapis.com
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ |

```

### Examine dataset

```

vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ wget https://thefengs.com/wuchang/courses/cs430/yob2014.txt
--2022-11-27 09:34:52-- https://thefengs.com/wuchang/courses/cs430/yob2014.txt
Resolving thefengs.com (thefengs.com)... 131.252.220.66
Connecting to thefengs.com (thefengs.com)|131.252.220.66|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 425485 (416K) [text/plain]
Saving to: 'yob2014.txt'

yob2014.txt
100%[=====] 415.51K --KB/s in 0.04s

2022-11-27 09:34:52 (11.9 MB/s) - 'yob2014.txt' saved [425485/425485]

vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ head -3 yob2014.txt
Em,F,20799
Qiana,F,19074
Sophia,F,18490
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ wc -l yob2014.txt
30044 yob2014.txt
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ |

vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gsutil mb -l us-west1 gs://vkarinje-bucket
Creating gs://vkarinje-bucket/...
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gsutil cp yob2014.txt gs://vkarinje-bucket/
Copying file:///yob2014.txt [Content-Type=text/plain]...
/ [1 files][415.5 KiB/415.5 KiB]
Operation completed over 1 objects/415.5 KiB.
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ |

```

# Create dataset

The screenshot shows the Google Cloud Platform console interface. At the top, there's a search bar and a navigation bar with 'cloud-CS-530-karinje-vkarinje' selected. The main area is divided into two panels. The left panel, titled 'Explorer', shows a search bar and a list of resources under 'cloud-CS-530-karinje-vkarinje', including 'yob'. The right panel, titled 'Dataset info', displays details for the 'yob' dataset:

Dataset info	
Dataset ID	cloud-CS-530-karinje-vkarinje.yob
Created	Nov 26, 2022, 7:42:09 PM UTC-8
Default table expiration	Never
Last modified	Nov 26, 2022, 7:42:09 PM UTC-8
Data location	us-west1
Description	
Default collation	

- Take a screenshot of the table's details that includes the number of rows in the table.

The screenshot shows the Google Cloud Platform console interface. At the top, there's a search bar and a navigation bar with 'cloud-CS-530-karinje-vkarinje' selected. The main area is divided into two panels. The left panel, titled 'Explorer', shows a search bar and a list of resources under 'cloud-CS-530-karinje-vkarinje', including 'yob' and 'baby\_names'. The right panel, titled 'baby\_names', displays details for the 'baby\_names' table. The 'PREVIEW' tab is selected, showing a table with columns 'Row', 'name', 'gender', and 'count'.

Row	name	gender	count
1	Emma	F	20799
2	Olivia	F	19674
3	Sophia	F	18490
4	Isabella	F	16950
5	Ava	F	15386
6	Mia	F	13442
7	Emily	F	12562
8	Abigail	F	11985
9	Madison	F	10247
10	Charlotte	F	10048
11	Harper	F	9564
12	Sofia	F	9542
13	Avery	F	9517
14	Elizabeth	F	9492
15	Amelia	F	8727
16	Evelyn	F	8692
17	Ella	F	8489
18	Chloe	F	8469
19	Victoria	F	7955
20	Isabella	F	7500

At the bottom of the table, it says 'Results per page: 50' and '1 - 50 of 33044'. Below the table, there are tabs for 'PERSONAL HISTORY' and 'PROJECT HISTORY'.

The screenshot shows the Google Cloud BigQuery console interface. At the top, there's a search bar and a dropdown menu showing 'cloud-CS-530-karinje-vkarinje'. Below this, the 'Explorer' panel on the left shows a tree view of resources: 'cloud-cs-530-karinje-vkarinje' > 'yob' > 'baby\_names'. The 'baby\_names' table is selected and highlighted. The main panel on the right shows the 'DETAILS' tab for the 'baby\_names' table. It includes a 'Table info' section with various attributes and a 'Storage info' section with row and byte counts.

**Table info**

Table ID	cloud-cs-530-karinje-vkarinje.yob.baby_names
Created	Nov 26, 2022, 7:53:11 PM UTC-8
Last modified	Nov 26, 2022, 7:53:11 PM UTC-8
Table expiration	NEVER
Data location	us-west1
Default collation	
Description	

**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	0 B

At the bottom of the console, there are tabs for 'PERSONAL HISTORY' and 'PROJECT HISTORY'.

## Query data

- Screenshot your results and include it in your lab notebook

The screenshots show the Azure Data Explorer interface for the 'cloud-CS-530-karinje-vkarinje' workspace. The left pane shows the Explorer with the 'baby\_names' table selected. The middle pane shows the 'Table info' for 'baby\_names', including details like Table ID, Created, Last modified, Table expiration, Data location, Default collation, and Description. The right pane shows the 'Query results' for a query. The query results are displayed in a table with columns 'name' and 'count'.

**Table info:**

- Table ID: cloud-cs-530-karinje-vkarinje.yob.baby\_names
- Created: Nov 26, 2022, 7:53:11 PM UTC-8
- Last modified: Nov 26, 2022, 7:53:11 PM UTC-8
- Table expiration: NEVER
- Data location: us-west1
- Default collation: us-west1
- Description:

**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: 0 B

**Query results (Top 19 names):**

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

**Query results (Top 20 names):**

Row	name	count
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

- Screenshot your results and include it in your lab notebook

```

vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje) $ bq query "SELECT name, count FROM [cloud-cs-530-karinje-vkarinje.yob.baby_names] WHERE gender='M' ORDER BY count ASC LIMIT 10"
+-----+-----+
| name | count |
+-----+-----+
| Aari | 5 |
| Aaliyah | 5 |
| Aadian | 5 |
| Aaroh | 5 |
| Aarib | 5 |
| Aadiv | 5 |
| Aadhi | 5 |
| Aarohan | 5 |
| Aariyan | 5 |
| Aamer | 5 |
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje) $

```

At the prompt, you can then enter your query. Run a query to find the 10 most popular male names in 2014.

- Screenshot your results and include it in your lab notebook

```
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ bq shell
Welcome to BigQuery! (Type help for more information.)
cloud-cs-530-karinje-vkarinje> SELECT name, count from [cloud-cs-530-karinje-vkarinje:yob.baby_names] WHERE gender='M' ORDER BY count ASC LIMIT 10
+-----+-----+
| name | count |
+-----+-----+
| Aari | 5 |
| Aliyah | 5 |
| Adian | 5 |
| Aaroh | 5 |
| Aarit | 5 |
| Aadiw | 5 |
| Aadhi | 5 |
| Aarohan | 5 |
| Ariyan | 5 |
| Amer | 5 |
+-----+-----+
cloud-cs-530-karinje-vkarinje> ||
```

Finally, run a query on your name. How popular was it?

My name appeared 20 times.

- Screenshot your results and include it in your lab notebook

```
cloud-cs-530-karinje-vkarinje> SELECT name, count from [cloud-cs-530-karinje-vkarinje:yob.baby_names] WHERE name='Varsha'
+-----+-----+
| name | count |
+-----+-----+
| Varsha | 20 |
+-----+-----+
cloud-cs-530-karinje-vkarinje> ||
```

## BigQuery, Notebooks Lab #2 (Natality)

```
Google Cloud cloud-cs-530-karinje-vkarinje Search for resources, docs, products, and more (/)
CLOUD SHELL Terminal (cloud-cs-530-karinje-vkarinje) + + Open Editor

Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to cloud-cs-530-karinje-vkarinje.
Use "gcloud config set project [PROJECT_ID]" to change to a different project.
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud services enable notebooks.googleapis.com
Operation "operations/acmt.pl-791612085972-34b47971-9f13-4d1e-a6cf-03ed7cd1e0a1" finished successfully.
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud iam service-accounts create cs430jupyter
Created service account [cs430jupyter].
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud projects add-iam-policy-binding $(GOOGLE_CLOUD_PROJECT) --member serviceAccount:cs430jupyter@$(GOOGLE_CLOUD_PROJECT).iam.gserviceaccount.com --role roles/bigquery.user
Updated IAM policy for project [cloud-cs-530-karinje-vkarinje].
bindings:
- members:
  - serviceAccount:service-791612085972@gcp-gae-service.iam.gserviceaccount.com
  role: roles/appengine.serviceAgent
- members:
  - serviceAccount:cs430jupyter@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com
  role: roles/bigquery.user
- members:
  - serviceAccount:791612085972@cloudbuild.gserviceaccount.com
  role: roles/cloudbuild.builder
- members:
  - serviceAccount:service-791612085972@gcp-na-cloudbuild.iam.gserviceaccount.com
  role: roles/cloudbuild.serviceAgent
- members:
  - serviceAccount:service-791612085972@gcp-admin-robot.iam.gserviceaccount.com
  role: roles/cloudfunctions.serviceAgent
- members:
  - serviceAccount:service-791612085972@gcp-na-cloudscheduler.iam.gserviceaccount.com
  role: roles/cloudscheduler.serviceAgent
- members:
  - serviceAccount:service-791612085972@compute-system.iam.gserviceaccount.com
  role: roles/compute.serviceAgent
- members:
  - deleted:serviceAccount:gcp-lab@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com?uid=10229672572951542201
  role: roles/compute.viewer
- members:
  - serviceAccount:service-791612085972@container-engine-robot.iam.gserviceaccount.com
  role: roles/container.serviceAgent
- members:
  - serviceAccount:service-791612085972@containerregistry.iam.gserviceaccount.com
  role: roles/containerregistry.serviceAgent
- members:
  - deleted:serviceAccount:cs430flee@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com?uid=114128102459650149292
  - serviceAccount:cloud-cs-530-karinje-vkarinje@aggspt.gserviceaccount.com
```

```

vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje) $ gcloud notebooks instances create bq-jupyter-instance \
--vm-image=project:deeplearning-platform-release \
--vm-image-family=ml-2-2-tp4 \
--machine-type=ml-standard-1 \
--location=us-west1-b \
--service-account=cs121jupyter@gcp.gcpcloud.com
Waiting for operation on instance [bq-jupyter-instance] to be created with [projects/cloud-cs-530-karinje-vkarinje/locations/us-west1-b/operations/operation-1669529588225-5ee6ffa1960a4-024157a7-42dc51fa]...done.
Created notebook instance bq-jupyter-instance [https://notebooks.googleapis.com/v1/projects/cloud-cs-530-karinje-vkarinje/locations/us-west1-b/operations/operation-1669529588225-5ee6ffa1960a4-024157a7-42dc51fa].
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje) $

```

## Launch notebook

The screenshot displays the Google Cloud Vertex AI Workbench interface. The top navigation bar includes the Google Cloud logo, a dropdown menu for the project 'cloud-cs-530-karinje-vkarinje', and a search bar. The main content area is divided into a left sidebar with navigation options (Tools, DATA, MODEL DEVELOPMENT, DEPLOY AND USE) and a central workspace. The 'MANAGED NOTEBOOKS' tab is active, showing a table of notebook instances. The table has columns for Notebook name, Zone, Auto-upgrade, Environment, Machine type, GPUs, and Owner. A single instance named 'bq-jupyter-instance' is listed, with a status of 'OPEN JUPYTERLAB'. The right sidebar contains an 'Info panel' with links to documentation and troubleshooting. The bottom of the interface shows a browser window with the URL '51d21d08bef21c79-dot-us-west1.notebooks.googleusercontent.com/lab?authuser=0&username=Varsha\_Karinje' and a file explorer showing the 'Untitled.ipynb' file.

## BigQuery query

The screenshot shows the Google Cloud BigQuery Explorer interface. The top navigation bar includes the Google Cloud logo, a dropdown menu for the project 'cloud-cs-530-karinje-vkarinje', and a search bar. The main content area is divided into a left sidebar with navigation options (Explorer, + ADD DATA) and a central workspace. The 'Explorer' tab is active, showing a list of datasets. The 'cloud-cs-530-karinje-vkarinje' dataset is selected, and the 'job' sub-dataset is visible. The central workspace displays a SQL query: 'SELECT \* FROM bigquery-public-data.samples.natality'. The bottom of the interface shows a status bar indicating 'This query will process 21.94 GB when run.'

Answer the following question for your lab notebook:

- How many twins were born during this time?



The screenshot shows the Google Cloud BigQuery console. At the top, there's a search bar with 'big' and a 'cloud-CS-530-karinje-vkarinje' dropdown. Below the search bar, there's a toolbar with 'RUN', 'SAVE', 'SHARE', 'SCHEDULE', and 'MORE' buttons. The main area displays a SQL query:

```

1 SELECT
2   plurality,
3   COUNT(1) AS num_babies,
4   AVG(weight_pounds) AS avg_wt
5 FROM
6   bigquery-public-data.samples.natality
7 WHERE
8   year between 2001 and 2003
9 GROUP BY
10  plurality

```

Below the query, the 'Query results' section is visible, showing a table with 5 rows and 4 columns: 'plurality', 'num\_babies', 'avg\_wt', and an unnamed column. The results are as follows:

Row	plurality	num_babies	avg_wt
1	4	1407	2.85196658...
2	5	239	2.69670186...
3	2	375362	5.17462027...
4	1	11757058	7.34578073...
5	3	20933	3.70977157...

There were 375362 twins born during this time.

## Jupyter notebook query

The screenshot shows a Jupyter Notebook interface. The top bar displays the URL '51d21d08bef21c79-dot-us-west1.notebooks.googleusercontent.com/lab?authuser=0&username=Varsha\_Karinje'. Below the URL bar, there's a toolbar with 'File', 'Edit', 'View', 'Run', 'Kernel', 'Git', 'Tabs', 'Settings', and 'Help' menus. The main area shows a Python 3 notebook with the following code:

