CS 530 INTERNET WEB AND CLOUD SYSTEMS

Name : Varsha Karinje
PSU ID: 925923534

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-optimized-arco-era5.endpoints.bigquery-public-data.cloud.goog
TITLE: Analysis-Ready, Cloud Optimized (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: bqdts-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
```

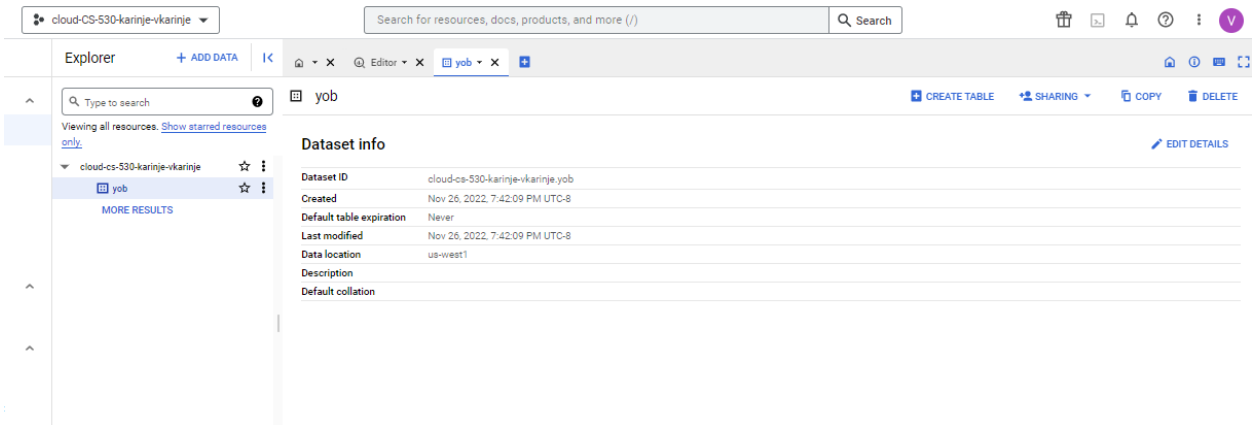
```
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud services enable  bigqueryconnection.googleapis.com
Operation "operations/acat.p2-791612085972-ec2ebed1-9ec2-4c83-91a7-116c1c0841ca" finished successfully.
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud services enable  bigquery.googleapis.com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

Examine dataset

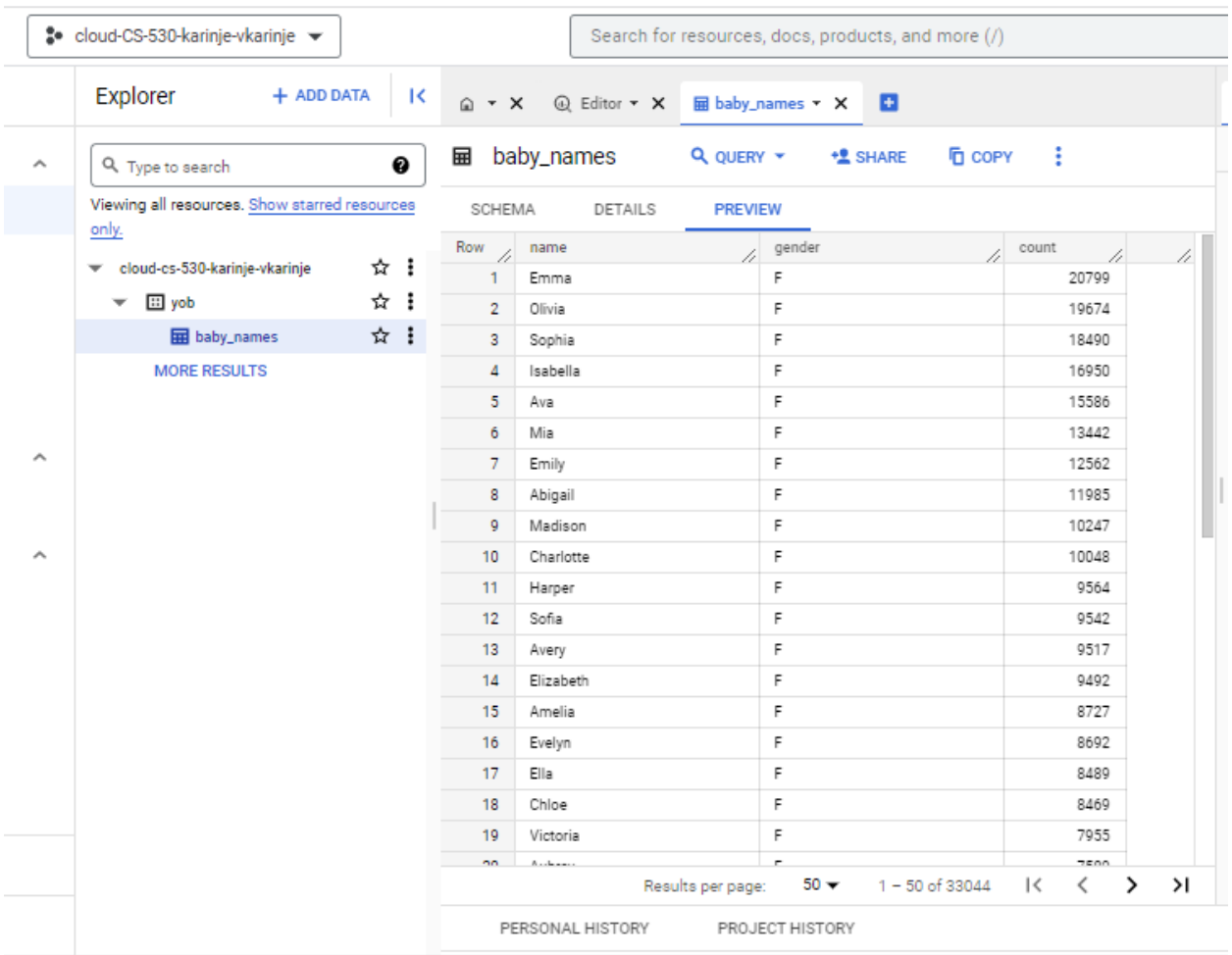
```
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ wget https://thefengs.com/wuchang/courses/cs430/yob2014.txt
--2022-11-27 03:34:52--  https://thefengs.com/wuchang/courses/cs430/yob2014.txt
Resolving thefengs.com (thefengs.com)... 131.252.220.66
Connecting to thefengs.com (thefengs.com)|131.252.220.66|:443... 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 03:34:52 (11.5 MB/s) - 'yob2014.txt' saved [425485/425485]

vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ head -3 yob2014.txt
Emma,F,20799
Olivia,F,19674
Sophia,F,18490
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ wc -l yob2014.txt
33044 yob2014.txt
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$
```

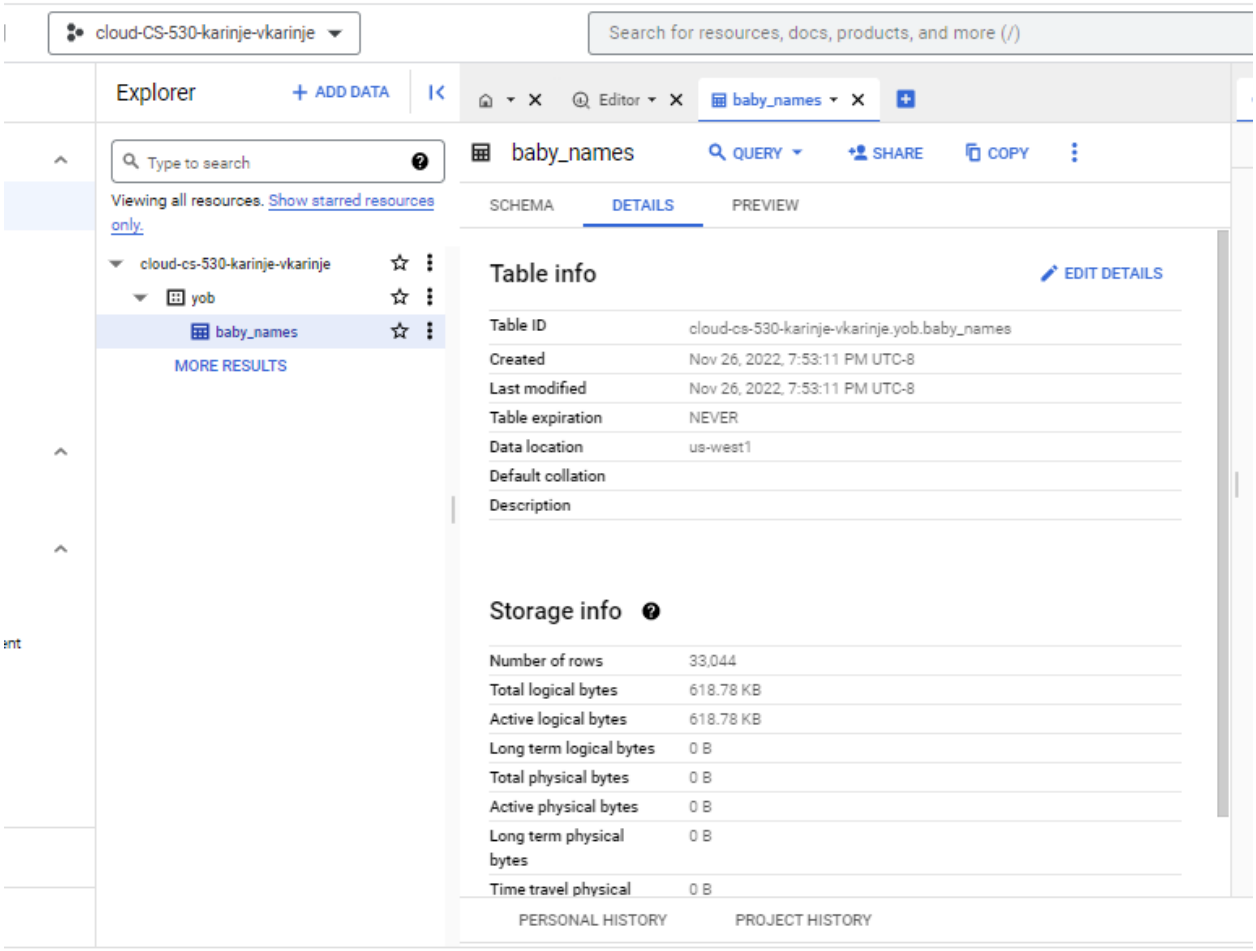
```
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gsutil mb -l us-west1 gs://vkarinje-bucket
Creating gs://vkarinje-bucket/...
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gsutil cp yob2014.txt gs://vkarinje-bucket/
Copying file://yob2014.txt [Content-Type=text/plain]...
/ [1 files][415.5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

Create dataset



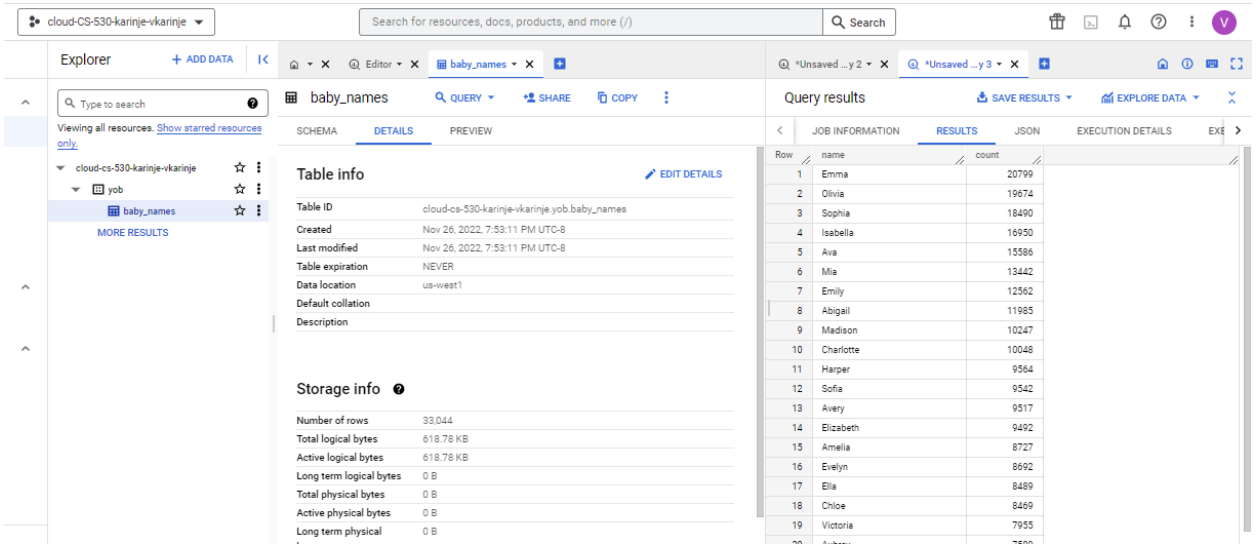
- Take a screenshot of the table's details that includes the number of rows in the table.