```

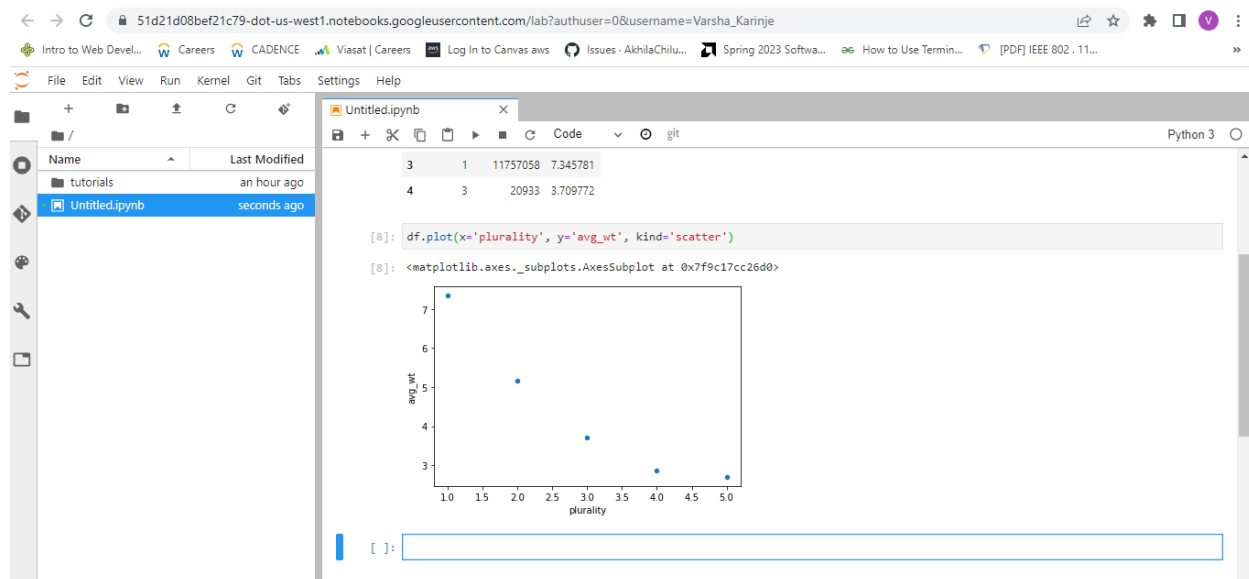
[6]: query_string = "SELECT plurality, COUNT(1) AS num_babies, AVG(weight_pounds) AS avg_wt FROM bigquery-public-data.samples.natality WHERE year between 2001 and 2003 GROUP BY plurality"

[7]: from google.cloud import bigquery
df = bigquery.Client().query(query_string).to_dataframe()
df.head()

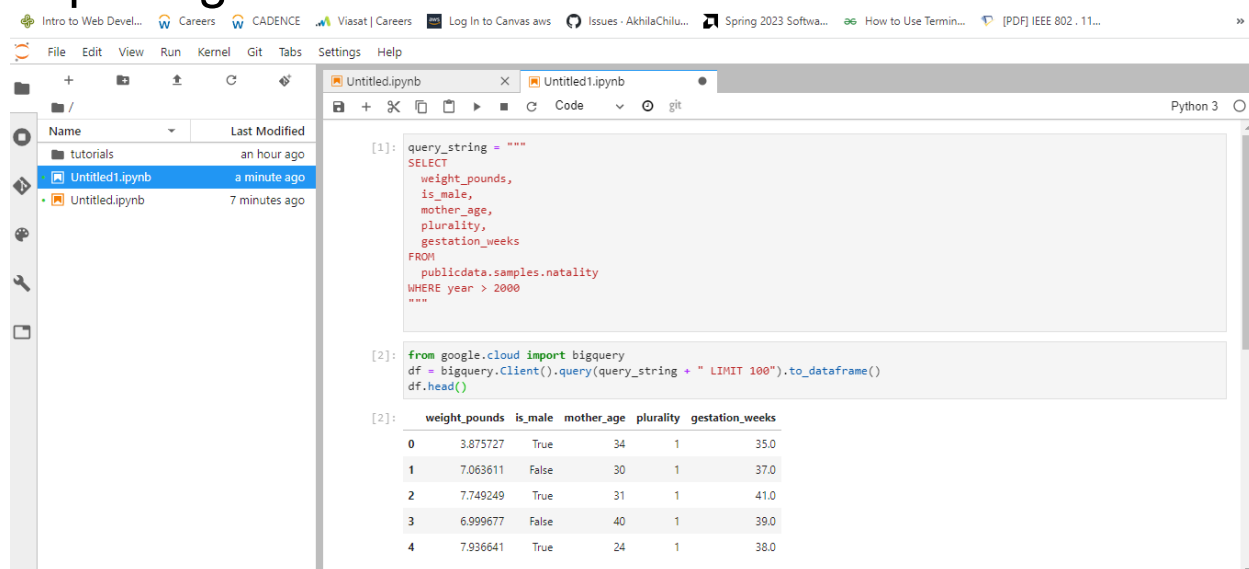
```

The output of the query is displayed as a table with 5 rows and 4 columns: 'plurality', 'num\_babies', 'avg\_wt', and an unnamed column. The results are as follows:

	plurality	num_babies	avg_wt
0	4	1407	2.851967
1	5	239	2.696702
2	2	375362	5.174620
3	1	11757058	7.345781
4	3	20933	3.709772



## Exploring the dataset



## Run queries

51d21d08bef21c79-dot-us-west1.notebooks.googleusercontent.com/lab?authuser=0&username=Varsha\_Karinje

Intro to Web Devel... Careers CADENCE Viasat | Careers Log In to Canvas aws Issues · AkhilaChilu... Spring 2023 Softwa... How to Use Termin... [PDF] IEEE 802.11...

File Edit View Run Kernel Git Tabs Settings Help


Untitled.ipynb x Untitled1.ipynb x Untitled2.ipynb x

Python 3

```
[4]: def get_distinct_values(column_name):
      query_string = f"""
      SELECT
      {column_name},
      COUNT(1) AS num_babies,
      AVG(weight_pounds) AS avg_wt
      FROM
      publicdata.samples.natality
      WHERE
      year > 2000
      GROUP BY
      {column_name}
      """
      return bigquery.Client().query(query_string).to_dataframe().sort_values(column_name)

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

[5]: <matplotlib.axes._subplots.AxesSubplot at 0x7f908e4167d0>
```



51d21d08bef21c79-dot-us-west1.notebooks.googleusercontent.com/lab?authuser=0&username=Varsha\_Karinje

Intro to Web Devel... Careers CADENCE Viasat | Careers Log In to Canvas aws Issues · AkhilaChilu... Spring 2023 Softwa... How to Use Termin... [PDF] IEEE 802.11...

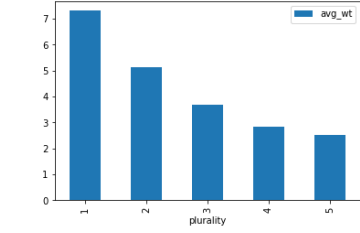
File Edit View Run Kernel Git Tabs Settings Help

Untitled.ipynb x Untitled1.ipynb x Untitled2.ipynb x

Python 3

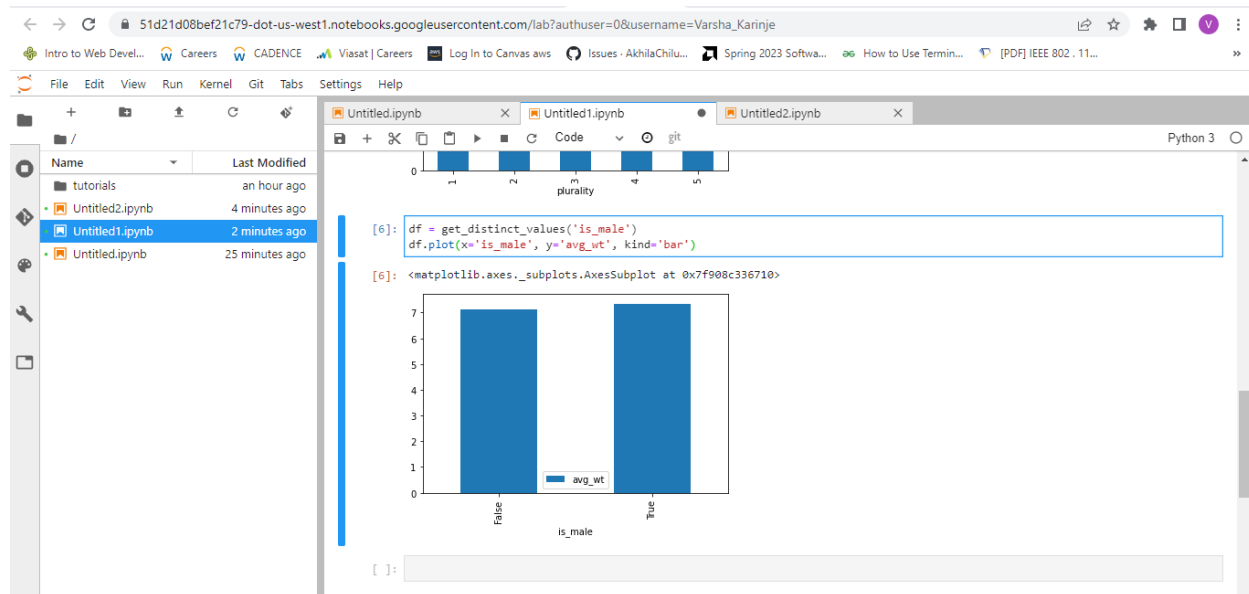
```
[5]: df = get_distinct_values('plurality')
      df.plot(x='plurality', y='avg_wt', kind='bar')

[5]: <matplotlib.axes._subplots.AxesSubplot at 0x7f908e4167d0>
```

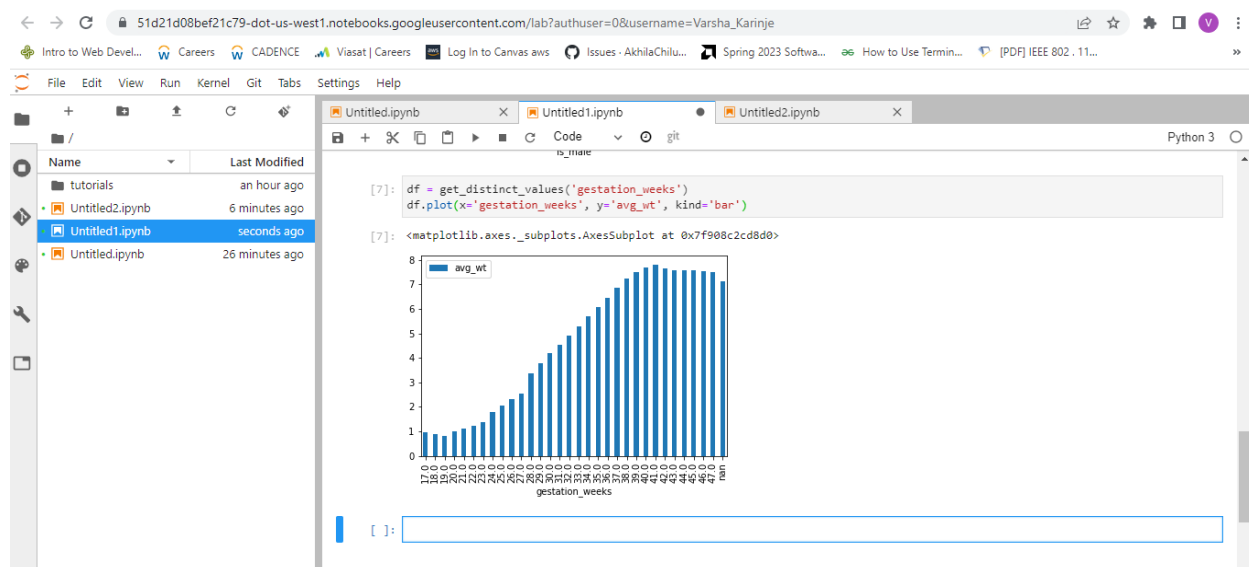


[ ]:

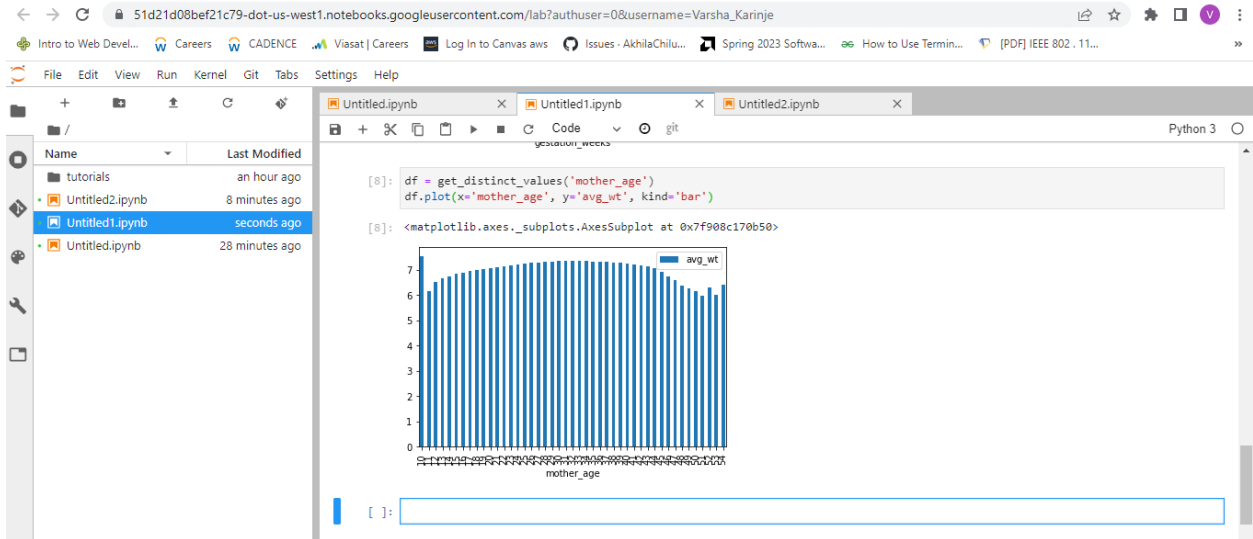
Then, run the query using gender:



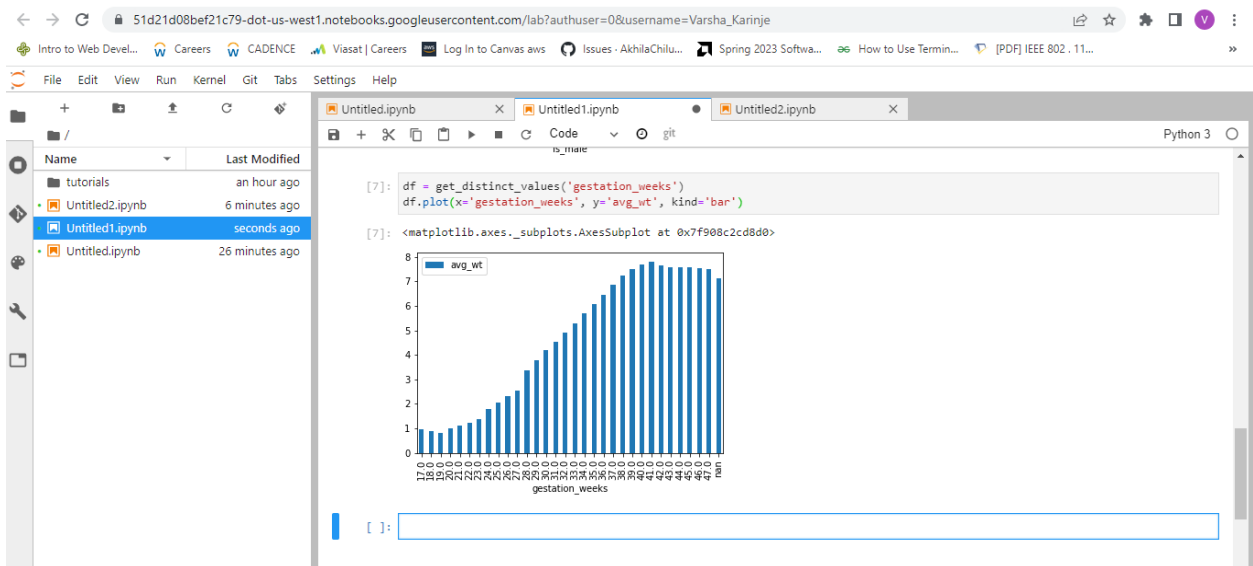
Then, run the query using gestation time:

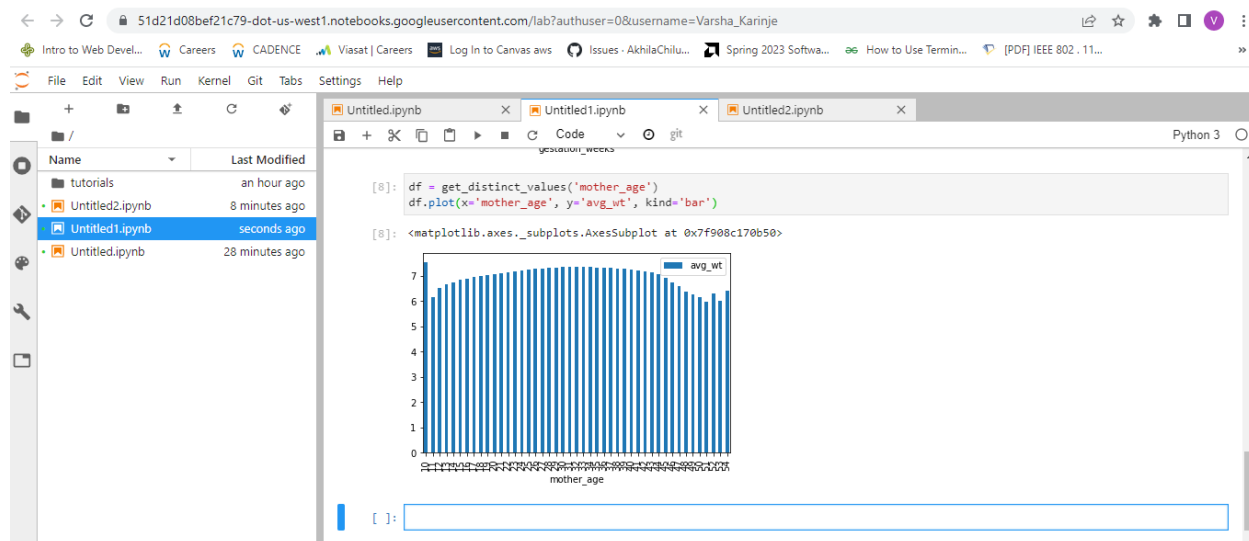


Finally, run the query using the mother's age:



- In examining the plots, which two features are the strongest predictors for a newborn baby's weight?
- Ans: The gestation time and the mother's age are the two most important features that are the strong predictors for a newborn baby's weight.
- Show the plots generated for the two most important features for your lab notebook





## BigQuery, Notebooks Lab #3 (COVID-19 Mobility)

Find the link that documents what the dataset measures and answer the following question:

- What dates are used as a baseline for the mobility data?

Google Cloud Explorer showing the schema for the `mobility_report` dataset.

Field name	Type	Mode	Collation	Default Value	Policy Tags	Description
<code>country_region_code</code>	STRING	NULLABLE				2 letter alpha code for the country/region in which changes are measured relative to the baseline. These values correspond with the ISO 3166-1 alpha-2 codes
<code>country_region</code>	STRING	NULLABLE				The country/region in which changes are measured relative to the baseline
<code>sub_region_1</code>	STRING	NULLABLE				First geographic sub-region in which the data is aggregated. This varies by country/region to ensure privacy and public health value in consultation with local public health authorities
<code>sub_region_2</code>	STRING	NULLABLE				Second geographic sub-region in which the data is aggregated. This varies by country/region to ensure privacy and public health value in consultation with local public health authorities
<code>metro_area</code>	STRING	NULLABLE				A specific metro area to measure mobility within a given city/metro area. This varies by country/region to ensure privacy and public health value in consultation with local public health authorities
<code>iso_3166_2_code</code>	STRING	NULLABLE				Unique identifier for the geographic region as defined by ISO Standard 3166-2
<code>census_fips_code</code>	STRING	NULLABLE				Unique identifier for each US county as defined by the US Census Bureau. Maps to <code>county_fips_code</code> in other tables
<code>place_id</code>	STRING	NULLABLE				A textual identifier that uniquely identifies a place in the Google Places

BigQuery-public-data was starred.

← → ↻ google.com/covid19/mobility/data\_documentation.html

Intro to Web Devel... Careers CADENCE Viasat | Careers Log In to Canvas aws Issues - AkhilaChilu... Spring 2023 Softwa... How to Use Termin... [PDF] IEEE 802 . 11...

## About this data

These datasets show how visits and length of stay at different places change compared to a baseline. We calculate these changes using the same kind of aggregated and anonymized data used to show [popular times](#) for places in Google Maps.

Changes for each day are compared to a baseline value for that day of the week:

- The baseline is the median value, for the corresponding day of the week, during the 5-week period Jan 3–Feb 6, 2020.
- The datasets show trends over several months with the most recent data representing approximately 2-3 days ago—this is how long it takes to produce the datasets.

What data is included in the calculation depends on user settings, connectivity, and whether it meets our privacy threshold. When the data doesn't meet quality and privacy thresholds, you might see empty fields for certain places and dates.

We include categories that are useful to social distancing efforts as well as access to essential services.

We calculate these insights based on data from users who have opted-in to Location History for their Google Account, so the data represents a sample of our users. As with all samples, this may or may not represent the exact behavior of a wider population.

Ans: The baseline is the median value, for the corresponding day of the week, during the 5-week period Jan 3–Feb 6, 2020.

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

Google Cloud cloud-CS-530-karinje-vkarinje Search for resources, docs, products, and more (/) Search

Explorer + ADD DATA

mobility\_report

SCHEMA DETAILS PREVIEW

Created Nov 26, 2022, 4:11:11 PM UTC-8

Last modified Nov 26, 2022, 4:11:51 PM UTC-8

Table expiration NEVER

Data location US

Default collation

Description Terms of use By downloading or using the data, you agree to Google's Terms of Service: <https://policies.google.com/terms> Description This dataset aims to provide insights into what has changed in response to policies aimed at combating COVID-19. It reports movement trends over time by geography, across different categories of places such as retail and recreation, groceries and

Query results

Row	census_fips_code	place_id	date	retail_and_recrea	grocery_and_ph	parks_percent	transit_stations	workplaces
1	41051	ChIJsbYckvDIVQR6bqX-gieH8	2020-03-01	19	9	23	7	
2	41051	ChIJsbYckvDIVQR6bqX-gieH8	2020-03-02	4	7	5	0	
3	41051	ChIJsbYckvDIVQR6bqX-gieH8	2020-03-03	5	12	36	2	
4	41051	ChIJsbYckvDIVQR6bqX-gieH8	2020-03-04	8	7	49	4	
5	41051	ChIJsbYckvDIVQR6bqX-gieH8	2020-03-05	5	7	37	0	

bigquery-public-data was starred.

Results per page: 50 1 - 31 of 31

mobility\_report

Created: Nov 26, 2022, 4:11:11 PM UTC-8  
Last modified: Nov 26, 2022, 4:11:51 PM UTC-8  
Table expiration: NEVER  
Data location: US  
Default collation: US  
Description: Terms of use By downloading or using the data, you agree to Google's Terms of Service: https://policies.google.com/terms This dataset aims to provide insights into what has changed in response to policies aimed at combating COVID-19. It reports movement trends over time by geography, across different categories of places such as retail and

Query results

Row	census_fips_code	place_id	date	retail_and_recreation	grocery_and_pharmacy	parks_percent	transit_stations	workplaces_percent
12	41051	CHJabYckvDIVQR8boKj-gien8	2020-03-12	0	16	43	-9	-5
13	41051	CHJabYckvDIVQR8boKj-gien8	2020-03-13	-11	17	-21	-17	-10
14	41051	CHJabYckvDIVQR8boKj-gien8	2020-03-14	-30	3	-23	-22	-11
15	41051	CHJabYckvDIVQR8boKj-gien8	2020-03-15	-17	1	13	-20	-8
16	41051	CHJabYckvDIVQR8boKj-gien8	2020-03-16	-14	15	50	-25	-28

Ans: 2020-03-13 was the day that saw the largest spike in trips to grocery and pharmacy stores

- On the day the stay-at-home order took effect (3/23/2020), what was the total impact on workplace trips?

Ans: There was a 49% decrease in workplace trips on the day the stay-at-home order took effect (3/23/2020)

Find the column in this table that gives us information on the traffic impact.

airport\_traffic

This is a partitioned table. [Learn more](#)

Filter: Enter property name or value

Field name	Type	Mode	Collation	Default Value	Policy Tags	Description
aggregation_method	STRING	NULLABLE				Aggregation period used to compute this metric
date	DATE	NULLABLE				Date of the data
version	STRING	NULLABLE				Version of the table
airport_name	STRING	NULLABLE				Aggregation period used to compute this metric
percent_of_baseline	FLOAT	NULLABLE				Proportion of trips on this date as compared to Avg number of trips on the same day of week in baseline period i.e 1st February 2020 - 15th March 2020
center_point_geom	GEOGRAPHY	NULLABLE				Geographic representation of the centroid of the Airport polygon
city	STRING	NULLABLE				City within which the Airport is located
state_region	STRING	NULLABLE				State within which the Airport is located
country_iso_code_2	STRING	NULLABLE				ISO 3166-2 code representing the country and subdivision within which the Airport is located
country_name	STRING	NULLABLE				Full text name of the country within which the Airport is located
airport_geom	GEOGRAPHY	NULLABLE				Geographic representation of the Airport polygon

percent\_of\_baseline is the table that gives us information on the traffic impact

Then, adapt the query below to find the following



- Which three airports were impacted the most in April 2020 (the month when lockdowns became widespread)?

Google Cloud interface showing a BigQuery query for April 2020 airport traffic. The query filters for the month of April and orders results by traffic fraction. The results table shows 11 airports, with the top three being Detroit Metropolitan Wayne County, McCarran International, and San Francisco International.

Row	airport_name	traffic_fraction
1	Detroit Metropolitan Wayne Co...	45.416000...
2	McCarran International	45.599999...
3	San Francisco International	47.266666...
4	Washington Dulles International	51.233333...
5	Denver International	53.449999...
6	LaGuardia	55.933333...
7	Hartsfield-Jackson Atlanta Inte...	60.149999...
8	Boston Logan International	61.75
9	John F. Kennedy International	69.733333...
10	Dallas/Fort Worth International	70.733333...
11	Seattle-Tacoma International	71.850000...

Low traffic fraction: Detroit Metropolitan Wayne Cou, McCarran International, San Francisco International

High traffic fraction: Chicago OHare International, Daniel K. Inouye International, Newark Liberty International

- Run the query again using the month of August 2020. Which three airports were impacted the most?

Google Cloud interface showing a BigQuery query for August 2020 airport traffic. The query filters for the month of August and orders results by traffic fraction. The results table shows 16 airports, with the top three being Dallas/Fort Worth International, Charlotte Douglas International, and Newark Liberty International.

Row	airport_name	traffic_fraction
1	McCarran International	44.2
2	Detroit Metropolitan Wayne Co...	45.100000...
3	San Francisco International	53.025000...
4	Washington Dulles International	58.574999...
5	LaGuardia	64.024999...
6	Boston Logan International	64.650000...
7	Hartsfield-Jackson Atlanta Inte...	64.682926...
8	John F. Kennedy International	65.902499...
9	Denver International	66.075000...
10	Miami International	69.975000...
11	Seattle-Tacoma International	73.774999...
12	Los Angeles International	74.583456...
13	Chicago O'Hare International	79.149999...
14	Daniel K. Inouye International	79.599999...
15	Dallas/Fort Worth International	81.487804...
16	Charlotte Douglas International	89.175000...

Low traffic fraction: McCarran International, Detroit Metropolitan Wayne Coun, San Francisco International

High traffic fraction: Dallas/Fort Worth International, Charlotte Douglas International, Newark Liberty International

## BigQuery, Notebooks Lab #4 (COVID-19 NYT)

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

The screenshot shows the Google Cloud BigQuery Explorer interface. On the left, the 'Explorer' pane lists various datasets, with 'excess\_deaths' selected under the 'covid19' dataset. The main pane displays the schema for the 'excess\_deaths' table. The schema table is as follows:

Field name	Type	Mode	Collation	Default Value	Policy Tags	Description
country	STRING	NULLABLE				The country reported
placename	STRING	NULLABLE				The place in the country reported
frequency	STRING	NULLABLE				Weekly or monthly, depending on how the data is recorded
start_date	DATE	NULLABLE				The first date included in the period
end_date	DATE	NULLABLE				The last date included in the period
year	STRING	NULLABLE				Year reported
month	INTEGER	NULLABLE				Numerical month
week	INTEGER	NULLABLE				Epidemiological week, which is a standardized way of counting weeks to allow for year-over-year comparisons. Mc
deaths	INTEGER	NULLABLE				The total number of confirmed deaths recorded from any cause
expected_deaths	INTEGER	NULLABLE				The baseline number of expected deaths, calculated from a historical average
excess_deaths	INTEGER	NULLABLE				The number of deaths minus the expected deaths
baseline	STRING	NULLABLE				The years used to calculate expected_deaths

Ans: The table name is excess\_deaths and the columns are placename, start\_date and excess\_deaths respectively.

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

The screenshot shows the Google Cloud BigQuery Explorer interface. On the left, the 'Explorer' pane lists various datasets, with 'us\_counties' selected under the 'covid19' dataset. The main pane displays the schema for the 'us\_counties' table. The schema table is as follows:

Field name	Type	Mode	Collation	Default Value	Policy Tags	Description
date	DATE	NULLABLE				Date reported
county	STRING	NULLABLE				County in the specified state
state_name	STRING	NULLABLE				State reported
county_fips_code	STRING	NULLABLE				Standard geographic identifier for the county
confirmed_cases	INTEGER	NULLABLE				The total number of confirmed cases of COVID-19
deaths	INTEGER	NULLABLE				The total number of confirmed deaths of COVID-19

Ans: The table is us\_counties and the columns are date, county and deaths

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

The screenshot shows the Google Cloud BigQuery Explorer interface. On the left, the 'Explorer' pane lists various datasets, with 'us\_states' selected. The main pane displays the 'us\_states' table schema. The schema table has columns: date (DATE), state\_name (STRING), state\_fips\_code (STRING), confirmed\_cases (INTEGER), and deaths (INTEGER). Each column has a description: 'Date reported', 'State reported', 'Standard geographic identifier for the state', 'The total number of confirmed cases of COVID-19', and 'The total number of confirmed deaths of COVID-19' respectively.

Field name	Type	Mode	Collation	Default Value	Policy Tags	Description
date	DATE	NULLABLE				Date reported
state_name	STRING	NULLABLE				State reported
state_fips_code	STRING	NULLABLE				Standard geographic identifier for the state
confirmed_cases	INTEGER	NULLABLE				The total number of confirmed cases of COVID-19
deaths	INTEGER	NULLABLE				The total number of confirmed deaths of COVID-19

Ans: The table name is us\_states and the columns are date, state\_name and confirmed\_cases

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

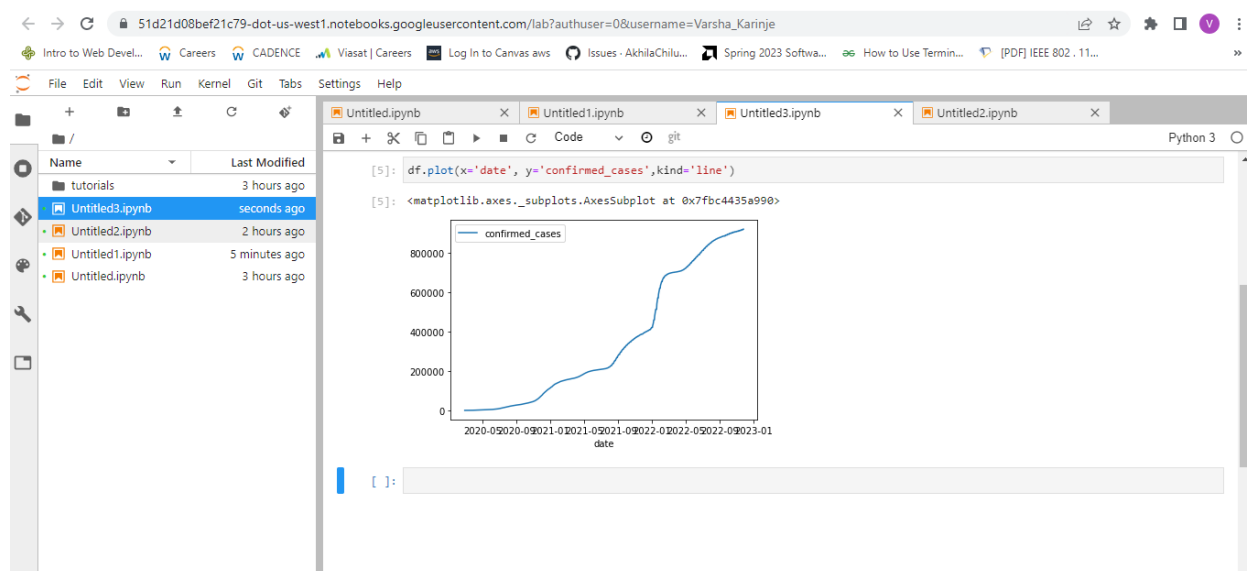
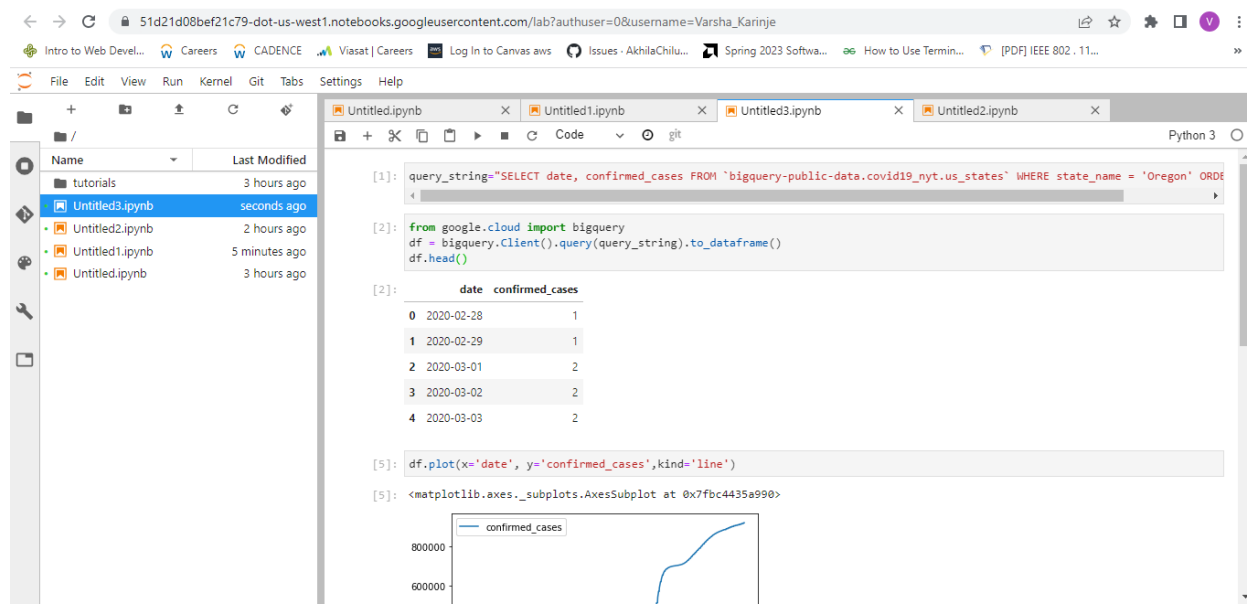
The screenshot shows the Google Cloud BigQuery Explorer interface. On the left, the 'Explorer' pane lists various datasets, with 'mask\_use\_by\_county' selected. The main pane displays the 'mask\_use\_by\_county' table schema. The schema table has columns: county\_fips\_code (STRING), never (FLOAT), rarely (FLOAT), sometimes (FLOAT), frequently (FLOAT), and always (FLOAT). Each column has a description: 'Standard geographic identifier for the county', 'The estimated share of people in this county who would say never in response to the question "How often do you v', 'The estimated share of people in this county who would say rarely', 'The estimated share of people in this county who would say sometimes', 'The estimated share of people in this county who would say frequently', and 'The estimated share of people in this county who would say always' respectively.

Field name	Type	Mode	Collation	Default Value	Policy Tags	Description
county_fips_code	STRING	NULLABLE				Standard geographic identifier for the county
never	FLOAT	NULLABLE				The estimated share of people in this county who would say never in response to the question "How often do you v
rarely	FLOAT	NULLABLE				The estimated share of people in this county who would say rarely
sometimes	FLOAT	NULLABLE				The estimated share of people in this county who would say sometimes
frequently	FLOAT	NULLABLE				The estimated share of people in this county who would say frequently
always	FLOAT	NULLABLE				The estimated share of people in this county who would say always

Ans: The table is mask\_use\_by\_county and the columns are county\_fips\_code and always

## Run example queries

- Show a screenshot of the plot and the code used to generate it for your lab notebook



- From within your Jupyter notebook, run the query and write code that shows the first 10 states that reached 1000 deaths from COVID-19. Take a screenshot for your lab notebook.

The screenshot shows a Google Colab notebook with the following code and output:

```
[3]: 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"""

[4]: from google.cloud import bigquery
df = bigquery.Client().query(query_string).to_dataframe()
df.head(10)
```

	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

- Take a screenshot for your lab notebook of the Top 5 counties and the states they are located in.

The screenshot shows a Google Colab notebook with the following code and output:

```
[5]: query_string="""SELECT DISTINCT mu.county_fips_code, mu.always, ct.county, ct.state_name
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"""

[6]: from google.cloud import bigquery
df = bigquery.Client().query(query_string).to_dataframe()
df.head(5)
```

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

## Write queries

Construct a query string that obtains the number of deaths from COVID-19 that have occurred in Multnomah county for each day in the dataset, ensuring the data is returned in ascending order of date. Run the query and obtain the results.

The image displays two screenshots of a Google Colab notebook interface. The top screenshot shows the execution of a BigQuery query to select COVID-19 deaths from the 'bigquery-public-data.covid19\_nyt.us\_counties' dataset, filtered by 'Multnomah' county and ordered by date. The bottom screenshot shows the resulting DataFrame with columns 'deaths', 'date', and 'county', displaying rows from index 0 to 990. The interface includes a file explorer on the left, a code editor in the center, and a terminal at the bottom.

**Top Screenshot:**

```
[33]: query_string="""Select deaths, date, county from `bigquery-public-data.covid19_nyt.us_counties` where county='Multnomah'
order by date asc """

[34]: from google.cloud import bigquery
df = bigquery.Client().query(query_string).to_dataframe()
df.head(100000)
```

	deaths	date	county
0	0	2020-03-10	Multnomah
1	0	2020-03-11	Multnomah
2	0	2020-03-12	Multnomah
3	0	2020-03-13	Multnomah
4	1	2020-03-14	Multnomah
...	...	...	...
986	1400	2022-11-21	Multnomah
987	1400	2022-11-22	Multnomah
988	1407	2022-11-23	Multnomah
989	1407	2022-11-24	Multnomah
990	1407	2022-11-25	Multnomah

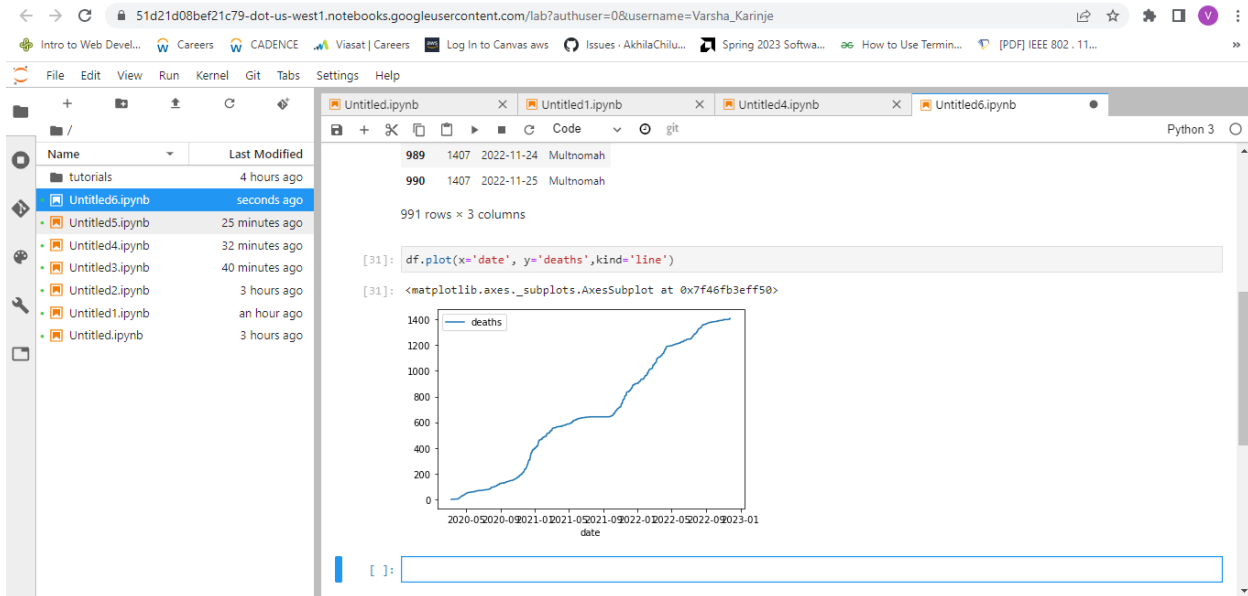
**Bottom Screenshot:**

	deaths	date	county
0	0	2020-03-10	Multnomah
1	0	2020-03-11	Multnomah
2	0	2020-03-12	Multnomah
3	0	2020-03-13	Multnomah
4	1	2020-03-14	Multnomah
...	...	...	...
986	1400	2022-11-21	Multnomah
987	1400	2022-11-22	Multnomah
988	1407	2022-11-23	Multnomah
989	1407	2022-11-24	Multnomah
990	1407	2022-11-25	Multnomah

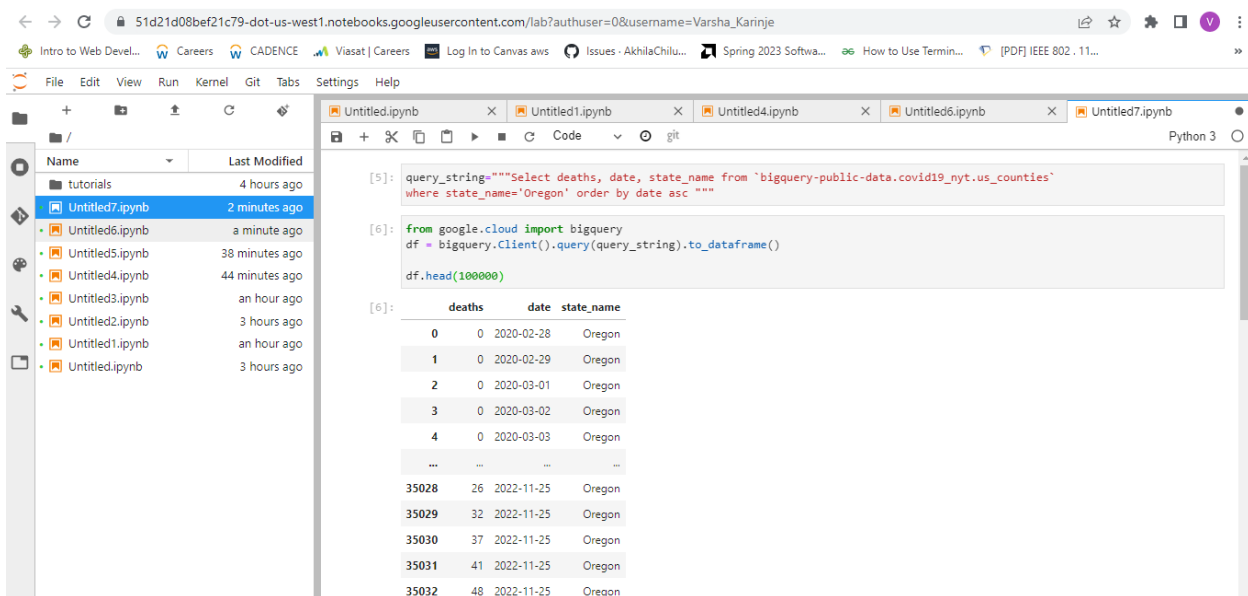
991 rows x 3 columns