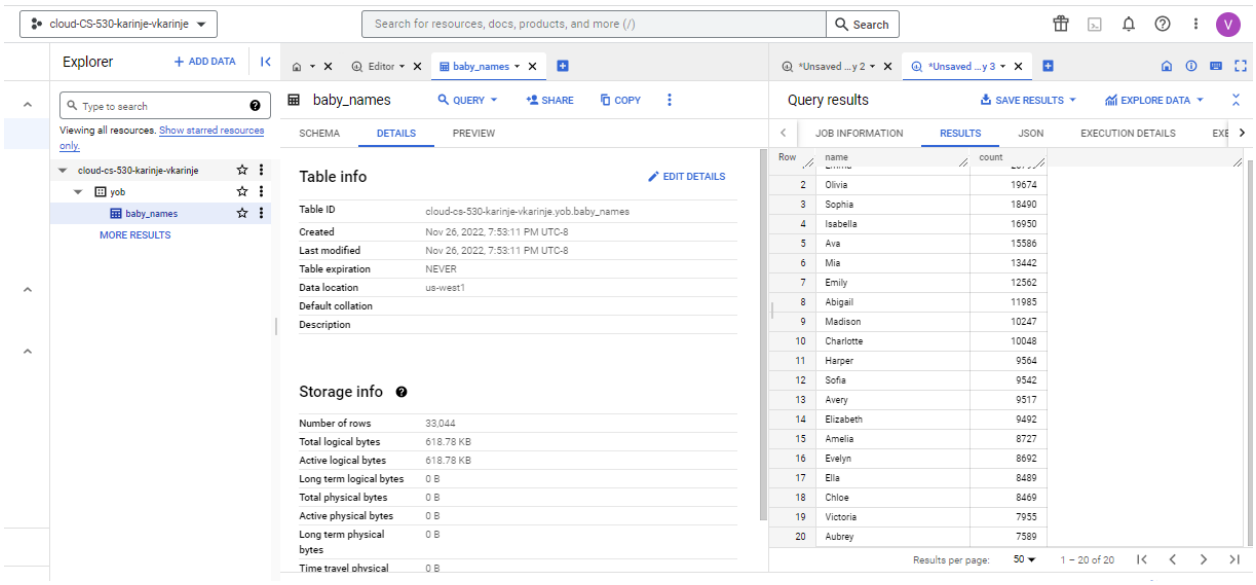




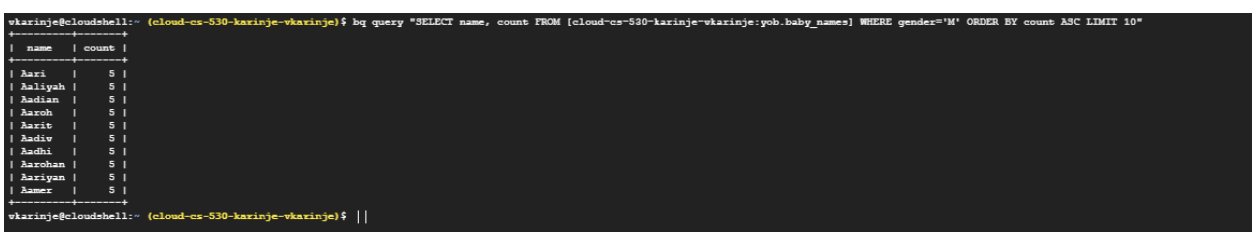
Query data

- Screenshot your results and include it in your lab notebook





- 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 |
| Aarit   |     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 |
| Aaliyah |     5 |
| Aadian  |     5 |
| Aaroh   |     5 |
| Aarit   |     5 |
| Aadiv   |     5 |
| Aadhi   |     5 |
| Aarohan |     5 |
| Aariyan |     5 |
| Aamer   |     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

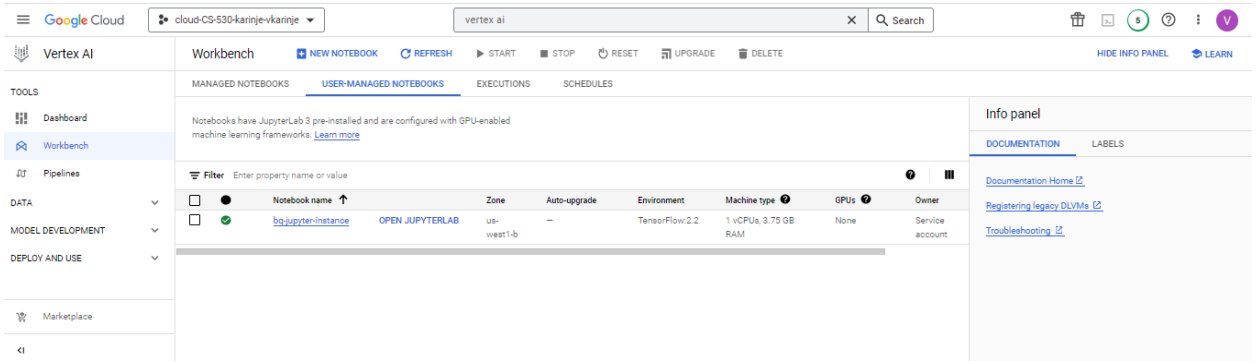
```
cloud-cs-530-karinje-vkarinje> SELECT name, count from [cloud-cs-530-karinje-vkarinje:yob.baby_names] WHERE name='Varsha'
+--------+-------+
|  name  | count |
+--------+-------+
| Varsha |    20 |
+--------+-------+
cloud-cs-530-karinje-vkarinje>
```

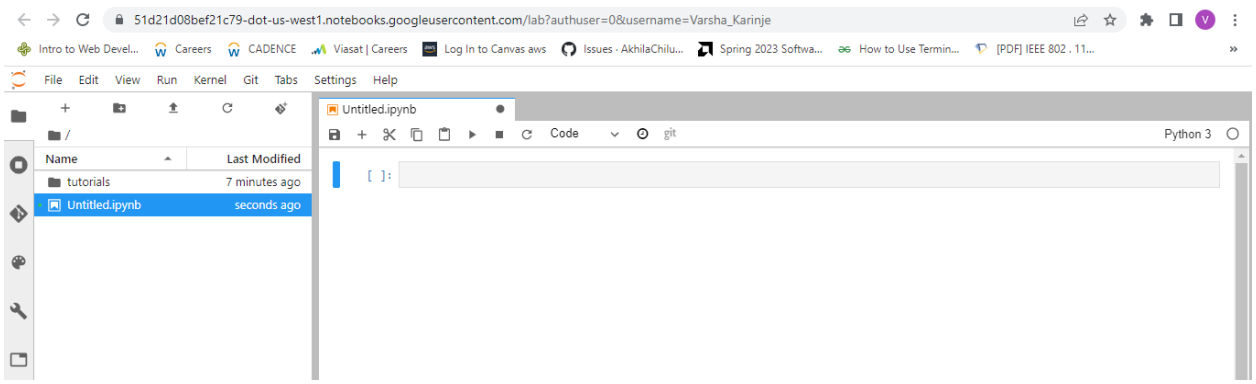
BigQuery, Notebooks Lab #2 (Natality)

```
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/acat.p2-791612085972-56b67571-9f13-461e-a6cf-03cd7cd1e0a1" 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.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:
  - 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:cs420flex@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com?uid=114138103453650149292
  - serviceAccount:cloud-cs-530-karinje-vkarinje@appspot.gserviceaccount.com
```

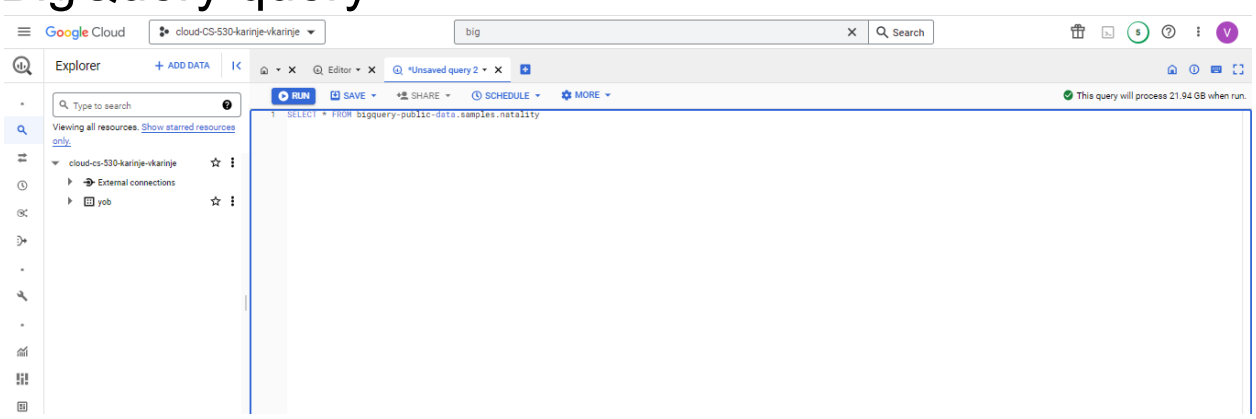
```
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud notebooks instances create bq-jupyter-instance \
  --vm-image-project=deeplearning-platform-release \
  --vm-image-family=tf2-2-2-cpu \
  --machine-type=n1-standard-1 \
  --location=us-west1-b \
  --service-account=cs430jupyter@${GOOGLE_CLOUD_PROJECT}.iam.gserviceaccount.com
Waiting for operation on Instance [bq-jupyter-instance] to be created with [projects/cloud-cs-530-karinje-vkarinje/locations/us-west1-b/operations/operation-1669539588235-5ee6ffa1860a4-024157a7-42dc51fa]...done.
Created notebooks instance bq-jupyter-instance [https://notebooks.googleapis.com/v1/projects/cloud-cs-530-karinje-vkarinje/locations/us-west1-b/operations/operation-1669539588235-5ee6ffa1860a4-024157a7-42dc51fa].
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$
```

Launch notebook



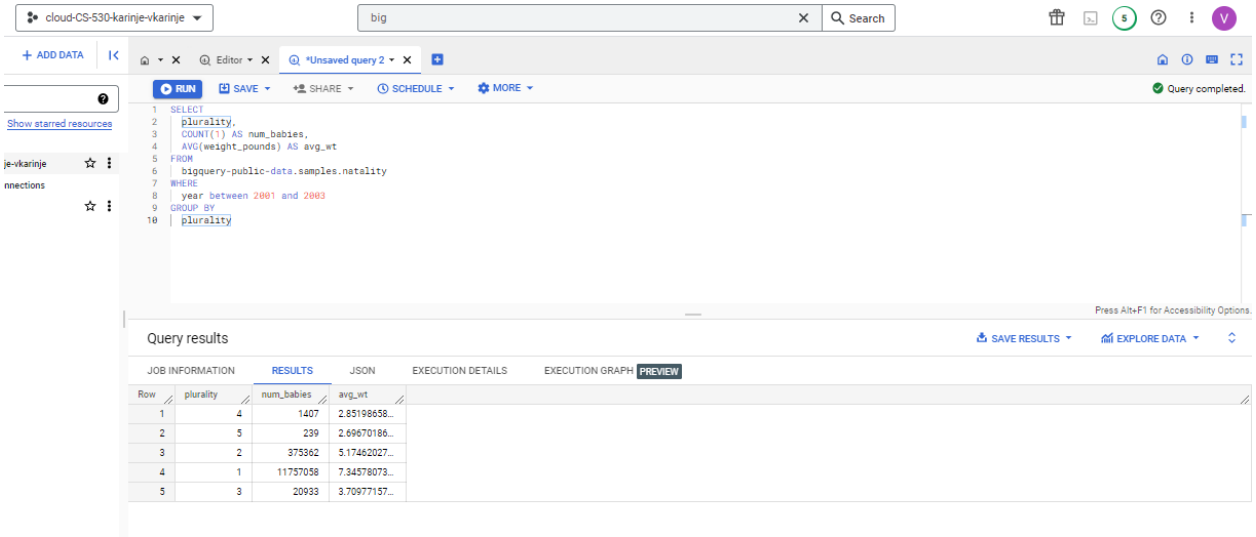


BigQuery query



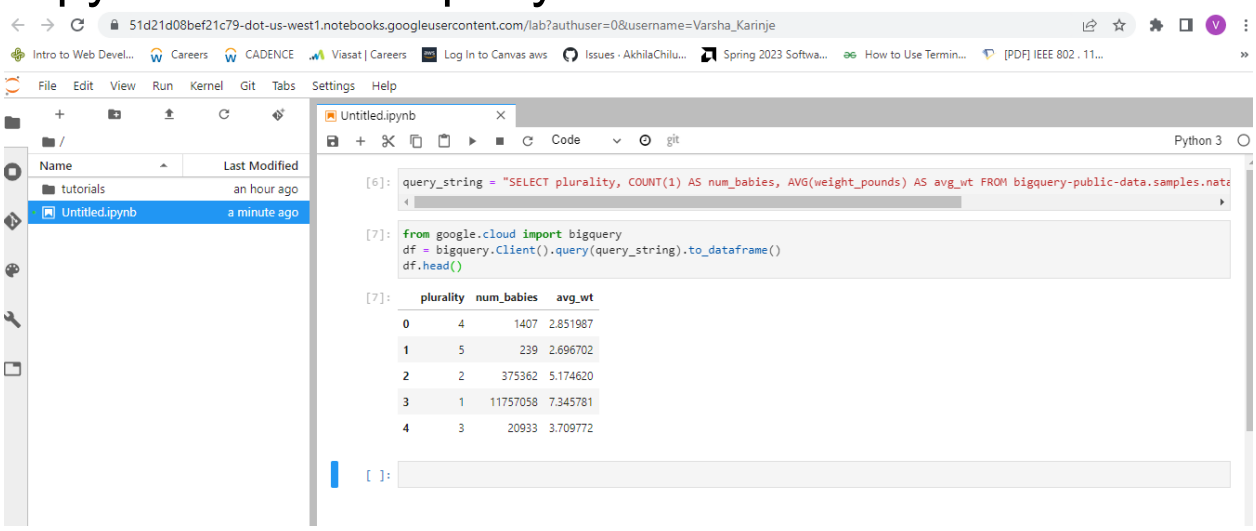
Answer the following question for your lab notebook:

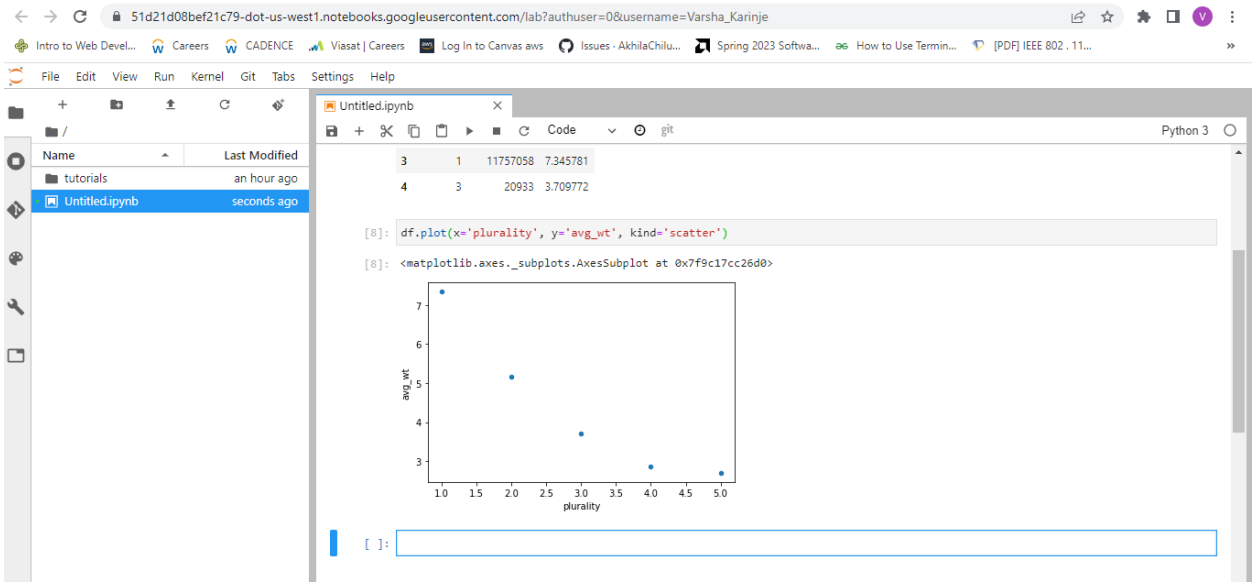
- How many twins were born during this time?



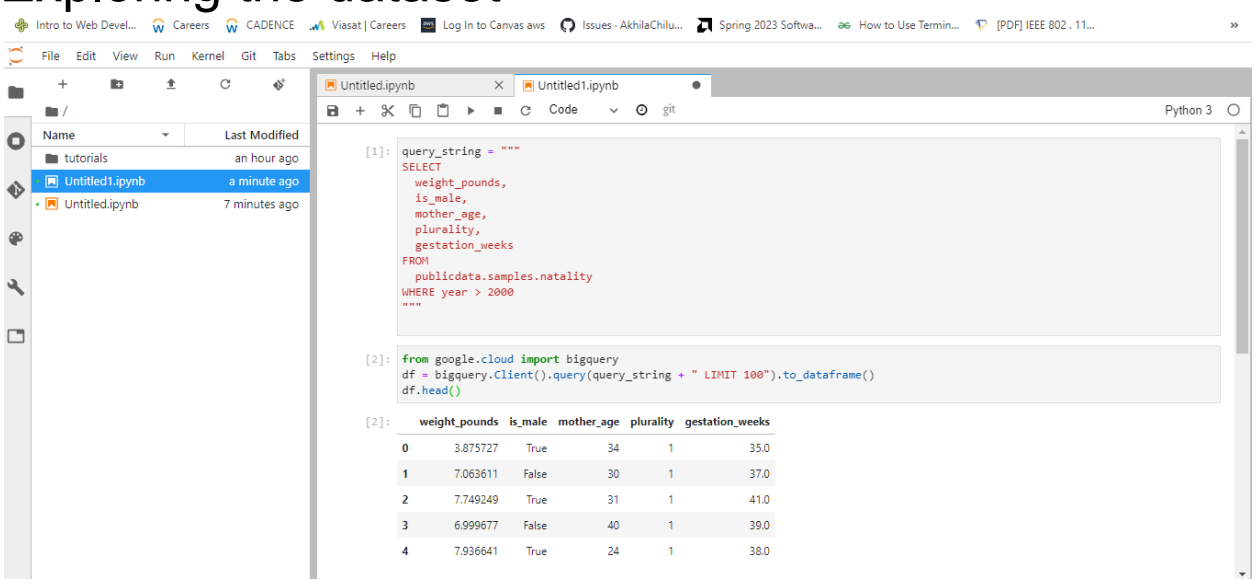
There were 375362 twins born during this time.