```
[ ]:
```

- Plot the results and take a screenshot for your lab notebook.



Construct a query string that obtains the number of deaths from COVID-19 that have occurred in Oregon for each day in the dataset, ensuring the data is returned in ascending order of date. Run the query and obtain the results.



The screenshot shows a Google Colab notebook with a file explorer on the left and a code editor on the right. The code editor displays a table of data for Oregon, with columns labeled 'date', 'deaths', and 'state\_name'. The data shows a steady increase in deaths from February 2020 to November 2022.

date	deaths	state_name
2020-02-28	0	Oregon
2020-02-29	1	Oregon
2020-03-01	2	Oregon
2020-03-02	3	Oregon
2020-03-03	4	Oregon
...	...	...
2022-11-25	35028	Oregon
2022-11-25	35029	Oregon
2022-11-25	35030	Oregon
2022-11-25	35031	Oregon
2022-11-25	35032	Oregon

35033 rows x 3 columns

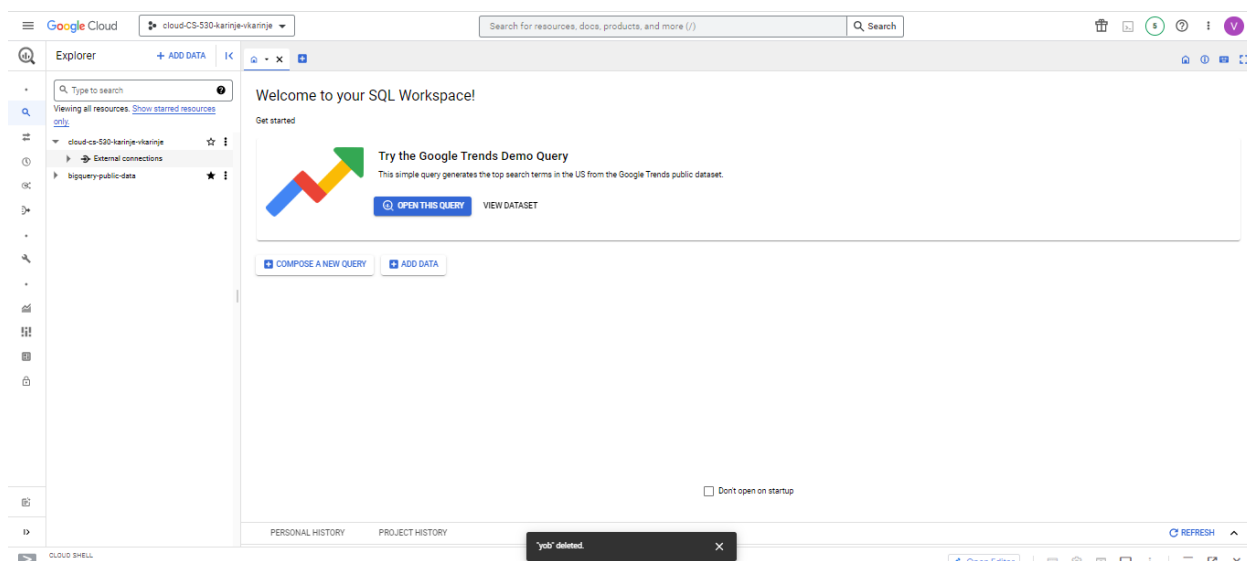
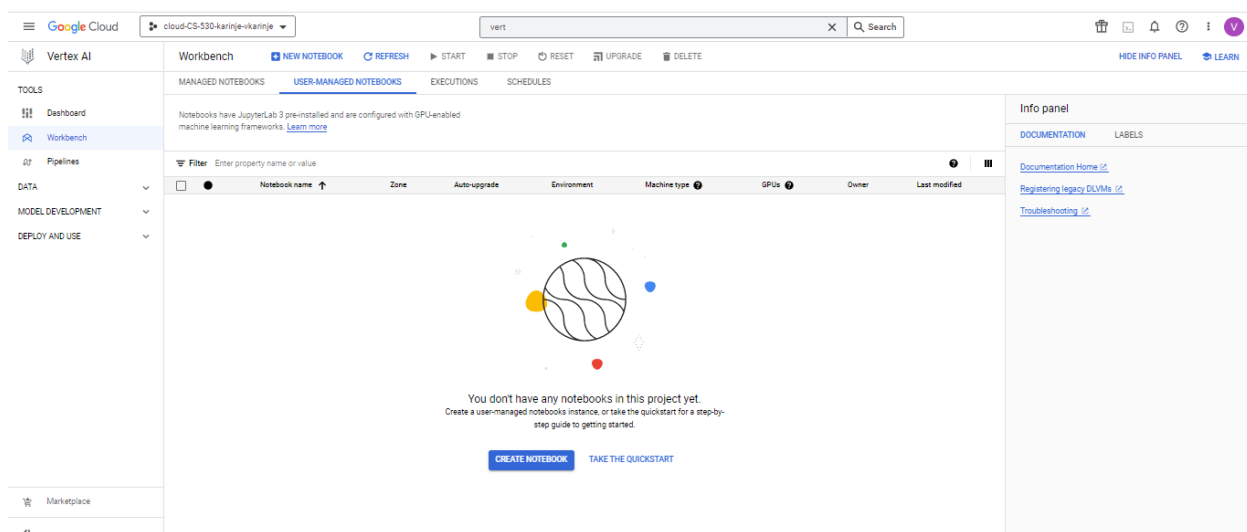
- Plot the results and take a screenshot for your lab notebook.

The screenshot shows a Google Colab notebook with a file explorer on the left and a code editor on the right. The code editor displays a line plot of COVID-19 deaths in Oregon. The x-axis is labeled 'date' and ranges from 2020-09 to 2022-01. The y-axis is labeled 'deaths' and ranges from 0 to 1400. The plot shows a steady increase in deaths over time.

```
[4]: df.plot(x='date', y='deaths', kind='line')
[4]: <matplotlib.axes._subplots.AxesSubplot at 0x7f2351bc2f50>
```

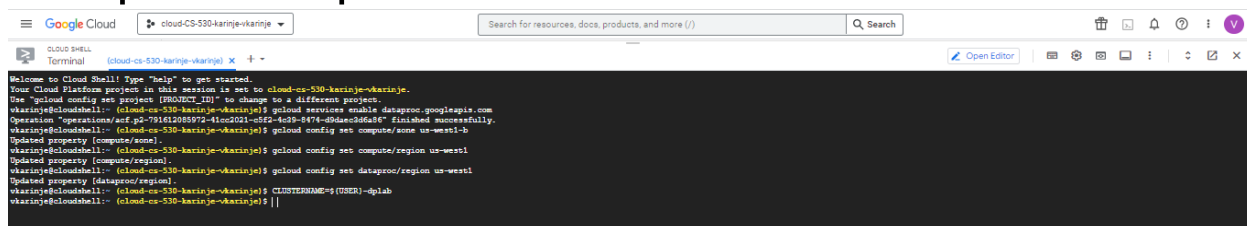
## Clean up





## 09.3g: Dataproc, Dataflow

### Dataproc setup



### Create Compute Engine cluster

The image shows a terminal window and two Google Cloud web console screenshots. The terminal window shows the command to create a Dataproc cluster named 'vkarinje-dplab' in the 'us-west1-b' zone. The console screenshots show the 'Dataproc Clusters' and 'Compute Engine VM instances' pages.

**Dataproc Clusters**

Name	Status	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created
vkarinje-dplab	Running	us-west1	us-west1-b	2	Off	dataproc-staging-us-west1-791612085972-2ajde05	Nov 28, 2022, 12:27:29 AM

**Compute Engine VM instances**

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
Running	vkarinje-dplab-m	us-west1-b			10.138.0.24 (nec)	34.168.203.236 (nec)	SSH
Running	vkarinje-dplab-e0	us-west1-b			10.138.0.26 (nec)	34.105.102.103 (nec)	SSH
Running	vkarinje-dplab-e1	us-west1-b			10.138.0.25 (nec)	34.168.107.250 (nec)	SSH

# Run computation

For your lab notebook:

- How long did the job take to execute?

Ans: The job took 30 seconds to execute

2 mins and 22 seconds

- Examine output.txt and show the estimate of  $\pi$  calculated.

Ans:

```
vkarinje@cloudshell:~$ (cloud-cs-530-karinje-vkarinje)$ date
Mon 28 Nov 2022 09:22:09 AM UTC
vkarinje@cloudshell:~$ (cloud-cs-530-karinje-vkarinje)$ gcloud dataproc jobs submit spark --cluster $(CLUSTERNAME) \
--class org.apache.spark.examples.SparkPi \
--jars file:///usr/lib/spark/examples/jars/spark-examples.jar -- 1000 \
>> output.txt &
[1] 1418
vkarinje@cloudshell:~$ (cloud-cs-530-karinje-vkarinje)$ gcloud dataproc jobs list --cluster $(CLUSTERNAME)
JOB ID: 9482291a6ab496db3fdae2fd2e44d
TYPE: spark
STATUS: DONE
[1]+  Done                  gcloud dataproc jobs submit spark --cluster $(CLUSTERNAME) --class org.apache.spark.examples.SparkPi --jars file:///usr/lib/spark/examples/jars/spark-examples.jar -- 1000 &> output.txt
vkarinje@cloudshell:~$ (cloud-cs-530-karinje-vkarinje)$ date
Mon 28 Nov 2022 09:25:21 AM UTC
vkarinje@cloudshell:~$ (cloud-cs-530-karinje-vkarinje)$ cat output.txt
Job [9482291a6ab496db3fdae2fd2e44d] submitted.
Waiting for job output...
22/11/28 09:23:22 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/11/28 09:23:22 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/11/28 09:23:22 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
22/11/28 09:23:22 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
22/11/28 09:23:22 INFO org.sparkproject.jetty.util.log: Logging initialised @6401ms to org.sparkproject.jetty.util.log.Slf4jLog
22/11/28 09:23:22 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git: b881a572662e194a1a4e12e7e1207989f212b74; jvm 1.8.0_352-b08
22/11/28 09:23:22 INFO org.sparkproject.jetty.server.Server: Started @6631ms
22/11/28 09:23:22 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@5fad41be(HTTP/1.1, (http/1.1))[(0.0.0.0:40279)]
22/11/28 09:23:24 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at vkarinje-dplab-m/10.138.0.35:8032
22/11/28 09:23:25 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at vkarinje-dplab-m/10.138.0.35:10200
22/11/28 09:23:26 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
22/11/28 09:23:26 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/11/28 09:23:26 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1669627200581_0001
22/11/28 09:23:29 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at vkarinje-dplab-m/10.138.0.35:8032
22/11/28 09:23:32 INFO com.google.cloud.hadoop.mapreduce.google.GoogleCloudHadoopImpl: Ignoring exception of type GoogleJsonResponseException: verified object already exists with desired state.
Pi is roughly 3.1415080314150803
22/11/28 09:23:57 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@5fad41be(HTTP/1.1, (http/1.1))[(0.0.0.0:0)]
Job [9482291a6ab496db3fdae2fd2e44d] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-west1-791612085972-2gjs6s95/google-cloud-dataproc-metainfo/b7aff59b-f4a3-4af1-bb1b-519d1ec5b1c/jobs/9482291a6ab496db3fdae2fd2e44d/driveroutput
driverOutputResourceUri: gs://dataproc-staging-us-west1-791612085972-2gjs6s95/google-cloud-dataproc-metainfo/b7aff59b-f4a3-4af1-bb1b-519d1ec5b1c/jobs/9482291a6ab496db3fdae2fd2e44d/driveroutput
jobUuid: 4cb765ae-116d-3ac8-a949-fcbb91fac114
placement:
  clusterName: vkarinje-dplab
  clusterUuid: b7aff59b-f4a3-4af1-bb1b-519d1ec5b1c
reference:
```

$\pi$  is roughly 3.1415080314150803

## Scale cluster