Jupyter notebook query

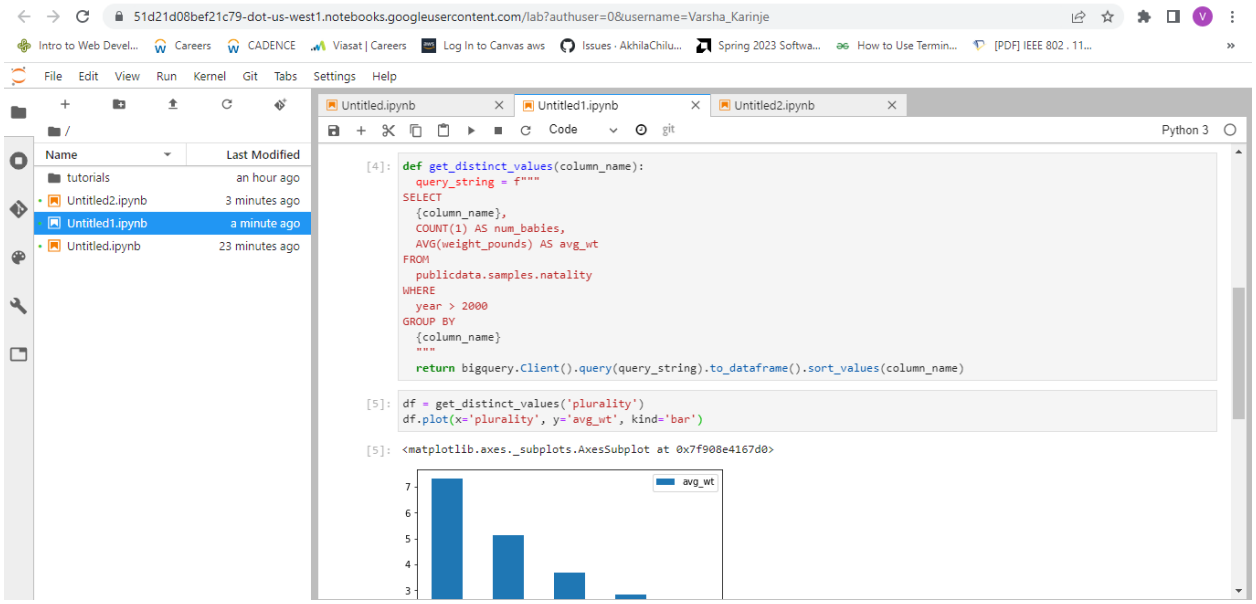


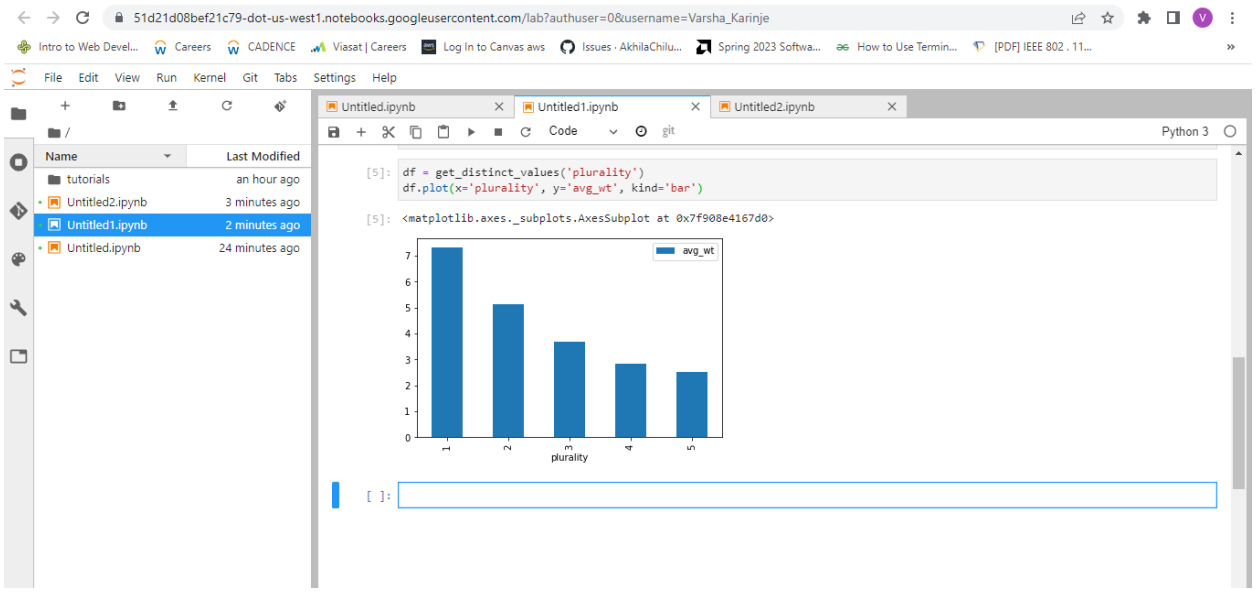


Exploring the dataset

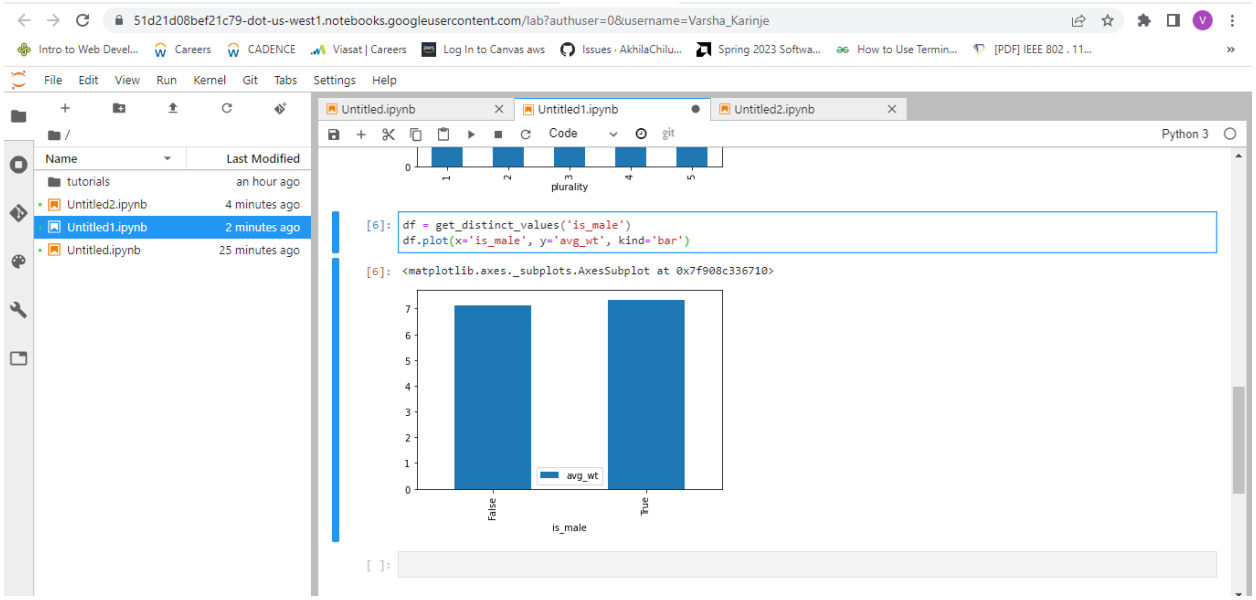


Run queries

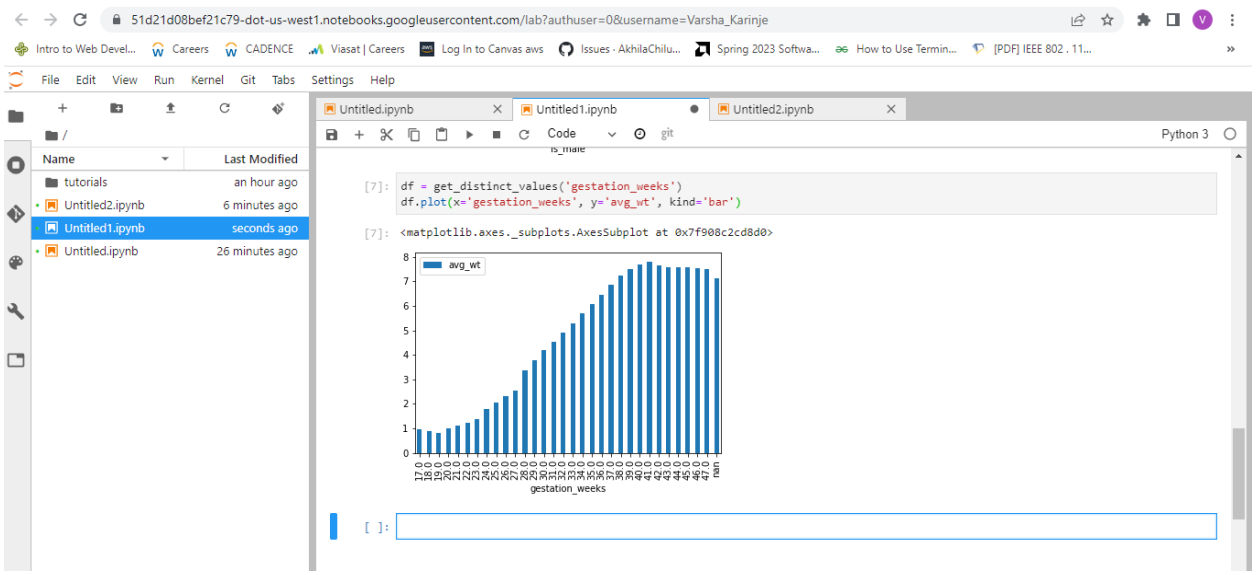




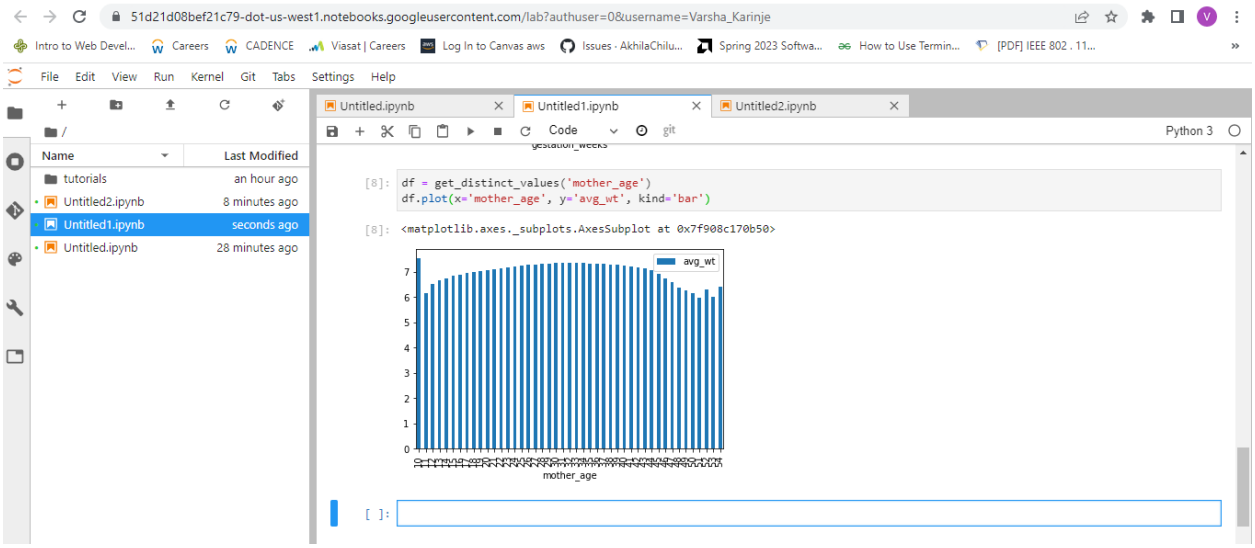
Then, run the query using gender:



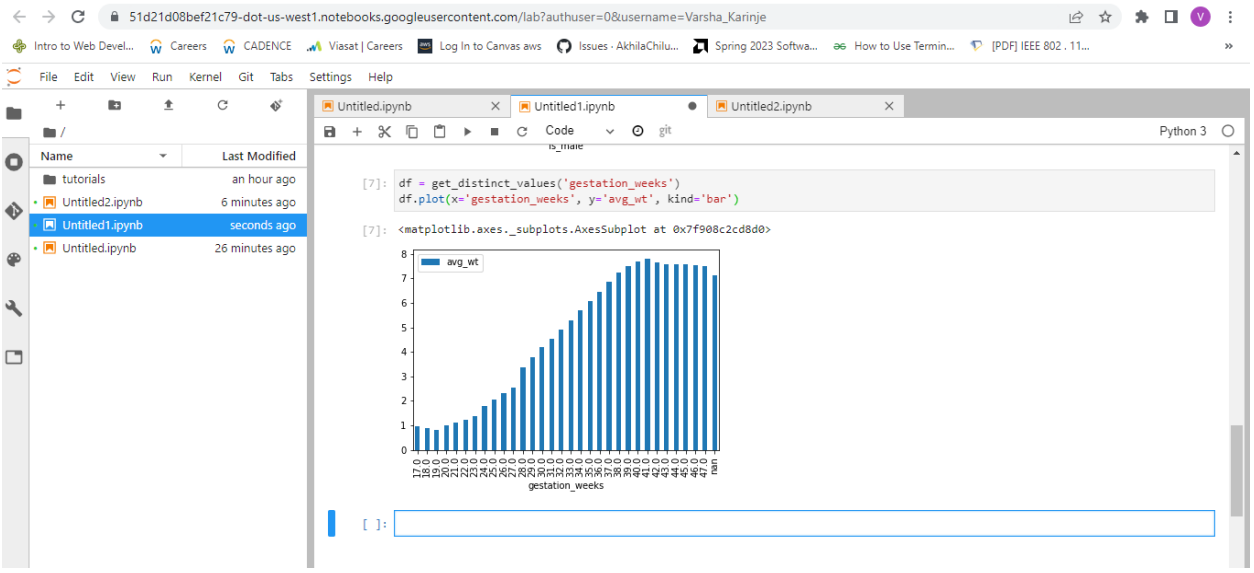
Then, run the query using gestation time:

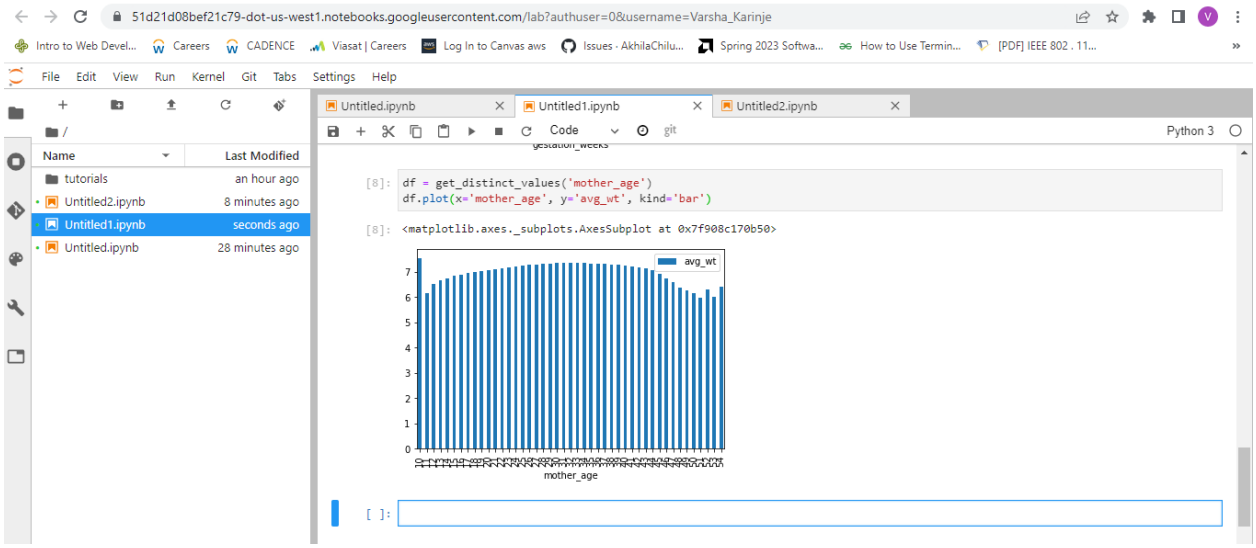


Finally, run the query using the mother's age:



- 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

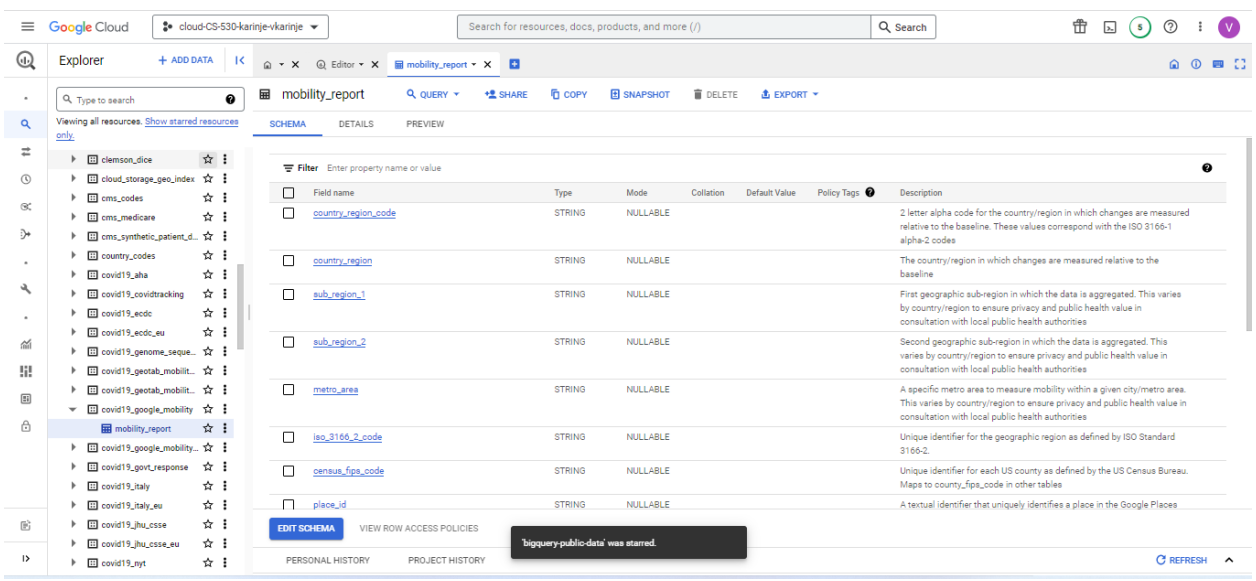


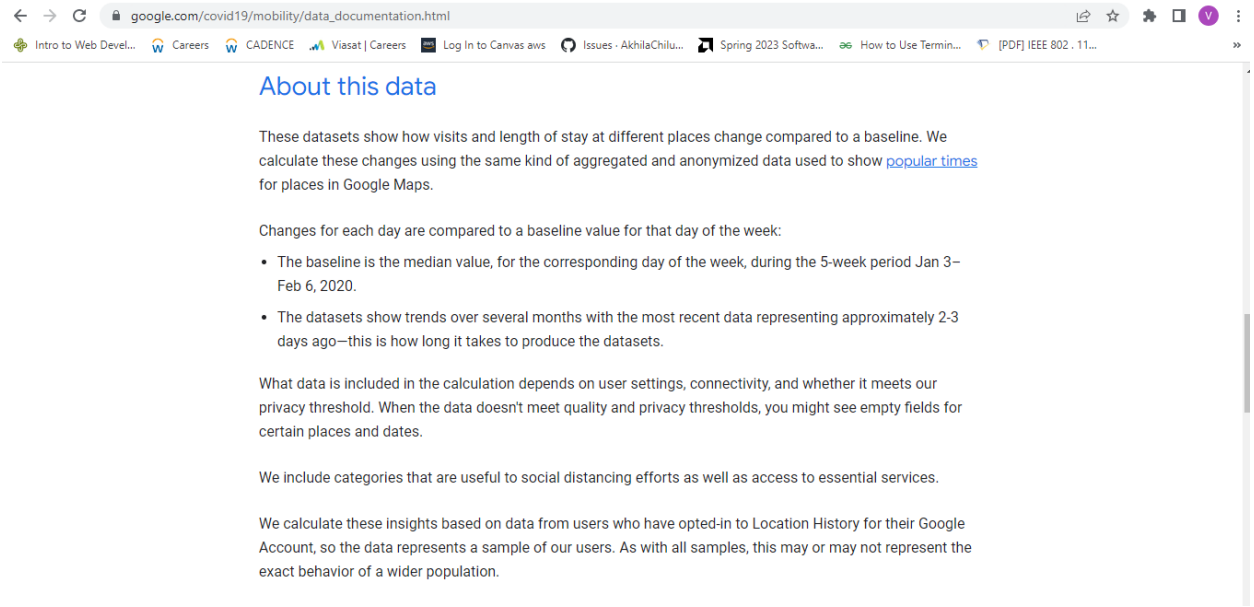


BigQuery, Notebooks Lab #3 (COVID-19 Mobility)

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.com/covid19/mobility/data_documentation.html

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.

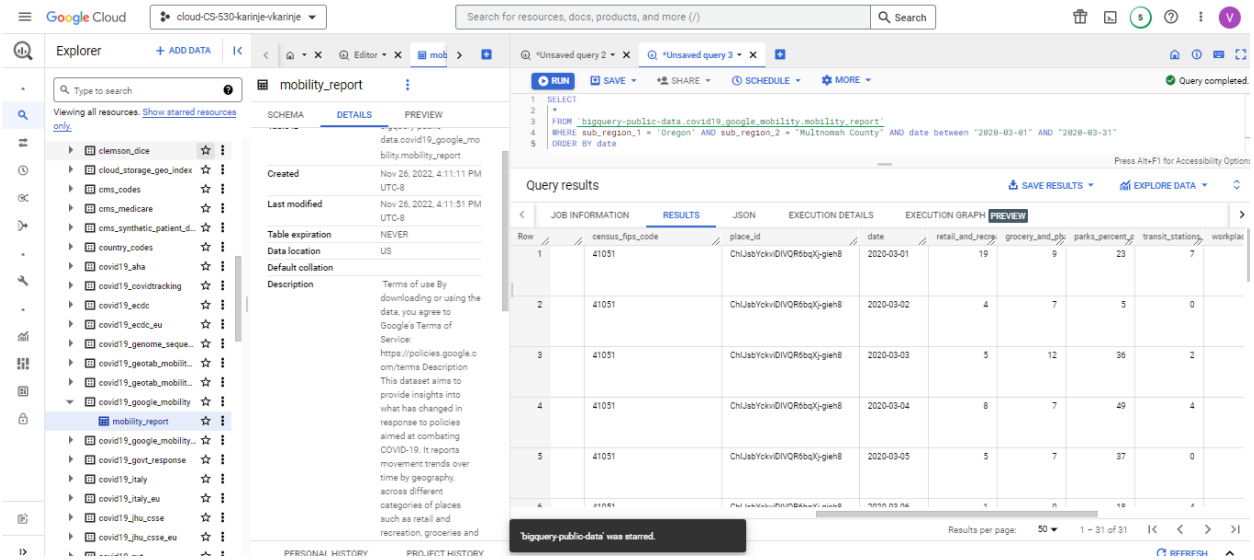
What data is included in the calculation depends on user settings, connectivity, and whether it meets our privacy threshold. When the data doesn't meet quality and privacy thresholds, you might see empty fields for certain places and dates.

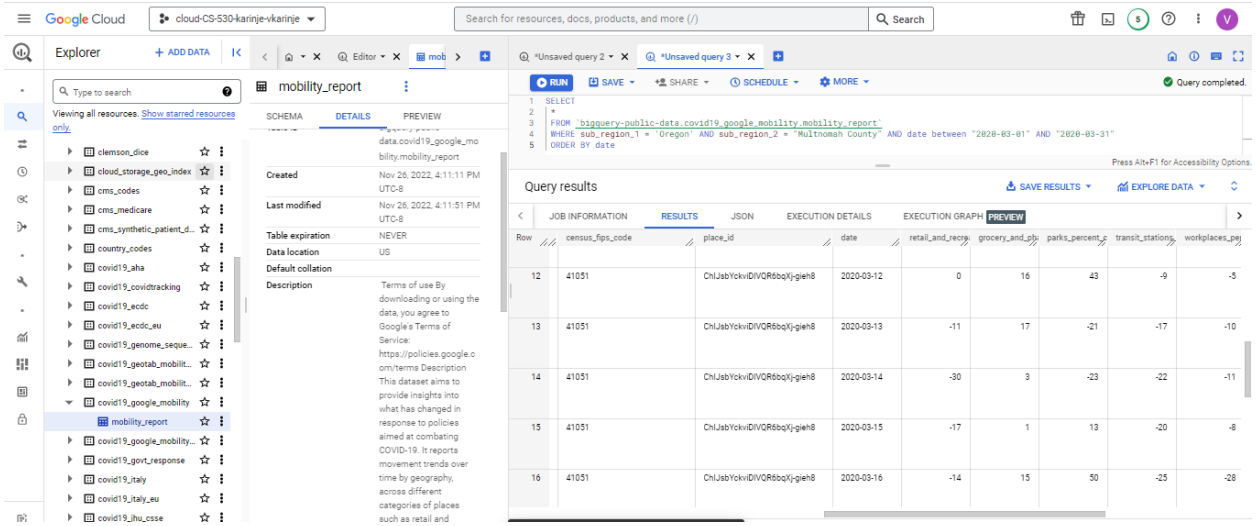
We include categories that are useful to social distancing efforts as well as access to essential services.

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?



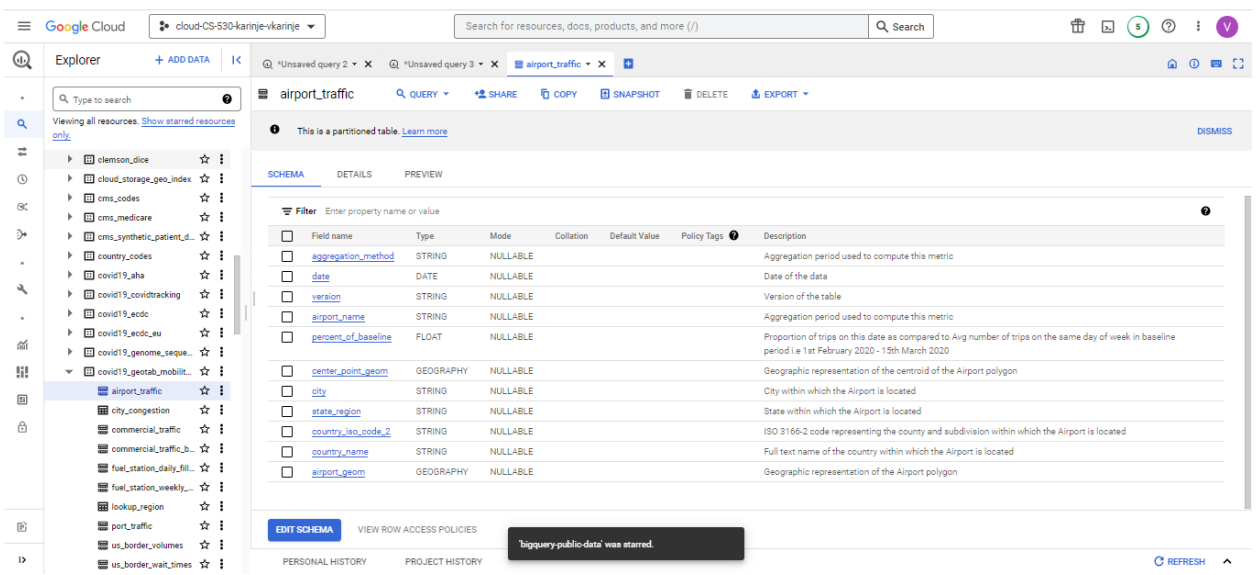


Ans: 2020-03-13 was the day that saw the largest spike in trips to grocery and pharmacy stores

- 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)

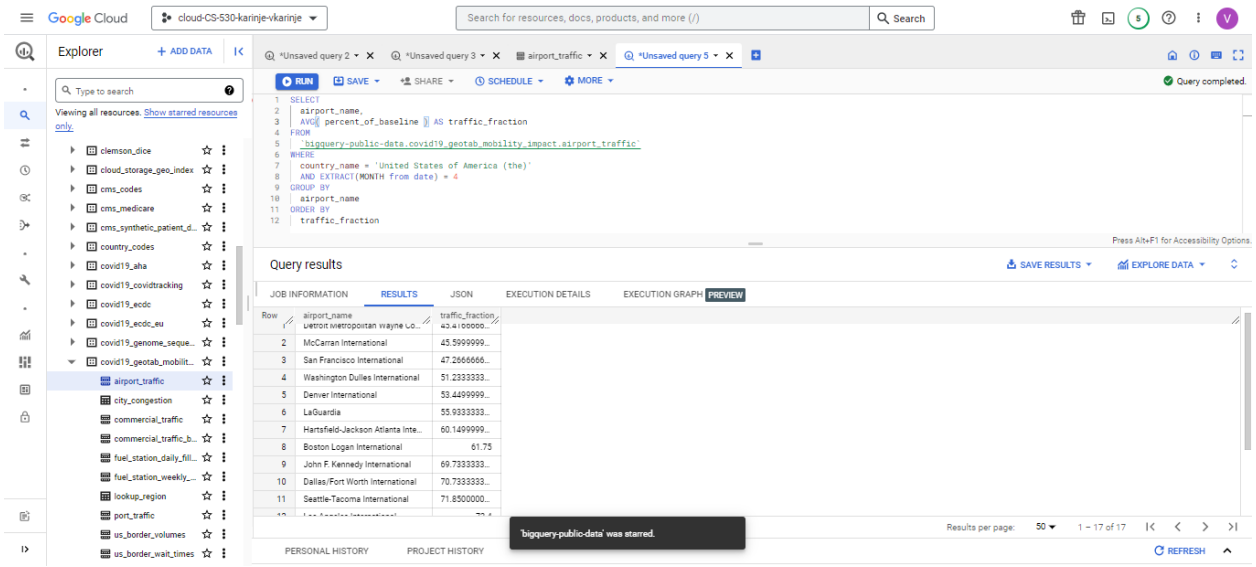
Find the column in this table that gives us information on the traffic impact.