```

vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud dataproc clusters describe $(CLUSTERNAME)
clusterName: vkarinje-dplab
clusterUuid: b7aff99b-f4a2-4af1-bb1b-519d1ec5bb1c
config:
  configBucket: dataproc-staging-us-west1-791612085972-2gja6s95
  endpointConfig: {}
  gceClusterConfig:
    internalIpOnly: false
    networkUri: https://www.googleapis.com/compute/v1/projects/cloud-cs-530-karinje-vkarinje/global/networks/default
    serviceAccountScopes:
      - https://www.googleapis.com/auth/cloud-platform
      - https://www.googleapis.com/auth/cloud.useraccounts.readonly
      - https://www.googleapis.com/auth/devstorage.read_write
      - https://www.googleapis.com/auth/logging.write
    tags:
      - codelab
    zoneUri: https://www.googleapis.com/compute/v1/projects/cloud-cs-530-karinje-vkarinje/zones/us-west1-b
  masterConfig:
    diskConfig:
      bootDiskSizeGb: 30
      bootDiskType: pd-standard
    imageUri: https://www.googleapis.com/compute/v1/projects/cloud-dataproc/global/images/dataproc-2-0-debian10-20221108-035100-rc01
    instanceNames:
      - vkarinje-dplab-m
    machineTypeUri: https://www.googleapis.com/compute/v1/projects/cloud-cs-530-karinje-vkarinje/zones/us-west1-b/machineTypes/e2-medium
    minCpuPlatform: AUTOMATIC
    numInstances: 1
    preemptibility: NON_PREEMPTIBLE
  softwareConfig:
    imageVersion: 2.0.51-debian10
    properties:
      capacity-scheduler.yarn.scheduler.capacity.root.default.ordering-policy: fair
      core:fs.gs.block.size: '134217728'
      core:fs.gs.metadata.cache.enable: 'false'
      core:hadoop.ssl.enabled.protocols: TLSv1,TLSv1.1,TLSv1.2
      distcp:mapreduce.map.java.opts: -Xmx576m
      distcp:mapreduce.map.memory.mb: '768'
      distcp:mapreduce.reduce.java.opts: -Xmx576m
      distcp:mapreduce.reduce.memory.mb: '768'
      hadoop-env:HADOOP_DATANODE_OPTS: -Xmx512m
      hdfs:dfs.datanode.address: 0.0.0.0:9866
      hdfs:dfs.datanode.http.address: 0.0.0.0:9864
      hdfs:dfs.datanode.https.address: 0.0.0.0:9865
      hdfs:dfs.datanode.ipc.address: 0.0.0.0:9867
      hdfs:dfs.namenode.handler.count: '20'
      hdfs:dfs.namenode.http.address: 0.0.0.0:9870
      hdfs:dfs.namenode.https.address: 0.0.0.0:9871

```

```

CLOUD SHEL
Terminal (cloud-cs-530-karinje-vkarinje) X +
Open Editor

mapred:mapreduce.job.reduces: '1'
mapred:mapreduce.jobhistory.recovery.store.class: org.apache.hadoop.mapreduce.v2.history.HistoryServerLocalStateStoreService
mapred:mapreduce.map.cpu.wcores: '1'
mapred:mapreduce.map.java.opts: -Xmx1311m
mapred:mapreduce.map.memory.mb: '1638'
mapred:mapreduce.reduce.cpu.wcores: '2'
mapred:mapreduce.reduce.java.opts: -Xmx622m
mapred:mapreduce.reduce.memory.mb: '2278'
mapred:mapreduce.task.io.sort.mb: '256'
mapred:yarn.app.mapreduce.am.command-opts: -Xmx1311m
mapred:yarn.app.mapreduce.am.resource.cpu-wcores: '1'
mapred:yarn.app.mapreduce.am.resource.mb: '1638'
spark:spark.driver.maxResultSize: 512m
spark:spark.driver.memory: 1024m
spark:spark.executor.cores: '1'
spark:spark.executor.instances: '1'
spark:spark.executor.memory: 1255m
spark:spark.executorEnv.OPENBLAS_NUM_THREADS: '1'
spark:spark.extraListeners: com.google.cloud.spark.performance.DataprocMetricsListener
spark:spark.scheduler.mode: FAIR
spark:spark.sql.cbo.enabled: 'true'
spark:spark.ui.port: '0'
spark:spark.yarn.am.memory: 512m
yarn-env:YARN_RESOURCEMANAGER_HEAPSIZE: '400'
yarn-env:YARN_RESOURCEMANAGER_HEAPSIZE: '1024'
yarn-env:YARN_RESOURCEMANAGER_HEAPSIZE: '1024'
yarn:yarn.nodemanager.address: 0.0.0.0:8026
yarn:yarn.nodemanager.resource.cpu-wcores: '2'
yarn:yarn.nodemanager.resource.memory.mb: '2278'
yarn:yarn.resource-manager.nodemanager-graceful-decommission-timeout-secs: '96400'
yarn:yarn.scheduler.maximum-allocation-mb: '2278'
yarn:yarn.scheduler.minimum-allocation-mb: '1'
tempLocation: dataproc-temp-us-west1-791612085972-wkxath3
workerConfig:
  diskConfig:
    bootDiskSizeGb: 30
    bootDiskType: pd-standard
  imageUri: https://www.googleapis.com/compute/v1/projects/cloud-dataproc/global/images/dataproc-2-0-debian10-20221108-035100-rc01
  instanceNames:
    - vkarinje-dplab-w-0
    - vkarinje-dplab-w-1
  machineTypeUri: https://www.googleapis.com/compute/v1/projects/cloud-cs-530-karinje-vkarinje/zones/us-west1-b/machineTypes/e2-medium
  minCpuPlatform: AUTOMATIC
  numInstances: 2
  preemptibility: NON_PREEMPTIBLE
labels:
  goog-dataproc-cluster-name: vkarinje-dplab

```

```

CLOUD SHELL
Terminal (cloud-cs-530-karinje-vkarinje) x +
minCpuPlatform: AUTOMATIC
numInstances: 2
preemptibility: NON_PREEMPTIBLE
labels:
  gcp-dataproc-cluster-name: vkarinje-dplab
  gcp-dataproc-cluster-uid: b7aff0b-f4a2-4af1-bb1b-519decbb1e
  gcp-dataproc-location: us-west1
metrics:
  hdfsMetrics:
    dfs-blocks-corrupt: '0'
    dfs-blocks-missing: '0'
    dfs-blocks-missing-repl-one: '0'
    dfs-blocks-pending-deletion: '0'
    dfs-block-under-replication: '0'
    dfs-capacity-percent: '3388898520'
    dfs-capacity-remaining: '8989815200'
    dfs-capacity-total: '62520839648'
    dfs-capacity-used: '72728'
    dfs-nodes-decommissioning: '0'
    dfs-nodes-running: '2'
  yarnMetrics:
    yarn-apps-completed: '1'
    yarn-apps-failed: '0'
    yarn-apps-killed: '0'
    yarn-apps-pending: '0'
    yarn-apps-running: '0'
    yarn-apps-submitted: '1'
    yarn-containers-allocated: '0'
    yarn-containers-pending: '0'
    yarn-containers-reserved: '0'
    yarn-memory-ab-allocated: '0'
    yarn-memory-ab-available: '6556'
    yarn-memory-ab-pending: '0'
    yarn-memory-ab-reserved: '0'
    yarn-memory-ab-total: '6556'
    yarn-nodes-active: '2'
    yarn-nodes-decommissioned: '0'
    yarn-nodes-lost: '0'
    yarn-nodes-rebooted: '0'
    yarn-nodes-subhealthy: '0'
    yarn-volumes-allocated: '0'
    yarn-volumes-available: '4'
    yarn-volumes-pending: '0'
    yarn-volumes-reserved: '0'
    yarn-volumes-total: '4'
projectId: cloud-cs-530-karinje-vkarinje
status:

```

Allocate two additional pre-emptible machines to the cluster

```

projectId: cloud-cs-530-karinje-vkarinje
status:
  state: RUNNING
  stateStartTime: '2022-11-28T09:21:13.468178Z'
statusHistory:
- state: CREATING
  stateStartTime: '2022-11-28T09:18:25.341759Z'
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud dataproc clusters update ${CLUSTERNAME} --num-secondary-workers=2
Waiting on operation [projects/cloud-cs-530-karinje-vkarinje/regions/us-west1/operations/1a258965-66a2-3cb5-94b9-a2457dec0411].
Waiting for cluster update operation...done.
Updated [https://dataproc.googleapis.com/v1/projects/cloud-cs-530-karinje-vkarinje/regions/us-west1/clusters/vkarinje-dplab].
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ ||

```

Repeat the listing to see that they show up in the Config section.

```

vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud dataproc clusters describe $(CLUSTERNAME)
clusterName: vkarinje-dplab
clusterUuid: b7aff99b-f4a3-4af1-bb1b-519d1ec5bb1c
config:
  configBucket: dataproc-staging-us-west1-791612085972-2gjs6s95
  endpointConfig: {}
  gceClusterConfig:
    internalIpOnly: false
    networkUri: https://www.googleapis.com/compute/v1/projects/cloud-cs-530-karinje-vkarinje/global/networks/default
    serviceAccountScopes:
      - https://www.googleapis.com/auth/cloud-platform
      - https://www.googleapis.com/auth/cloud.useraccounts.readonly
      - https://www.googleapis.com/auth/devstorage.read_write
      - https://www.googleapis.com/auth/logging.write
    tags:
      - codelab
    zoneUri: https://www.googleapis.com/compute/v1/projects/cloud-cs-530-karinje-vkarinje/zones/us-west1-b
  masterConfig:
    diskConfig:
      bootDiskSizeGb: 30
      bootDiskType: pd-standard
    imageUri: https://www.googleapis.com/compute/v1/projects/cloud-dataproc/global/images/dataproc-2-0-deb10-20221108-035100-rc01
    instanceNames:
      - vkarinje-dplab-m
    machineTypeUri: https://www.googleapis.com/compute/v1/projects/cloud-cs-530-karinje-vkarinje/zones/us-west1-b/machineTypes/e2-medium
    minCpuPlatform: AUTOMATIC
    numInstances: 1
    preemptibility: NON_PREEMPTIBLE
  secondaryWorkerConfig:
    diskConfig:
      bootDiskSizeGb: 30
      bootDiskType: pd-standard
    imageUri: https://www.googleapis.com/compute/v1/projects/cloud-dataproc/global/images/dataproc-2-0-deb10-20221108-035100-rc01
    instanceNames:
      - vkarinje-dplab-sw-n2n7
      - vkarinje-dplab-sw-t5fw
    isPreemptible: true
    machineTypeUri: https://www.googleapis.com/compute/v1/projects/cloud-cs-530-karinje-vkarinje/zones/us-west1-b/machineTypes/e2-medium
    managedGroupConfig:
      instanceGroupManagerName: dataproc-vkarinje-dplab-sw
      instanceTemplateName: dataproc-vkarinje-dplab-sw
      minCpuPlatform: AUTOMATIC
      numInstances: 2
      preemptibility: PREEMPTIBLE
  softwareConfig:
    imageVersion: 2.0.51-debian10
  properties:
    capacity-scheduler:yarn.scheduler.capacity.root.default.ordering-policy: fair

```

Then, visit Compute Engine to see the new nodes in the cluster.

The screenshot shows the Google Cloud Console interface. On the left, the 'Compute Engine' menu is open, showing 'VM instances' selected. The main panel displays a table of VM instances for the cluster 'vkarinje-dplab' in the 'us-west1-b' zone. The table has columns for Status, Name, Zone, Recommendations, In use by, Internal IP, External IP, and Connect. There are five instances listed: 'vkarinje-dplab-m', 'vkarinje-dplab-sw-n2n7', 'vkarinje-dplab-sw-t5fw', 'vkarinje-dplab-sw-0', and 'vkarinje-dplab-sw-1'. All instances are in a 'Running' state. Below the table, there are 'Related actions' such as 'Explore Backup and DR', 'View billing report', 'Monitor VMs', 'Explore VM logs', 'Set up firewall rules', and 'Patch management'. On the right, a 'Select an instance' panel is visible with tabs for 'PERMISSIONS', 'LABELS', and 'MONITORING'. A message at the bottom of this panel says 'Please select at least one resource.'

## Run computation again

For your lab notebook:

- How long did the job take to execute? How much faster did it take?

```

vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ date
Tue 29 Nov 2022 05:37:58 AM UTC
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud dataproc jobs submit spark --cluster ${CLUSTERNAME} \
--class org.apache.spark.examples.SparkPi \
--jars file:///usr/lib/spark/examples/jars/spark-examples.jar -- 1000 \
>& output2.txt &
[1] 1570
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud dataproc jobs list --cluster ${CLUSTERNAME}
JOB_ID: df53e4917fbc42e5bc8b8f9fd5d757e0
TYPE: spark
STATUS: SETUP_DONE

JOB_ID: a58b8dd8a69a48dd91b9174642fcf342
TYPE: spark
STATUS: DONE

JOB_ID: ee40011cf0ef40ed9cc5a5dd61b593f1
TYPE: spark
STATUS: DONE
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud dataproc jobs list --cluster ${CLUSTERNAME}
JOB_ID: df53e4917fbc42e5bc8b8f9fd5d757e0
TYPE: spark
STATUS: SETUP_DONE

JOB_ID: a58b8dd8a69a48dd91b9174642fcf342
TYPE: spark
STATUS: DONE

JOB_ID: ee40011cf0ef40ed9cc5a5dd61b593f1
TYPE: spark
STATUS: DONE
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud dataproc jobs list --cluster ${CLUSTERNAME}
JOB_ID: df53e4917fbc42e5bc8b8f9fd5d757e0
TYPE: spark
STATUS: SETUP_DONE

JOB_ID: a58b8dd8a69a48dd91b9174642fcf342
TYPE: spark

```

```

JOB_ID: ee40011cf0ef40ed9cc5a5dd61b593f1
TYPE: spark
STATUS: DONE
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud dataproc jobs list --cluster ${CLUSTERNAME}
JOB_ID: df53e4917fbc42e5bc8b8f9fd5d757e0
TYPE: spark
STATUS: SETUP_DONE

JOB_ID: a58b8dd8a69a48dd91b9174642fcf342
TYPE: spark
STATUS: DONE

JOB_ID: ee40011cf0ef40ed9cc5a5dd61b593f1
TYPE: spark
STATUS: DONE
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud dataproc jobs list --cluster ${CLUSTERNAME}
JOB_ID: df53e4917fbc42e5bc8b8f9fd5d757e0
TYPE: spark
STATUS: SETUP_DONE

JOB_ID: a58b8dd8a69a48dd91b9174642fcf342
TYPE: spark
STATUS: DONE

JOB_ID: ee40011cf0ef40ed9cc5a5dd61b593f1
TYPE: spark
STATUS: DONE
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ date
Tue 29 Nov 2022 05:38:35 AM UTC

```

Ans: It just took 37 seconds for the job to execute. It was 1 min and 45 seconds faster.

- Examine `output2.txt` and show the estimate of  $\pi$  calculated.

Ans: Pie is roughly 3.141625191416252