percent_of_baseline is the table that gives us information on the traffic impact

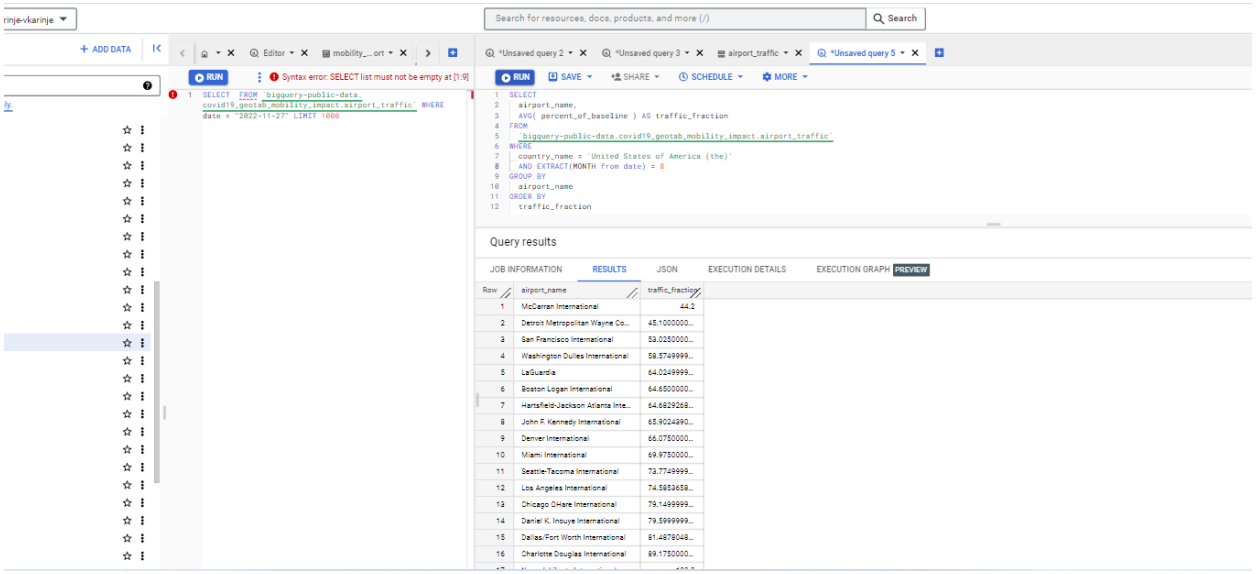
Then, adapt the query below to find the following

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



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

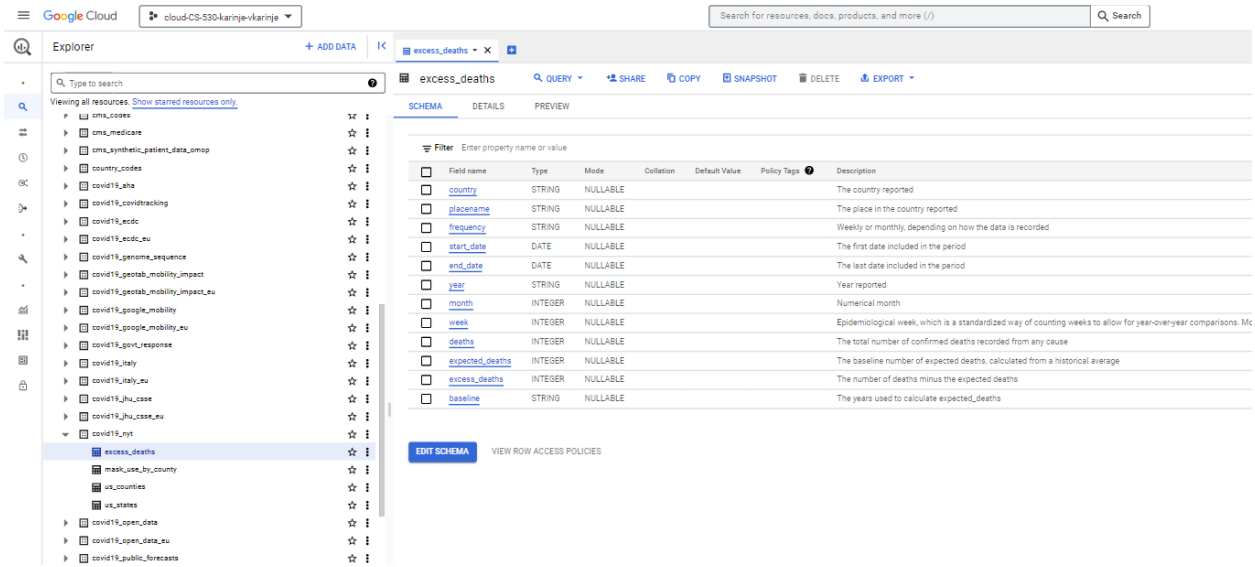
- Run the query again using the month of August 2020. Which three airports were impacted the most?



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

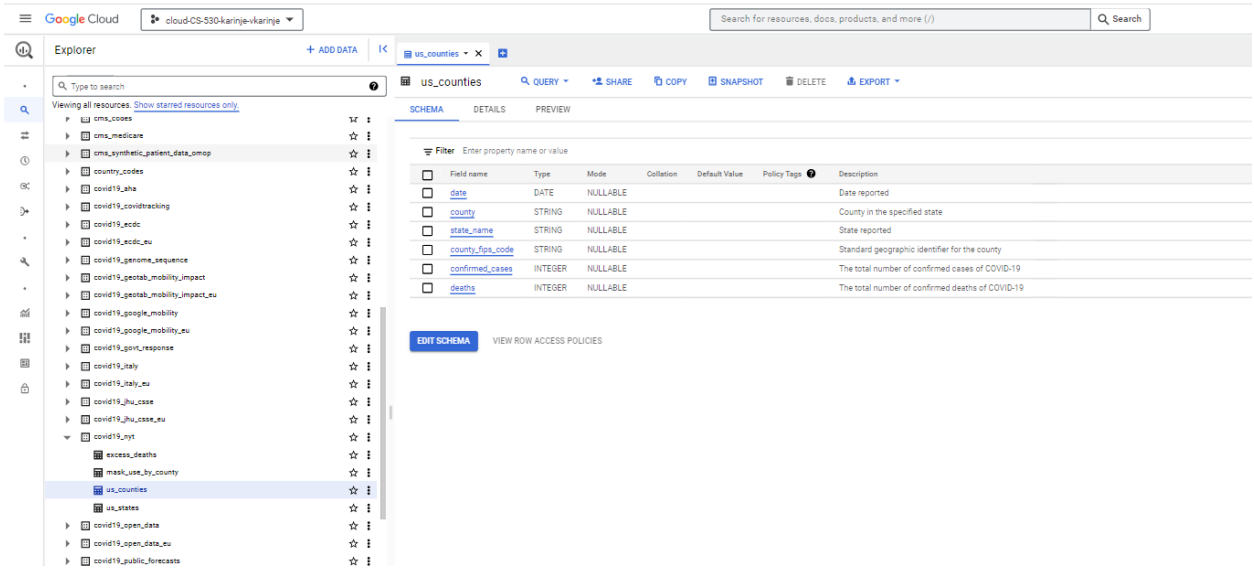
BigQuery, Notebooks Lab #4 (COVID-19 NYT)

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



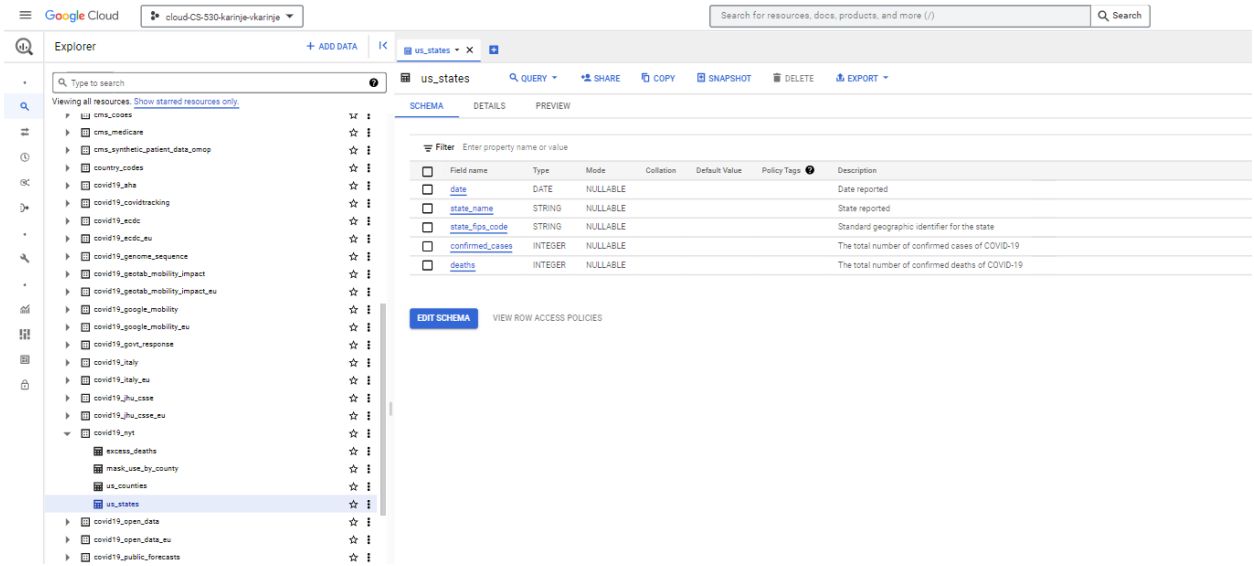
Ans: The table name is excess_deaths and the columns are placename, start_date and excess_deaths respectively.

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



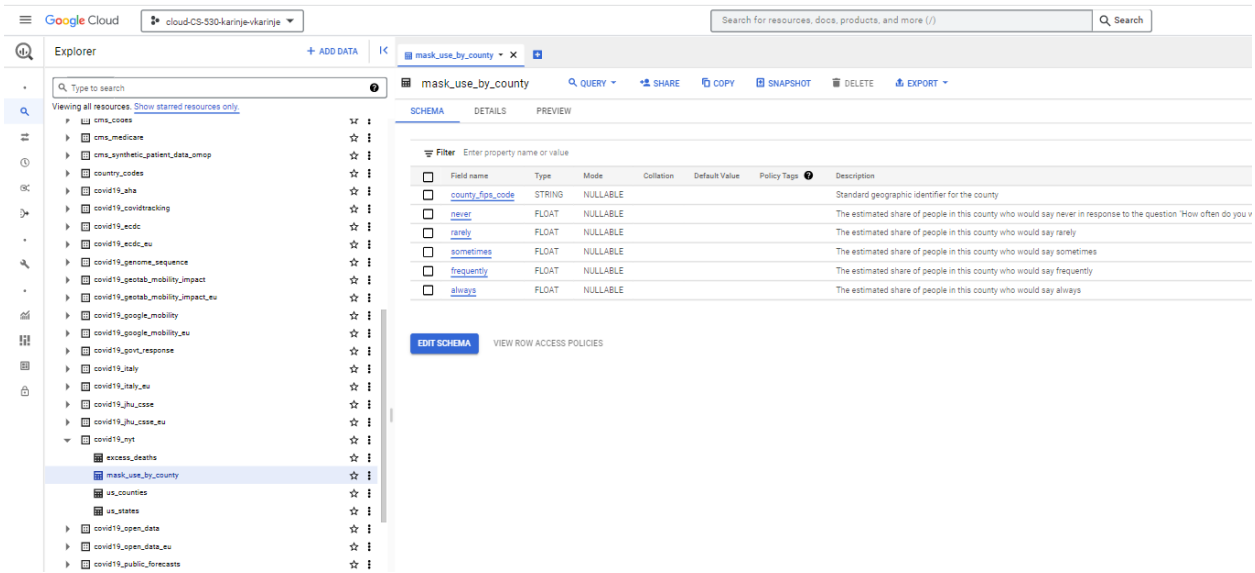
Ans: The table is us_counties and the columns are date, county and deaths

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



Ans:The table name is us_states and the columns are date, state_name and confirmed_cases

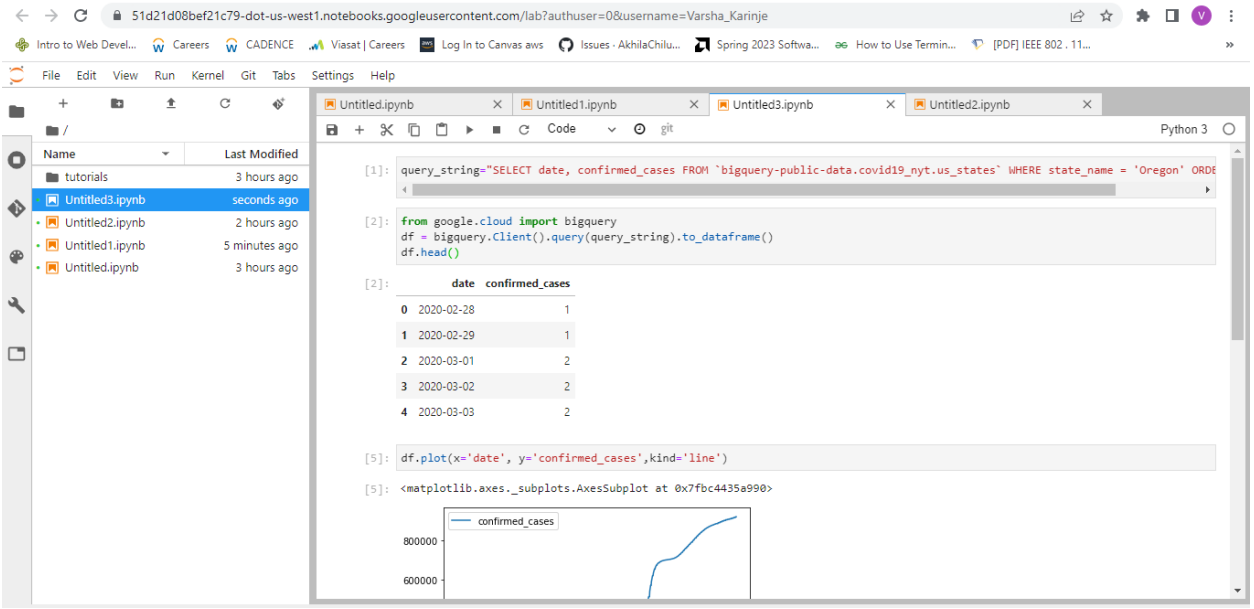
- What table and columns identify a county code and the percentage of its residents that report they always wear masks?