```

11 Welcome to Cloud Shell
12 is_popular.py
13 home > vkaranje > training-data-analyst > courses > data_analysis > lab2 > python > is_popular.py
14
15 if __name__ == '__main__':
16     parser = argparse.ArgumentParser(description='Find the most used Java packages')
17     parser.add_argument('--output_prefix', default='/tmp/output', help='Output prefix')
18     parser.add_argument('--input', default='../javahelp/src/main/java/com/google/cloud/training/dataanalyst/javahelp/', help='Input directory')
19
20     options, pipeline_args = parser.parse_known_args()
21     p = beam.Pipeline(argv=pipeline_args)
22
23     input = '{}*.java'.format(options.input)
24     output_prefix = options.output_prefix
25     keyword = 'import'
26
27     # find most used packages
28     (p
29      | 'GetJava' >> beam.io.ReadFromText(input)
30      | 'GetImports' >> beam.FlatMap(lambda line: startswith(line, keyword))
31      | 'PackageUse' >> beam.FlatMap(lambda line: packageUse(line, keyword))
32      | 'TotalUse' >> beam.CombinePerKey(sum)
33      | 'Top_5' >> beam.transforms.combiners.TopOf(5, key=lambda kv: kv[1])
34      | 'write' >> beam.io.WriteToText(output_prefix)
35     )
36
37     p.run().wait_until_finish()
38
39

```

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

Ans: Using the given line and keyword we iterate and find the package names using `getPackages` and `splitPackageName` functions. The `PackageUse` transform implements the yield operation. We then iterate over the generator that is returned by the `PackageUse` function to process each package.

- Look up Beam's `CombinePerKey`. What operation does the `TotalUse` operation implement?

Ans: The Beam's `CombinePerKey` function does combine all the elements for each key in a collection. For example, we can pass a function, `sum` which takes an iterable and sums up elements based on keys.

Answer the following question for your lab notebook.

- Which operations correspond to a "Map"?

Ans: `GetImports`, `PackageUse`-`beam.FlatMap`

- Which operation corresponds to a "Shuffle-Reduce"?

Ans: `TotalUse`- `beam.CombinePerKey`

- Which operation corresponds to a "Reduce"?

Ans: `Top_5`-`beam.transforms.combiners.Top.Of`

## Run pipeline locally

- Take a screenshot of its contents

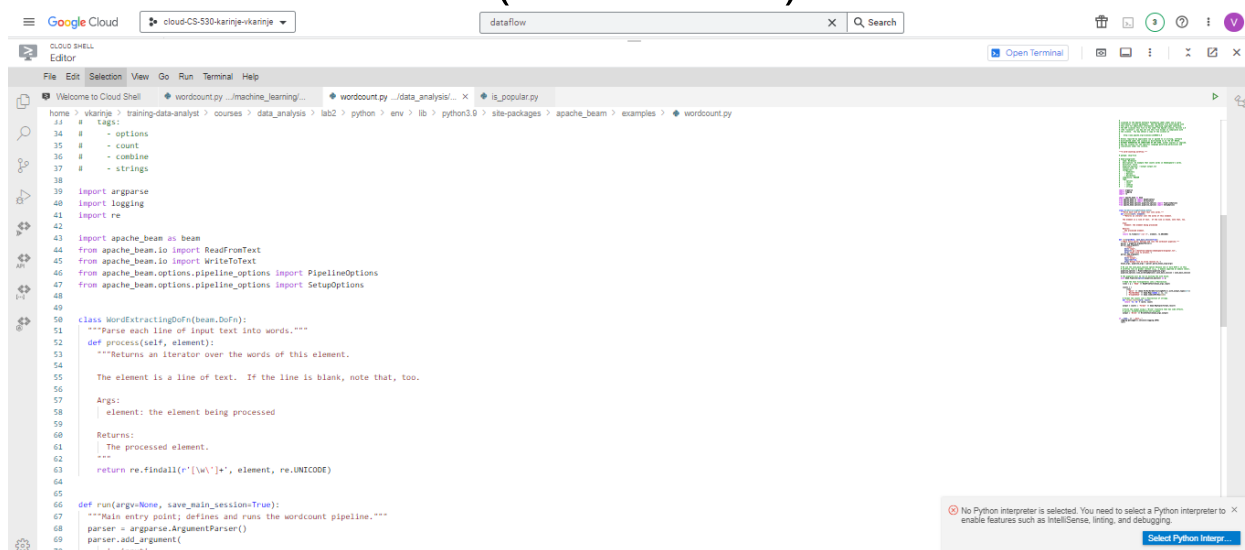
```
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje)$ edit is_popular.py
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje)$ python is_popular.py
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje)$ cd /tmp
(env) vkarinje@cloudshell:/tmp (cloud-cs-530-karinje-vkarinje)$ cat 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) vkarinje@cloudshell:/tmp (cloud-cs-530-karinje-vkarinje)$ ||
```

- Explain what the data in this output file corresponds to based on your understanding of the program.

Ans: The data in this output file corresponds to the most used packages in the java files present in the directory

```
../javahelp/src/main/java/com/google/cloud/training/dataanalyst/javahelp/'
```

## 14. Dataflow Lab #2 (Word count)



- What are the names of the stages in the pipeline?

Ans: The names of the stages in pipeline are Read, Split, PairWithOne, GroupAndSum, Format and Write

The screenshot shows the Google Cloud Shell Editor interface. The top bar includes the Google Cloud logo, a dropdown menu showing 'cloud-CS-530-karinje-vkarinje', and a search bar with 'dataflow'. The editor window displays a Python script for word counting using Apache Beam. The script is titled 'wordcount.py' and is located at '.../machine\_learning/...'. The script includes comments and code for reading input, splitting lines, extracting words, counting them, and writing the output. The script is as follows:

```

186 # The pipeline will be run on exiting the with block.
187 with beam.Pipeline(options=pipeline_options) as p:
188
189     # Read the text file[pattern] into a PCollection.
190     lines = p | 'Read' >> ReadFromText(known_args.input)
191
192     counts = (
193         lines
194         | 'Split' >> (beam.ParDo(WordExtractingDoFn()).with_output_types(str))
195         | 'PairWithOne' >> beam.Map(lambda x: (x, 1))
196         | 'GroupAndSum' >> beam.CombinePerKey(sum))
197
198     # Format the counts into a PCollection of strings.
199     def format_result(word, count):
200         return '%s: %d' % (word, count)
201
202     output = counts | 'Format' >> beam.MapTuple(format_result)
203
204     # Write the output using a "Write" transform that has side effects.
205     # pylint: disable-expression-not-assigned
206     output | 'Write' >> WriteToText(known_args.output)
207
208
209 if __name__ == '__main__':
210     logging.getLogger().setLevel(logging.INFO)
211     run()
212

```

- Describe what each stage does.

Ans: The Read stage reads and processes the input file in our default case

“gs://dataflow-samples/shakespeare/kinglear.txt”

The Split stage invokes the function process of the class WordExtractingDoFn . The process function for each element is run parallelly by the beam and returns any word element from it. The Pair with One stage, uses the map function which performs a mapping action to map a word string to (word,1).

The Group and Sum stage invokes the sum function to combine and group the words.

Finally, the Write stage prints the output.

Format stage does the formatting.

## Run code locally

- Use `wc` with an appropriate flag to determine the number of unique words in King Lear.

```

vkari@jupyterlab:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkari)$ source ~/env/activate
(env) vkari@jupyterlab:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkari)$ edit env/lib/python2.7/site-packages/apache_beam/examples/wordcount.py
(env) vkari@jupyterlab:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkari)$ python -m apache_beam.examples.wordcount \
--output outputs
INFO root:Running pipeline option (runners). Executing pipeline using the default runner: DirectRunner.
INFO apache_beam.internal.pyg_auth:Setting socket default timeout to 60 seconds.
INFO root:Default Python HLE lease for environment is apache/beam.pythond3.9.ssh:2.43.0
INFO apache_beam.runners.portability.fn_api_runner.translations: <function annotate_demonstrate_side_inputs at 0x7f592b67b0b>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function fix_side_input_parallel_orders at 0x7f592b67ca0>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function pack_combiners at 0x7f592b681f0>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function lift_combiners at 0x7f592b68580>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function expand_sdf at 0x7f592b68430>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function expand_gbk at 0x7f592b684c0>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function sink_elements at 0x7f592b685e0>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function greedily_flow at 0x7f592b68670>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function read_to_impulse at 0x7f592b69700>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function impulse_to_impulse at 0x7f592b69790>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function sort_stages at 0x7f592b69840>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function add_impulse_to_sampling_transform at 0x7f592b69af0>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function setup_timer_sampling at 0x7f592b69940>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function populate_data_channel_codes at 0x7f592b69a60>
INFO apache_beam.runners.worker.statecache:Creating state cache with size 104857600
INFO apache_beam.runners.portability.fn_api_runner.worker_handlers:Created Worker handler <apache_beam.runners.portability.fn_api_runner.worker_handlers.DatabricksWorkerHandler object at 0x7f592b297ad0> for environment ref.EnvironmentDefaultEnvironment_1 (on
an env:embedded python:v1. h'')
INFO apache_beam.io.filesystems:Starting finalizer write threads with num_shards: 1 (skipped: 0), batches: 1, num_threads: 1
INFO apache_beam.io.filesystems:Renamed 1 shards in 0.00 seconds.
(env) vkari@jupyterlab:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkari)$ mv wc
(env) vkari@jupyterlab:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkari)$ wc -w outputs-00000-of-00001
5068 outputs-00000-of-00001
(env) vkari@jupyterlab:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkari)$ ||

```

There are 9568 unique words in King Lear.

- Use sort with appropriate flags to perform a *numeric* sort on the *key field* containing the count for each word in *descending* order. Pipe the output into `head` to show the top 3 words in King Lear and the number of times they appear

```

~/env/outputs-00000-of-00001
(env) vkari@jupyterlab:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkari)$ sort -k2nr outputs-00000-of-00001 | head -3
the: 786
I: 622
and: 594
(env) vkari@jupyterlab:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkari)$ ||

```

The top 3 words in King Lear are the, I & and

the-786

I- 622

and- 594

Find a place in the pipeline that you can insert a stage that transforms all of the characters it receives into lowercase.

```

69 parser.add_argument(
70     '--input',
71     dest='input',
72     default='gs://dataflow-samples/shakespeare/kinglear.txt',
73     help='Input file to process.')
74 parser.add_argument(
75     '--output',
76     dest='output',
77     required=True,
78     help='Output file to write results to.')
79 known_args, pipeline_args = parser.parse_known_args(argv)
80
81 # We use the save_main_session option because one or more DoFn's in this
82 # workflow rely on global context (e.g., a module imported at module level).
83 pipeline_options = PipelineOptions(pipeline_args)
84 pipeline_options.view_as(SetupOptions).save_main_session = save_main_session
85
86 # The pipeline will be run on exiting the with block.
87 with beam.Pipeline(options=pipeline_options) as p:
88
89     # Read the text file[pattern] into a PCollection.
90     lines = p | 'Read' >> ReadFromText(known_args.input)
91
92     counts = (
93         lines
94         | 'lowercase' >> beam.Map(lambda x: x.lower())
95         | 'Split' >> (beam.ParDo(WordExtractingDoFn()).with_output_types(str))
96         | 'PairWithOne' >> beam.Map(lambda x: (x, 1))
97         | 'GroupAndSum' >> beam.CombinePerKey(sum))
98
99     # Format the counts into a PCollection of strings.
100     def format_result(word, count):
101         return '%s: %d' % (word, count)
102
103     output = counts | 'Format' >> beam.MapTuple(format_result)
104
105     # Write the output using a "Write" transform that has side effects.
106     output |>> beam.Map(lambda x: x)

```

Perform the following and show a screenshot of the results in your lab notebook:

- Use the previous method to show the top 3 words in King Lear, case-insensitive, and the number of times they appear.

Ans:

```

(vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdiver/04_features/dataflow/python) (cloud-cs-530-karinje-vkarinje) edit env/lib/python3.8/site-packages/apache_beam/examples/wordcount.py
(vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdiver/04_features/dataflow/python) (cloud-cs-530-karinje-vkarinje) python -m apache_beam.examples.wordcount --output outputs
INFO root:Default Python REPL env for environment is apache/beam:python3.8-2.42.0
INFO apache_beam.runners.portability.fn_api_runner.translations: <function annotate_downstream_side_inputs at 0x7f444f21d30>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function fix_side_input_pull_order at 0x7f444f21d60>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function pool_combiners at 0x7f444f21d90>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function life_combiners at 0x7f444f21da0>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function expand_self at 0x7f444f21db0>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function expand_gbk at 0x7f444f21dc0>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function sink_flattens at 0x7f444f21dd0>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function greedily_fuse at 0x7f444f21de0>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function read_to_impulse at 0x7f444f21df0>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function impulse_to_input at 0x7f444f21e00>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function sort_stager at 0x7f444f21e10>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function add_impulse_to_dangling_transforms at 0x7f444f21e20>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function setup_timer_mapping at 0x7f444f21e30>
INFO apache_beam.runners.portability.fn_api_runner.translations: <function populate_data_channel_order at 0x7f444f21e40>
INFO apache_beam.runners.portability.fn_api_runner.worker_handlers: Created Worker handler CapableBeamRunnersPortabilityFnApiRunnerWorkerHandler object at 0x7f444f21e50 for environment ref_environment_default_environment_1 (be
am-env:embedded-python:v1.1.0)
INFO apache_beam.io.filebasedsink: Starting finalise write threads with num_shards: 1 (skipped: 0), batches: 1, num_threads: 1
INFO apache_beam.io.filebasedsink: Finished 1 shards in 0.00 seconds
(vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdiver/04_features/dataflow/python) (cloud-cs-530-karinje-vkarinje) sort -k2nr outputs-0000-of-0001 | head -3
the: 908
and: 738
i: 622
(vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdiver/04_features/dataflow/python) (cloud-cs-530-karinje-vkarinje)

```

The top 3 words in King Lear, case-insensitive are the, and & i

the -908

and -738

i -622

## Setup for Cloud Dataflow