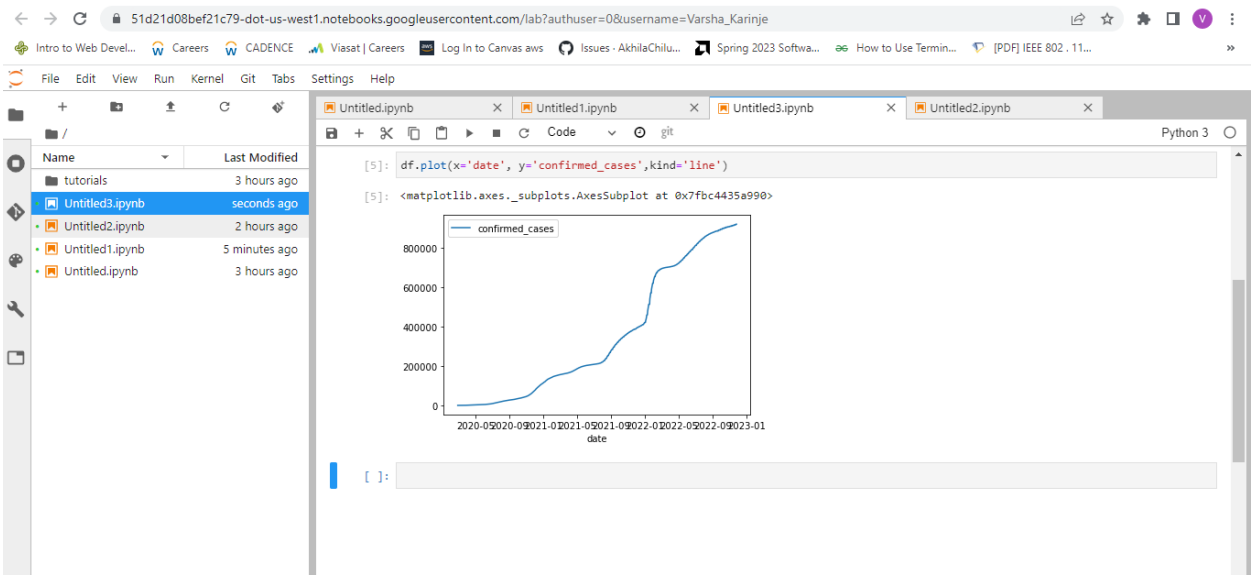


Ans: The table is mask_use_by_county and the columns are county_fips_code and always

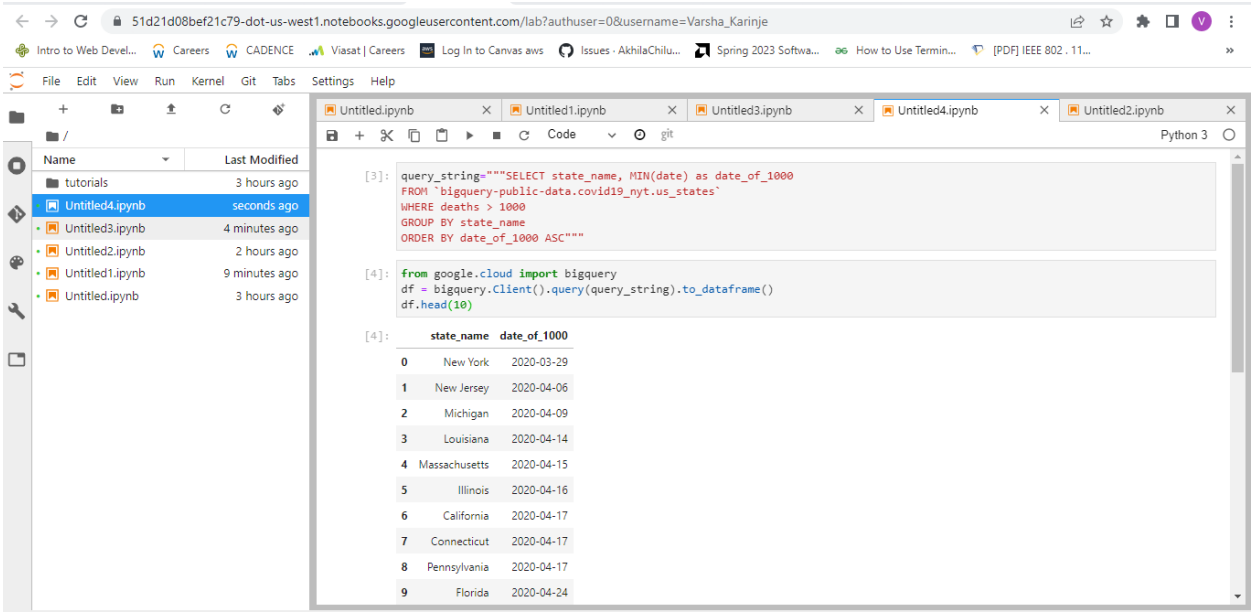
Run example queries

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

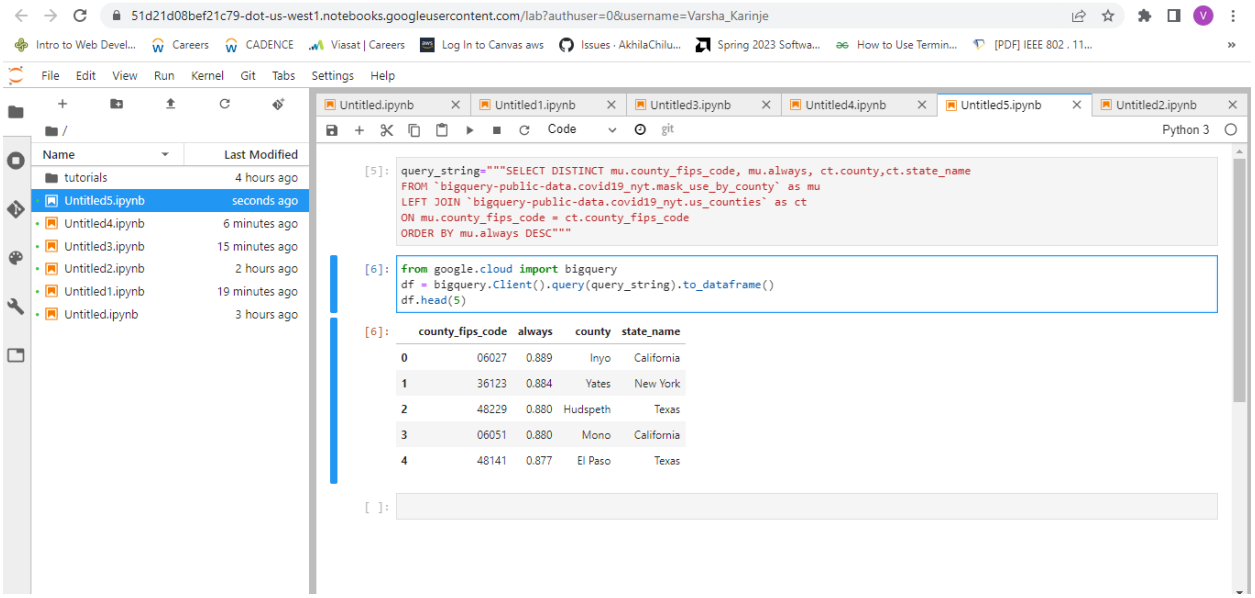




- From within your Jupyter notebook, run the query and write code that shows the first 10 states that reached 1000 deaths from COVID-19. Take a screenshot for your lab notebook.

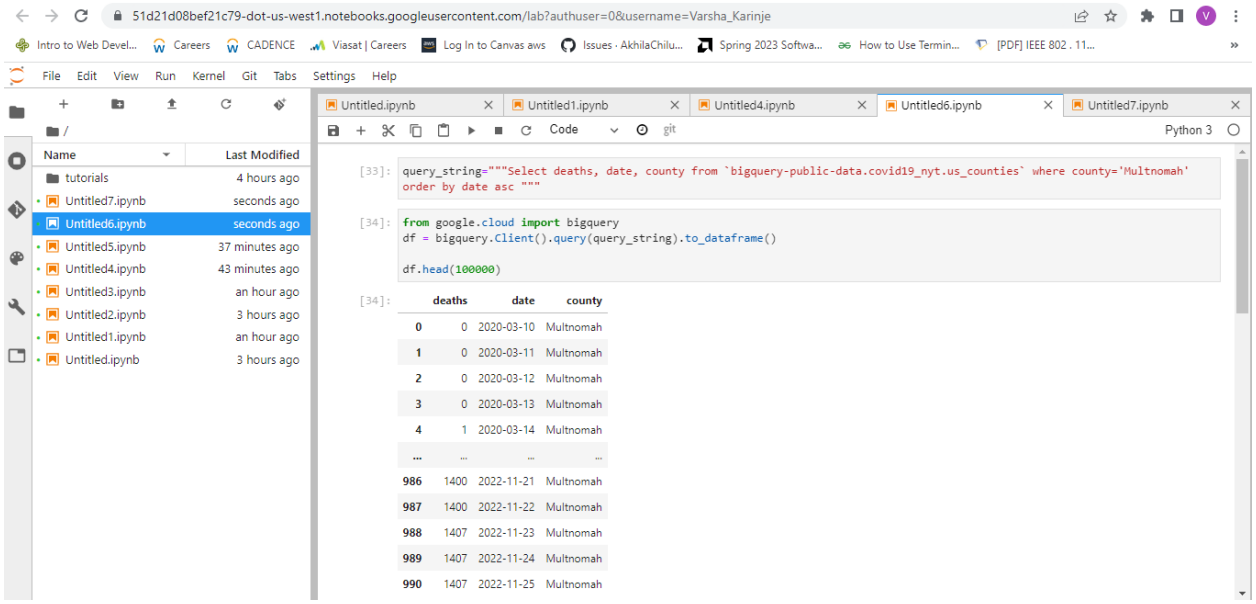


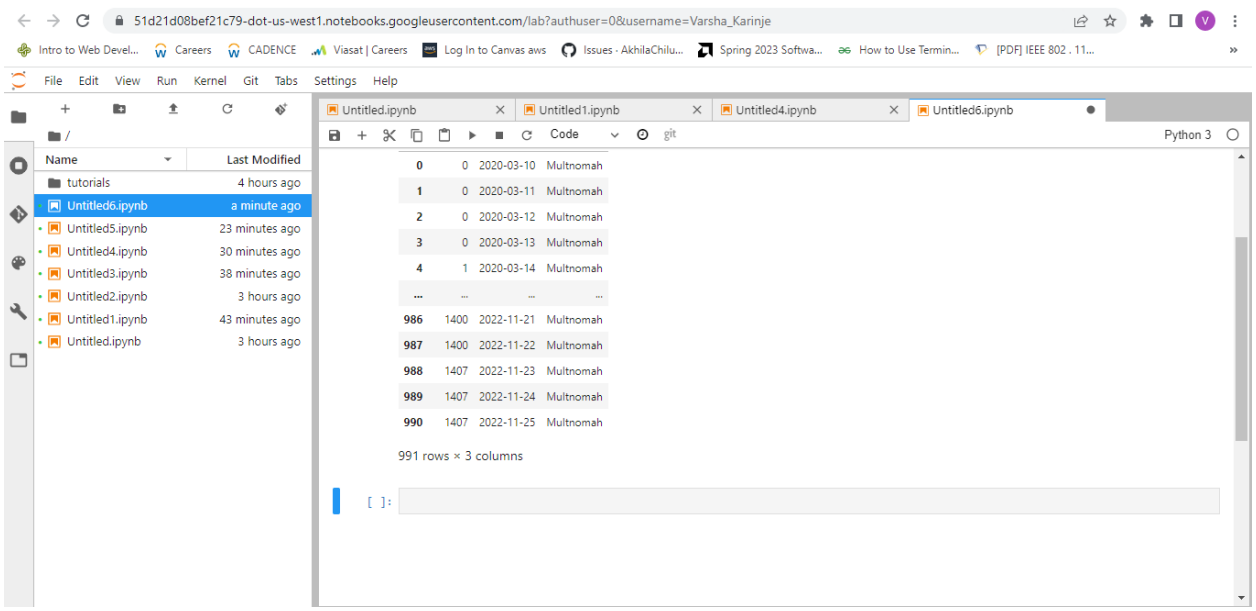
- Take a screenshot for your lab notebook of the Top 5 counties and the states they are located in.



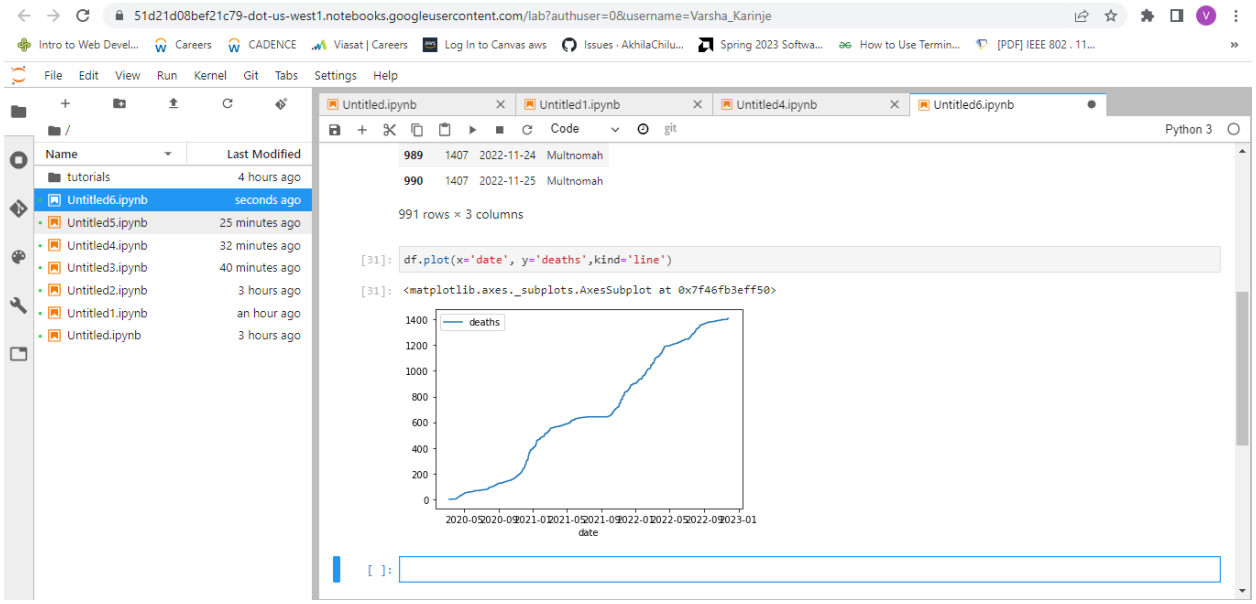
Write queries

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.

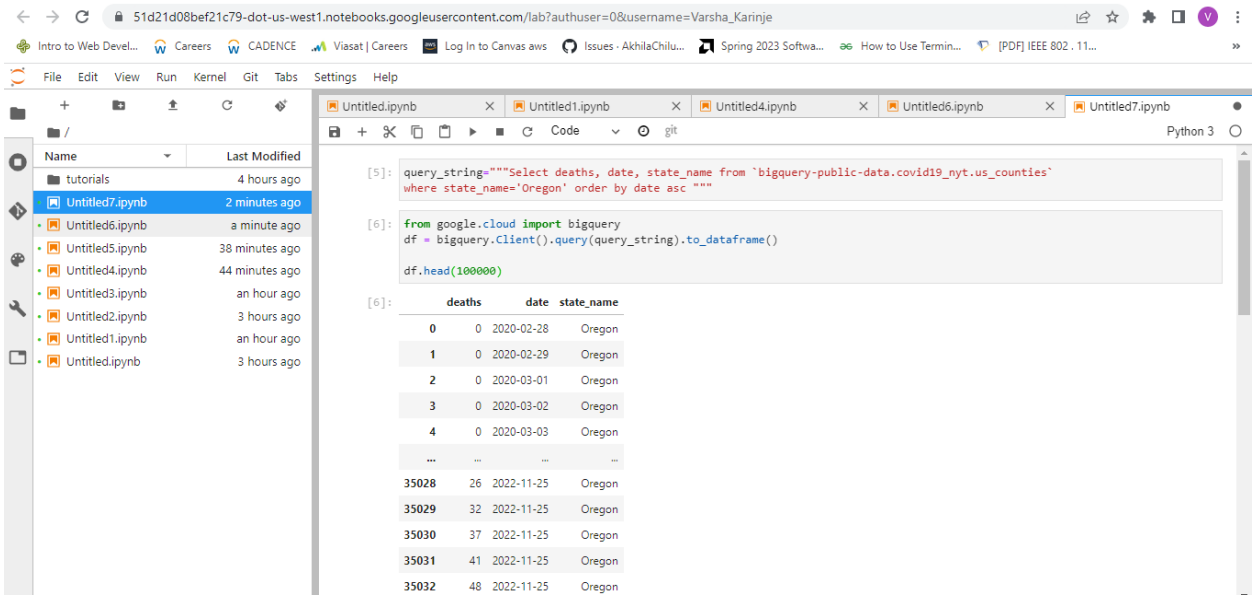


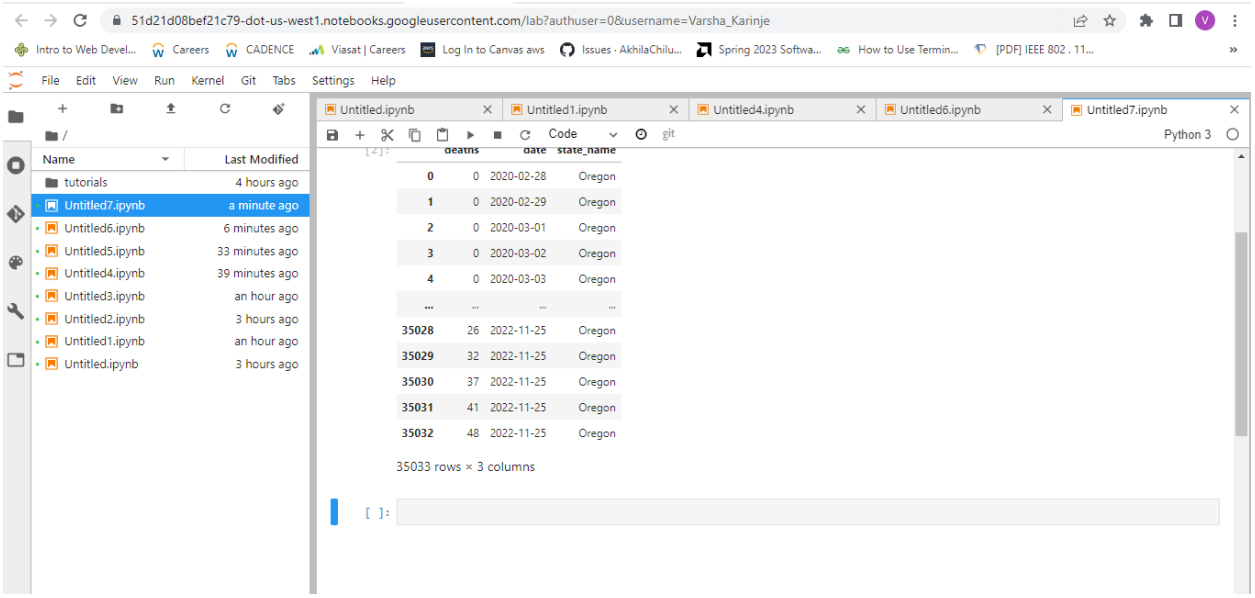


- Plot the results and take a screenshot for your lab notebook.

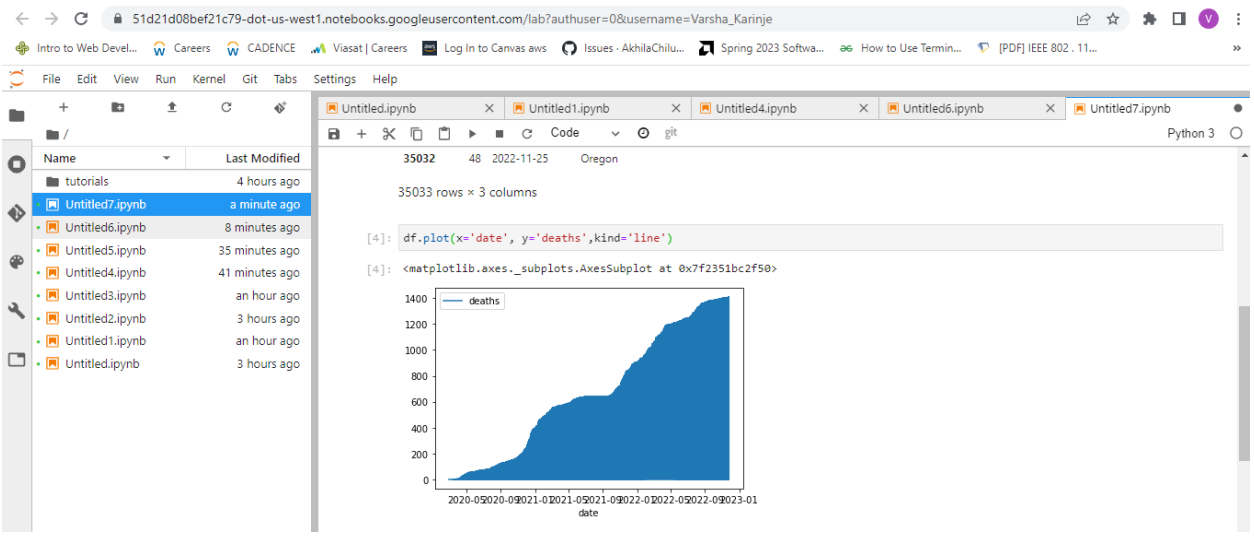


Construct a query string that obtains the number of deaths from COVID-19 that have occurred in Oregon for each day in the dataset, ensuring the data is returned in ascending order of date. Run the query and obtain the results.

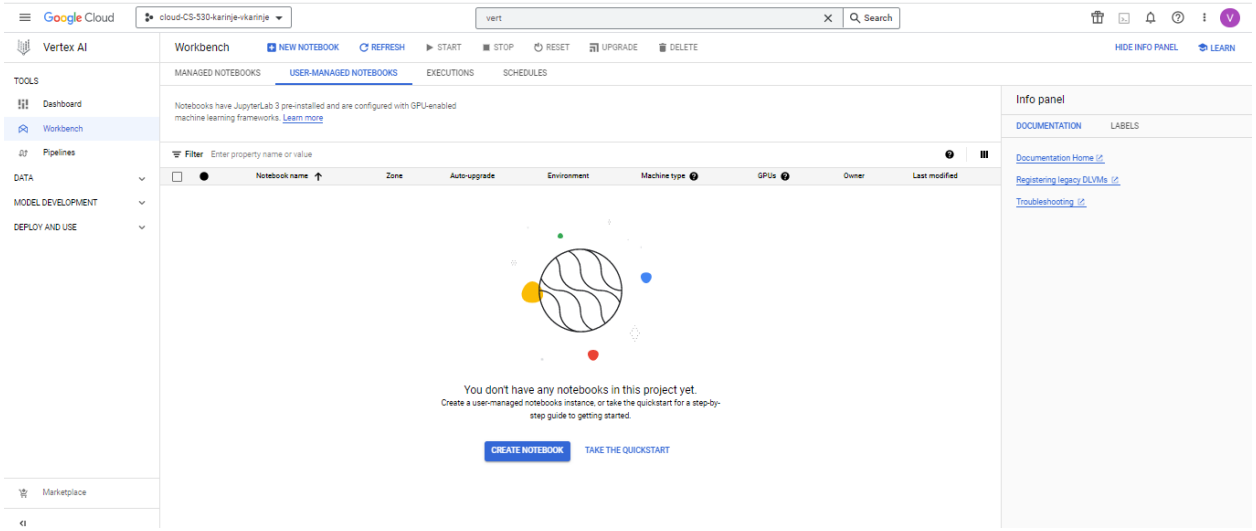


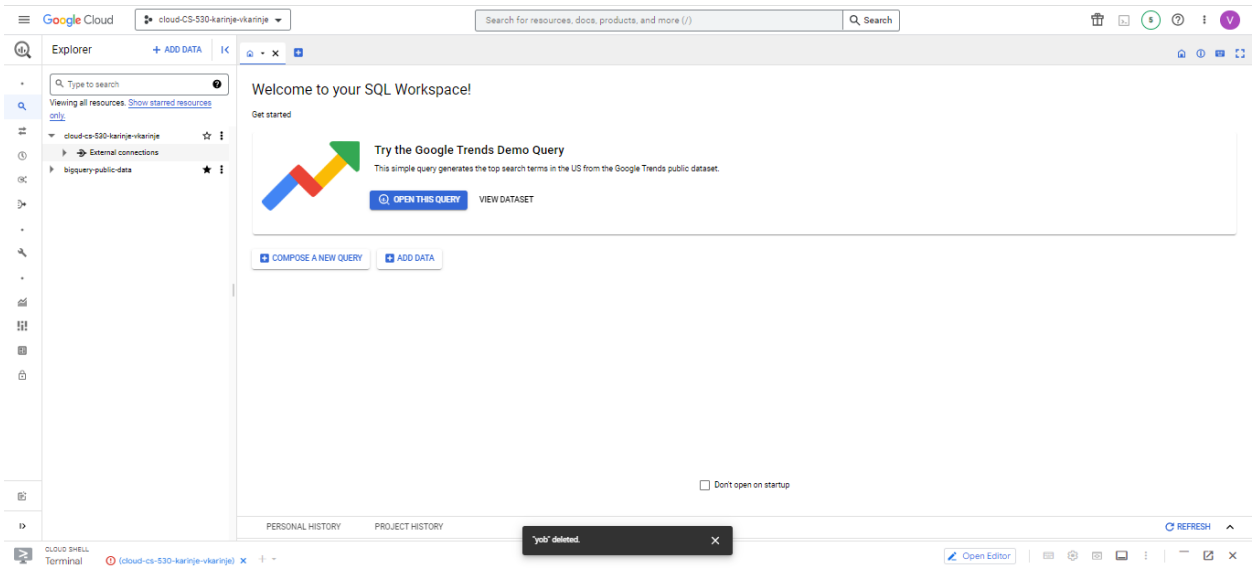


- Plot the results and take a screenshot for your lab notebook.



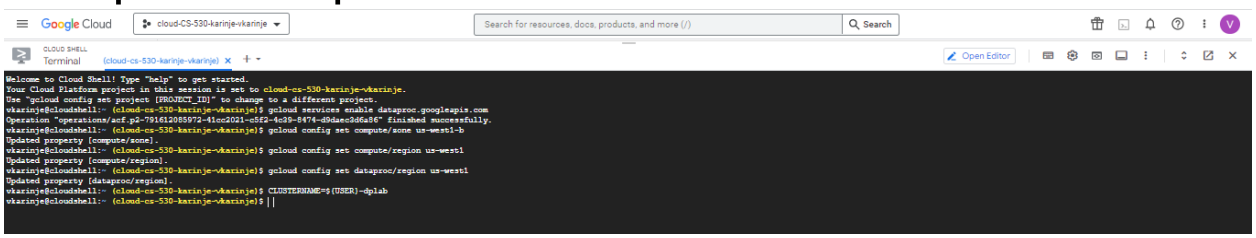
Clean up



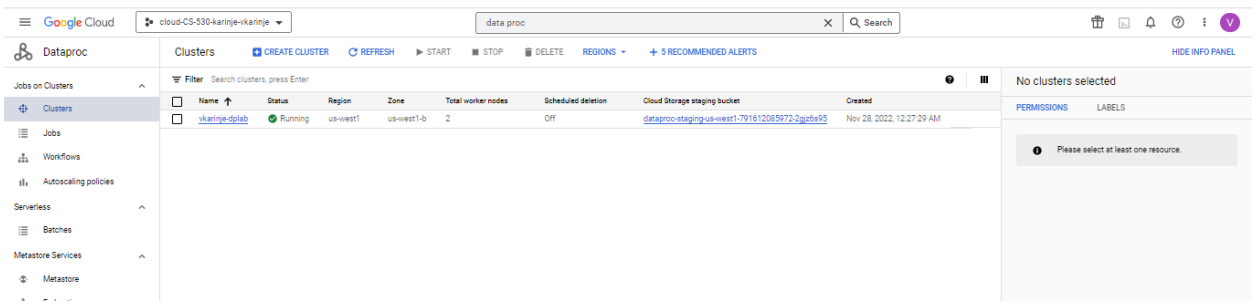


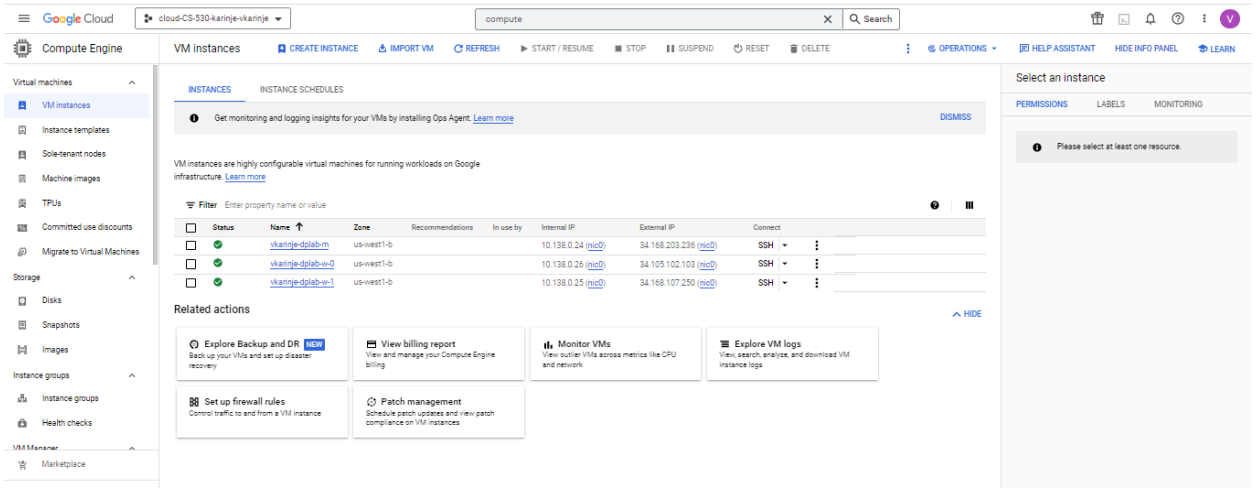
09.3g: Dataproc, Dataflow

Dataproc setup



Create Compute Engine cluster





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 π calculated.