```
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje) $ gcloud services enable dataflow compute_component storage_component storage_api
Operation "operations/act_p2-791612085972-6493282-e897-4078-9559-12ac5291ed94" finished successfully.
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje) $ export BUCKET=$(GOOGLE_CLOUD_PROJECT)
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje) $ export REGION=us-west1
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje) $ gsutil mb gs://$BUCKET
Creating gs://cloud-cs-530-karinje-vkarinje/...
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje) $ ||
```

## Service account setup

```
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje) $ cd
(env) vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje) $ gcloud iam service-accounts create df-lab
Created service account [df-lab].
(env) vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje) $ gcloud projects add-iam-policy-binding $(GOOGLE_CLOUD_PROJECT) \
--member serviceAccount:df-lab@$(GOOGLE_CLOUD_PROJECT).iam.gserviceaccount.com \
--role roles/dataflow.admin
Updated IAM policy for project [cloud-cs-530-karinje-vkarinje].
bindings:
- members:
  - serviceAccount:service-791612085972@gcp-sa-aiplatform.iam.gserviceaccount.com
    role: roles/aiplatform.serviceAgent
- members:
  - serviceAccount:service-791612085972@gcp-gae-service.iam.gserviceaccount.com
    role: roles/appengine.serviceAgent
- members:
  - serviceAccount:cs430jupyter@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com
    role: roles/bigquery.user
- members:
  - serviceAccount:791612085972@cloudbuild.gserviceaccount.com
    role: roles/cloudbuild.builds.builder
- members:
  - serviceAccount:service-791612085972@gcp-sa-cloudbuild.iam.gserviceaccount.com
    role: roles/cloudbuild.serviceAgent
- members:
  - serviceAccount:service-791612085972@gcf-admin-robot.iam.gserviceaccount.com
    role: roles/cloudfunctions.serviceAgent
```

```
Version: 1
(env) vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje) $ gcloud projects add-iam-policy-binding $(GOOGLE_CLOUD_PROJECT) \
--member serviceAccount:df-lab@$(GOOGLE_CLOUD_PROJECT).iam.gserviceaccount.com \
--role roles/dataflow.worker
Updated IAM policy for project [cloud-cs-530-karinje-vkarinje].
bindings:
- members:
  - serviceAccount:service-791612085972@gcp-sa-aiplatform.iam.gserviceaccount.com
    role: roles/aiplatform.serviceAgent
- members:
  - serviceAccount:service-791612085972@gcp-gae-service.iam.gserviceaccount.com
    role: roles/appengine.serviceAgent
- members:
  - serviceAccount:cs430jupyter@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com
    role: roles/bigquery.user
- members:
  - serviceAccount:791612085972@cloudbuild.gserviceaccount.com
    role: roles/cloudbuild.builds.builder
- members:
  - serviceAccount:service-791612085972@gcp-sa-cloudbuild.iam.gserviceaccount.com
    role: roles/cloudbuild.serviceAgent
- members:
  - serviceAccount:service-791612085972@gcf-admin-robot.iam.gserviceaccount.com
    role: roles/cloudfunctions.serviceAgent
- members:
  - serviceAccount:service-791612085972@gcp-sa-cloudscheduler.iam.gserviceaccount.com
    role: roles/cloudscheduler.serviceAgent
- members:
  - serviceAccount:service-791612085972@compute-system.iam.gserviceaccount.com
    role: roles/compute.serviceAgent
- members:
  - deleted:serviceAccount:gcs-lab@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com?uid=102296725729513422801
    role: roles/compute.viewer
- members:
```

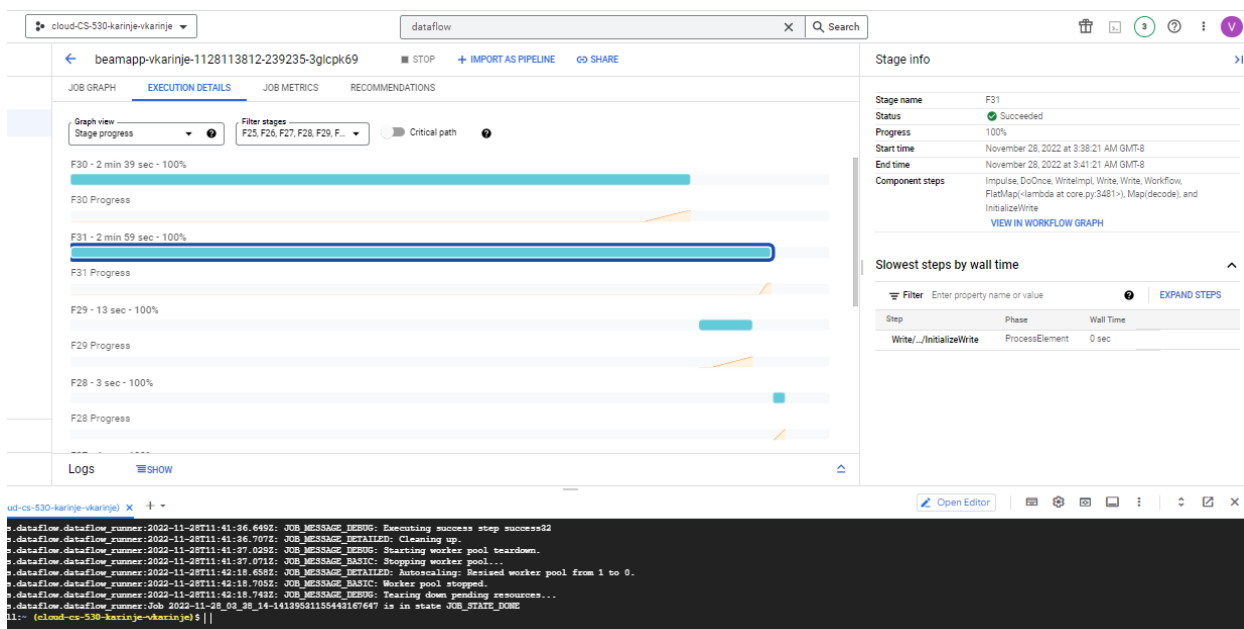
```
(env) vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje) $ gcloud iam service-accounts keys create df-lab.json --iam-account df-lab@$(GOOGLE_CLOUD_PROJECT).iam.gserviceaccount.com
created key [0c57fa0ac575b3099da3ce2c18316a9f509515da] of type [json] as [df-lab.json] for [df-lab@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com]
(env) vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje) $ ||
```

## Run code using Dataflow runner

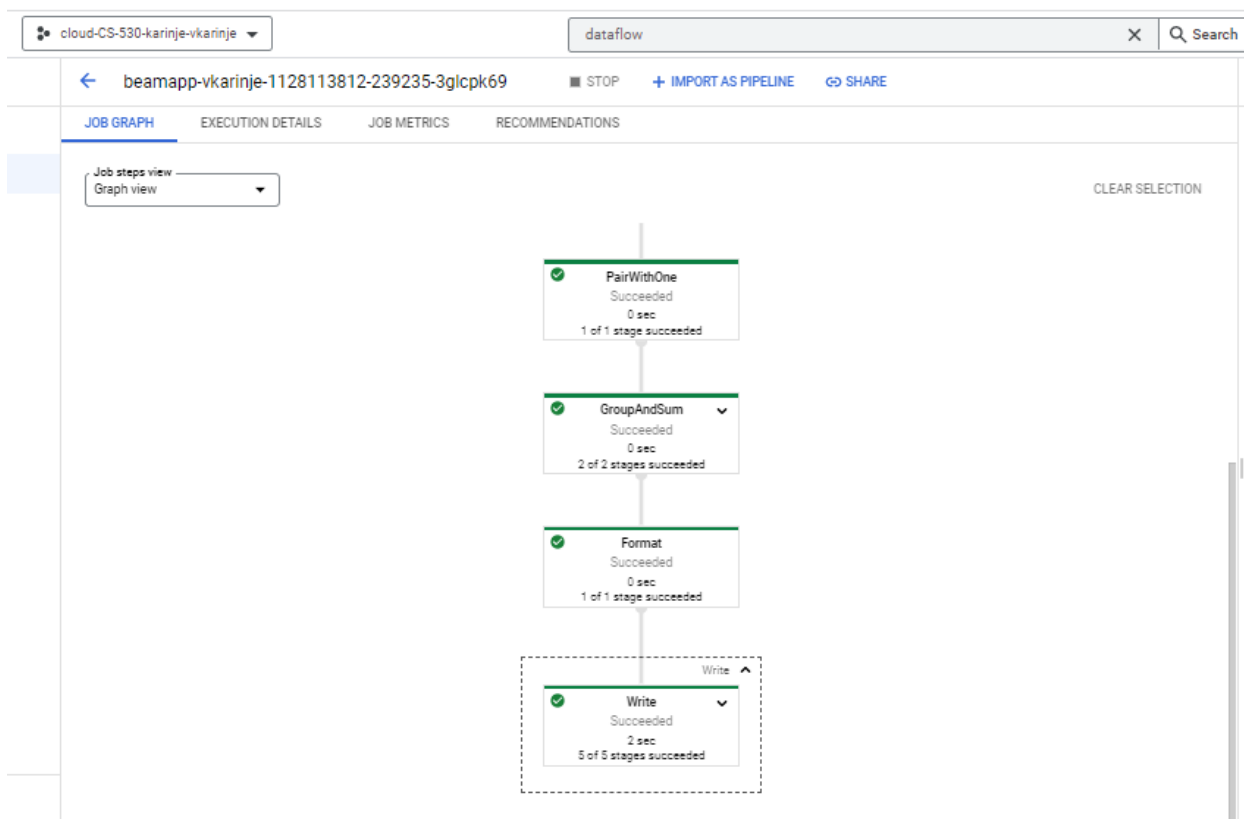
Include the following in your lab notebook:

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

Ans: F31

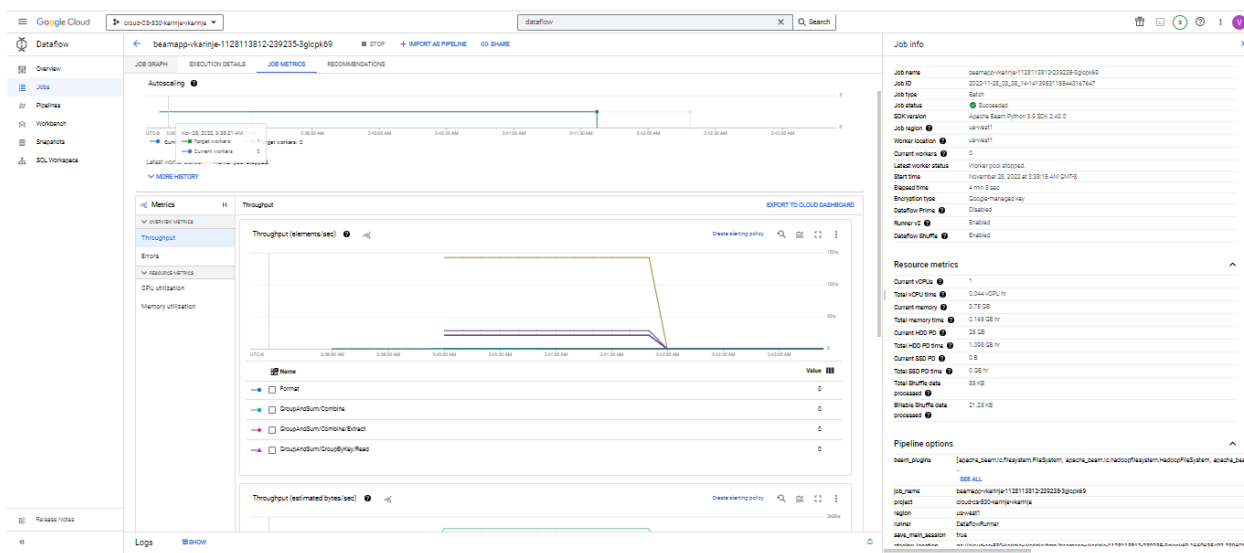


Write has taken the longest time to complete





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



- Examine the output directory in Cloud Storage. How many files has the final write stage in the pipeline created?

The screenshot displays the Google Cloud Storage console for a bucket named 'cloud-cs-530-karinje-vkarinje'. The bucket contains a single file named 'outputs-00000-e4-00001' with a size of 42.7 KB, created on November 28, 2022, at 9:41:35 AM.

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Retention expiration date	Hide
outputs-00000-e4-00001	42.7 KB	text/plain	Nov 28, 2022, 9:41:35 AM	Standard	Nov 28, 2022, 9:41:35 AM	Not public	—	Google-managed key	—	None

Ans: Only one file has been created by the write stage in the pipeline

## Clean up

```

deleted service account [df-lab@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com].
- serviceAccount: guestbook@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com
  role: roles/storage.objectViewer
  etag: BwKuhpd2i14=
  version: 1
(env) vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje) $ gcloud iam service-accounts delete df-lab@$(GOOGLE_CLOUD_PROJECT).iam.gserviceaccount.com
You are about to delete service account [df-lab@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com].

Do you want to continue (Y/n)? Y

deleted service account [df-lab@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com]
(env) vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje) $ ||

```

```

deleted service account [df-lab@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com]
(env) vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje) $ gcloud -a rm -r gs://$(BUCKET)
Removing gs://cloud-cs-530-karinje-vkarinje/results/outputs-00000-of-00001#1669635492574421...
Removing gs://cloud-cs-530-karinje-vkarinje/tmp/beamapp-vkarinje-1128113812-239235-3glcptk69.1669635492.239601/pickled_main_session#1669635492574421...
Removing gs://cloud-cs-530-karinje-vkarinje/tmp/beamapp-vkarinje-1128113812-239235-3glcptk69.1669635492.239601/apache_beam-2.43.0-cp39-manylinux_2_17_x86_64.whl#1669635492574421...
Removing gs://cloud-cs-530-karinje-vkarinje/tmp/beamapp-vkarinje-1128113812-239235-3glcptk69.1669635492.239601/dataflow_python_sdk.tar#1669635492574421...
Removing gs://cloud-cs-530-karinje-vkarinje/tmp/beamapp-vkarinje-1128113812-239235-3glcptk69.1669635492.239601/tmp-5416e3e582270c8d-00000-of-00001.sdfmeta#1669635492574421...
/ [6/6 objects] 100% Done
Operation completed over 6 objects.
Removing gs://cloud-cs-530-karinje-vkarinje/...
(env) vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje) $ ||

```



09.2g: BigQuery, JupyterLab