Ans:

```
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ date
Mon 28 Nov 2022 09:23:09 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: 9482291a6abd496db3fdae2fdfb2e44d
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 [9482291a6abd496db3fdae2fdfb2e44d] 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 09:23:23 INFO org.sparkproject.jetty.util.log: Logging initialized @6401ms to org.sparkproject.jetty.util.log.Slf4jLog
22/11/28 09:23:23 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git: b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8.0_352-b08
22/11/28 09:23:23 INFO org.sparkproject.jetty.server.Server: Started @6651ms
22/11/28 09:23:23 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@5fad41be{HTTP/1.1, (http/1.1)}{0.0.0.0:40379}
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 09:23:28 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/10.138.0.35:8030
22/11/28 09:23:32 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: 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 [9482291a6abd496db3fdae2fdfb2e44d] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-west1-791612085972-2gjs6s95/google-cloud-dataproc-metainfo/b7aff99b-f4a3-4af1-bb1b-519d1ec5bb1c/jobs/9482291a6abd496db3fdae2fdfb2e44d/
driverOutputResourceUri: gs://dataproc-staging-us-west1-791612085972-2gjs6s95/google-cloud-dataproc-metainfo/b7aff99b-f4a3-4af1-bb1b-519d1ec5bb1c/jobs/9482291a6abd496db3fdae2fdfb2e44d/driveroutput
jobUuid: 4cb765ae-114d-3ac8-a949-fcdk91fac114
placement:
  clusterName: vkarinje-dplab
  clusterUuid: b7aff99b-f4a3-4af1-bb1b-519d1ec5bb1c
reference:
```

Π is roughly 3.1415080314150803

Scale cluster

```
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
  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 SHELL
Terminal (cloud-cs-530-karinje-vkarinje)                                                    Open Editor
      mapred:mapreduce.job.reduces: '1'
      mapred:mapreduce.jobhistory.recovery.store.class: org.apache.hadoop.mapreduce.v2.hs.HistoryServerLeveldbStateStoreService
      mapred:mapreduce.map.cpu.vcores: '1'
      mapred:mapreduce.map.java.opts: -Xmx1311m
      mapred:mapreduce.map.memory.mb: '1639'
      mapred:mapreduce.reduce.cpu.vcores: '2'
      mapred:mapreduce.reduce.java.opts: -Xmx2622m
      mapred:mapreduce.reduce.memory.mb: '3278'
      mapred:mapreduce.task.io.sort.mb: '256'
      mapred:yarn.app.mapreduce.am.command-opts: -Xmx1311m
      mapred:yarn.app.mapreduce.am.resource.cpu-vcores: '1'
      mapred:yarn.app.mapreduce.am.resource.mb: '1639'
      spark-env:SPARK_DAEMON_MEMORY: 1024m
      spark:spark.driver.maxResultSize: 512m
      spark:spark.driver.memory: 1024m
      spark:spark.executor.cores: '1'
      spark:spark.executor.instances: '2'
      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: 640m
      yarn-env:YARN_NODEMANAGER_HEAPSIZE: '409'
      yarn-env:YARN_RESOURCEMANAGER_HEAPSIZE: '1024'
      yarn-env:YARN_TIMELINESERVER_HEAPSIZE: '1024'
      yarn:yarn.nodemanager.address: 0.0.0.0:8026
      yarn:yarn.nodemanager.resource.cpu-vcores: '2'
      yarn:yarn.nodemanager.resource.memory-mb: '3278'
      yarn:yarn.resourcemanager.nodemanager-graceful-decommission-timeout-secs: '86400'
      yarn:yarn.scheduler.maximum-allocation-mb: '3278'
      yarn:yarn.scheduler.minimum-allocation-mb: '1'
  tempBucket: dataproc-temp-us-west1-791612085972-w3exx5h3
  workerConfig:
    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-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)                                                    Open Editor
    minCpuPlatform: AUTOMATIC
    numInstances: 2
    preemptibility: NON_PREEMPTIBLE
  labels:
    goog-dataproc-cluster-name: vkarinje-dplab
    goog-dataproc-cluster-uuid: b7aff99b-f4a3-4af1-bb1b-519d1ec5bb1c
    goog-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-blocks-under-replication: '0'
      dfs-capacity-present: '3398898528'
      dfs-capacity-remaining: '3398891520'
      dfs-capacity-total: '6292889648'
      dfs-capacity-used: '7372'
      dfs-nodes-decommissioned: '0'
      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-mb-allocated: '0'
      yarn-memory-mb-available: '6556'
      yarn-memory-mb-pending: '0'
      yarn-memory-mb-reserved: '0'
      yarn-memory-mb-total: '6556'
      yarn-nodes-active: '2'
      yarn-nodes-decommissioned: '0'
      yarn-nodes-lost: '0'
      yarn-nodes-rebooted: '0'
      yarn-nodes-unhealthy: '0'
      yarn-vcores-allocated: '0'
      yarn-vcores-available: '4'
      yarn-vcores-pending: '0'
      yarn-vcores-reserved: '0'
      yarn-vcores-total: '4'
projectId: cloud-cs-530-karinje-vkarinje
status:
```

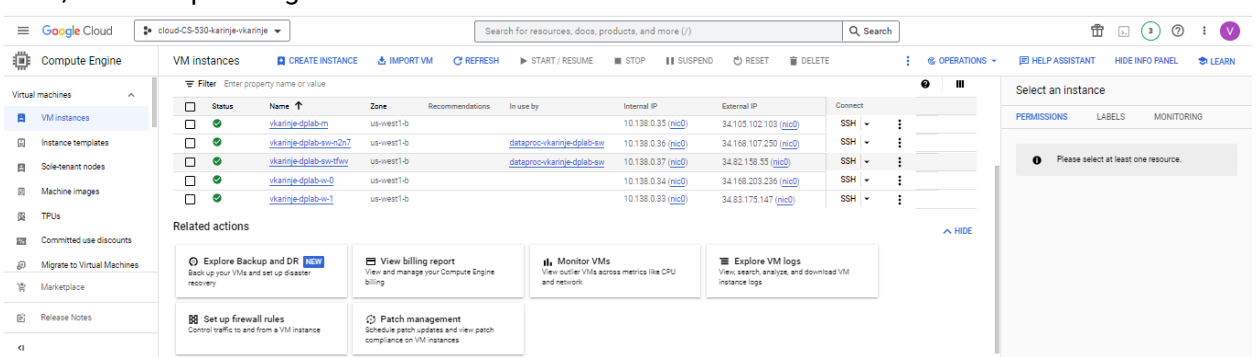
Allocate two additional pre-emptible machines to the cluster

```
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 ${CLUSTERNAME} --num-secondary-workers=2
Waiting on operation [projects/cloud-cs-530-karinje-vkarinje/regions/us-west1/operations/1a258965-66a2-3cb5-94b9-aa457decb041].
Waiting for cluster update operation...done.
Updated [https://dataproc.googleapis.com/v1/projects/cloud-cs-530-karinje-vkarinje/regions/us-west1/clusters/vkarinje-dplab].
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ ||
```

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-west1-791612085972-2gjz6s95
  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-tfwv
    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.



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 π calculated.

  Ans: Pie is roughly 3.141625191416252

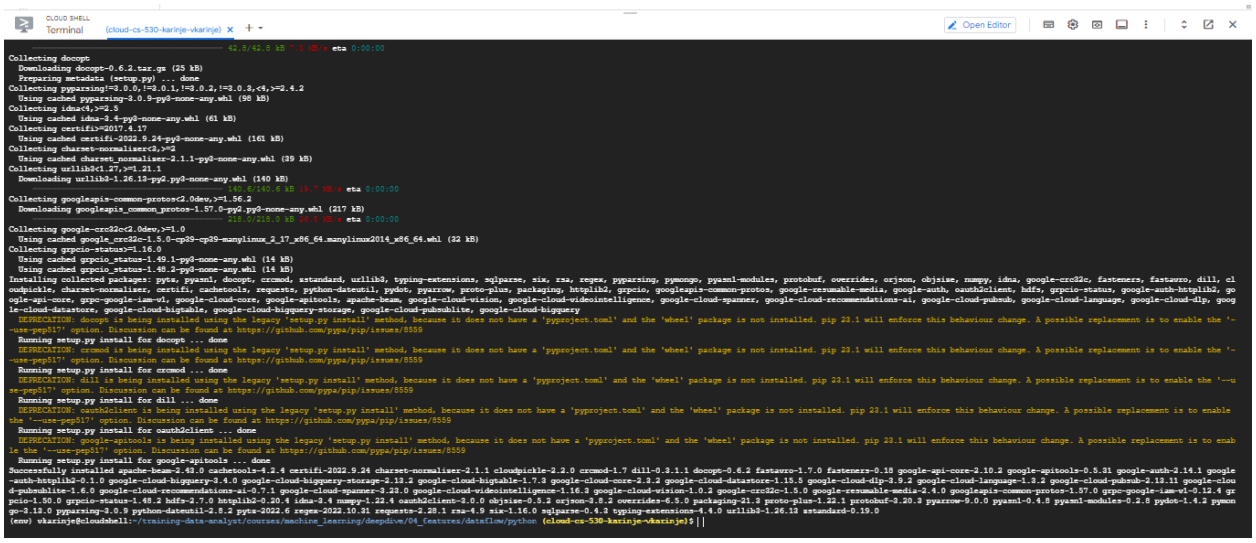
```
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ cat output2.txt
Job [a58b8dd8a69a48dd91b9174642fcf342] submitted.
Waiting for job output...
22/11/29 05:34:17 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/11/29 05:34:17 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/11/29 05:34:17 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
22/11/29 05:34:17 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
22/11/29 05:34:17 INFO org.sparkproject.jetty.util.log: Logging initialized @4121ms to org.sparkproject.jetty.util.log.Slf4jLog
22/11/29 05:34:17 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git: b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8.0_352-b08
22/11/29 05:34:17 INFO org.sparkproject.jetty.server.Server: Started @4224ms
22/11/29 05:34:17 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@73ffae32{HTTP/1.1, (http/1.1)}{0.0.0.0:44333}
22/11/29 05:34:18 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at vkarinje-dplab-m/10.138.0.53:8032
22/11/29 05:34:18 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at vkarinje-dplab-m/10.138.0.53:10200
22/11/29 05:34:19 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
22/11/29 05:34:19 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/11/29 05:34:20 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1669699902947_0002
22/11/29 05:34:21 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at vkarinje-dplab-m/10.138.0.53:8030
22/11/29 05:34:23 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
Pi is roughly 3.1416251191416252
22/11/29 05:34:38 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@73ffae32{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [a58b8dd8a69a48dd91b9174642fcf342] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-west1-791612085972-2gjs6s95/google-cloud-dataproc-metainfo/77cc77c3-e4c8-4286-a0a1-b75853f78592/jobs/a58b8dd8a69a48dd91b9174642fcf342/
driverOutputResourceUri: gs://dataproc-staging-us-west1-791612085972-2gjs6s95/google-cloud-dataproc-metainfo/77cc77c3-e4c8-4286-a0a1-b75853f78592/jobs/a58b8dd8a69a48dd91b9174642fcf342/driveroutput
jobUuid: 802035fa-6c3c-2771-b9b6-886631d6f355
placement:
  clusterName: vkarinje-dplab
  clusterUuid: 77cc77c3-e4c8-4286-a0a1-b75853f78592
reference:
  jobId: a58b8dd8a69a48dd91b9174642fcf342
  projectId: cloud-cs-530-karinje-vkarinje
sparkJob:
  args:
  - '1000'
  jarFileUris:
  - file:///usr/lib/spark/examples/jars/spark-examples.jar
  mainClass: org.apache.spark.examples.SparkPi
status:
  state: DONE
  stateStartTime: '2022-11-29T05:34:43.310510Z'
statusHistory:
- state: PENDING
  stateStartTime: '2022-11-29T05:34:10.351257Z'
```

Clean up

```
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud dataproc clusters delete $CLUSTERNAME
The cluster 'vkarinje-dplab' and all attached disks will be deleted.

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

Waiting on operation [projects/cloud-cs-530-karinje-vkarinje/regions/us-west1/operations/967ccd8f-2ed9-3255-a19d-69d62cfdaf76].
Waiting for cluster deletion operation...done.
Deleted [https://dataproc.googleapis.com/v1/projects/cloud-cs-530-karinje-vkarinje/regions/us-west1/clusters/vkarinje-dplab].
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ ||
```

Setup

```
vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ cd training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python/
vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje)$ python3 -m venv env
vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje)$ source env/bin/activate
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje)$ pip install -U pip
Requirement already satisfied: pip in ./env/lib/python3.9/site-packages (20.3.4)
Collecting pip
  Using cached pip-22.3.1-py3-none-any.whl (2.1 MB)
Installing collected packages: pip
  Attempting uninstall: pip
    Found existing installation: pip 20.3.4
    Uninstalling pip-20.3.4:
      Successfully uninstalled pip-20.3.4
Successfully installed pip-22.3.1
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje)$ pip install apache-beam[gcp] oauth2client==3.0.0
Collecting apache-beam[gcp]
  Downloading apache_beam-2.43.0-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (14.3 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 14.3/14.3 MB 44.0 MB/s eta 0:00:00
Collecting oauth2client==3.0.0
  Downloading oauth2client-3.0.0.tar.gz (77 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 77.2/77.2 kB 12.0 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting httplib2>=0.9.1
  Using cached httplib2-0.21.0-py3-none-any.whl (96 kB)
Collecting pyasn1-modules>=0.0.5
  Using cached pyasn1_modules-0.2.8-py2.py3-none-any.whl (155 kB)
Collecting pyasn1>=0.1.7
  Using cached pyasn1-0.4.8-py2.py3-none-any.whl (77 kB)
Collecting rsa>=3.1.4
  Using cached rsa-4.9-py3-none-any.whl (34 kB)
Collecting six>=1.6.1
  Using cached six-1.16.0-py2.py3-none-any.whl (11 kB)
Collecting proto-plus<2,>=1.7.1
  Using cached proto_plus-1.22.1-py3-none-any.whl (47 kB)
Collecting typing-extensions>=3.7.0
  Downloading typing_extensions-4.4.0-py3-none-any.whl (26 kB)
Collecting fasteners<1.0,>=0.3
  Downloading fasteners-0.18-py3-none-any.whl (18 kB)
Collecting pydot<2,>=1.2.0
  Downloading pydot-1.4.2-py2.py3-none-any.whl (21 kB)
Collecting protobuf<4,>3.12.2
  Using cached protobuf-3.20.3-cp39-cp39-manylinux_2_5_x86_64.manylinux1_x86_64.whl (1.0 MB)
Collecting requests<3.0.0,>=2.24.0
```

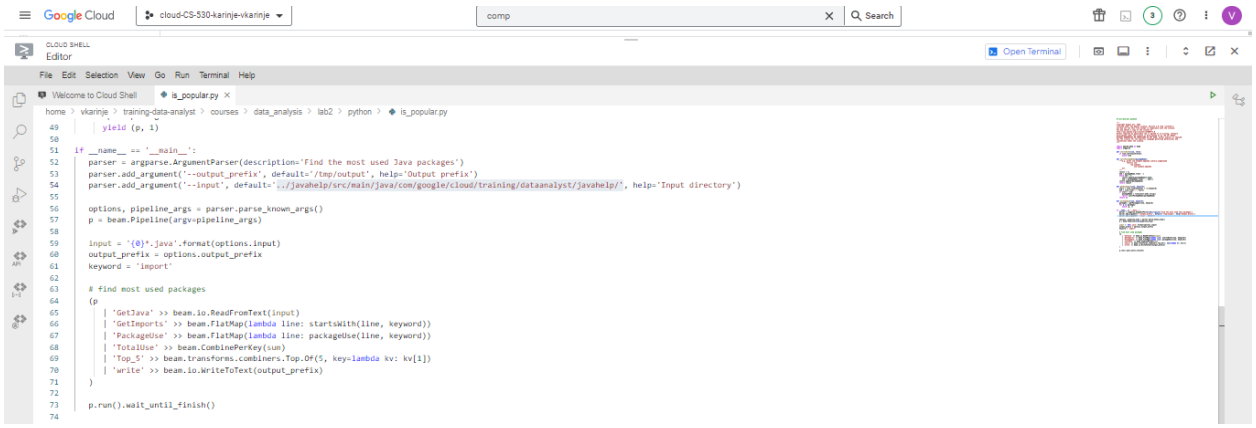


Beam code

Answer the following questions for your lab notebook.

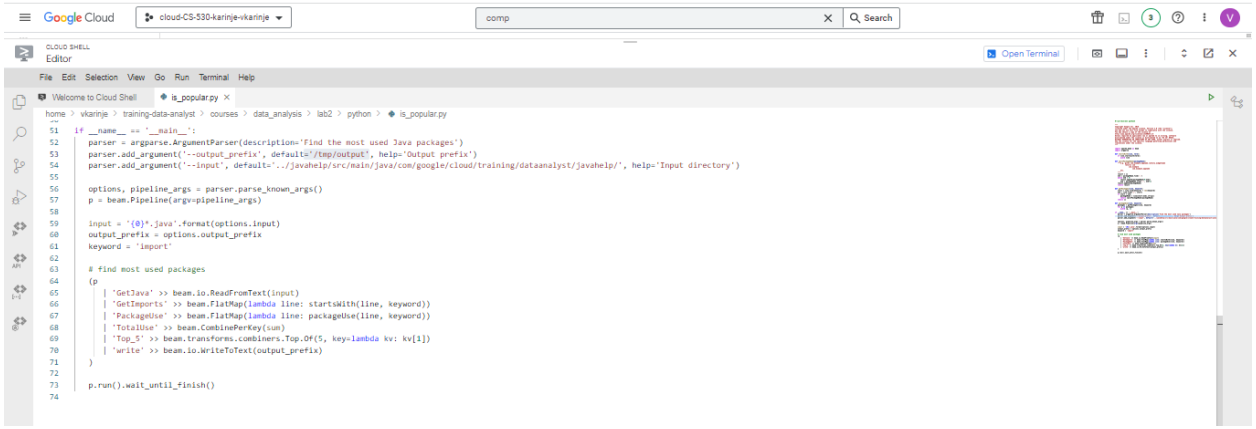
- Where is the input taken from by default?

Ans: The input is taken by default from `../javahelp/src/main/java/com/google/cloud/training/dataanalyst/javahelp/'` Directory



- Where does the output go by default?

`'/tmp/output` is where the output goes by default.



- 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

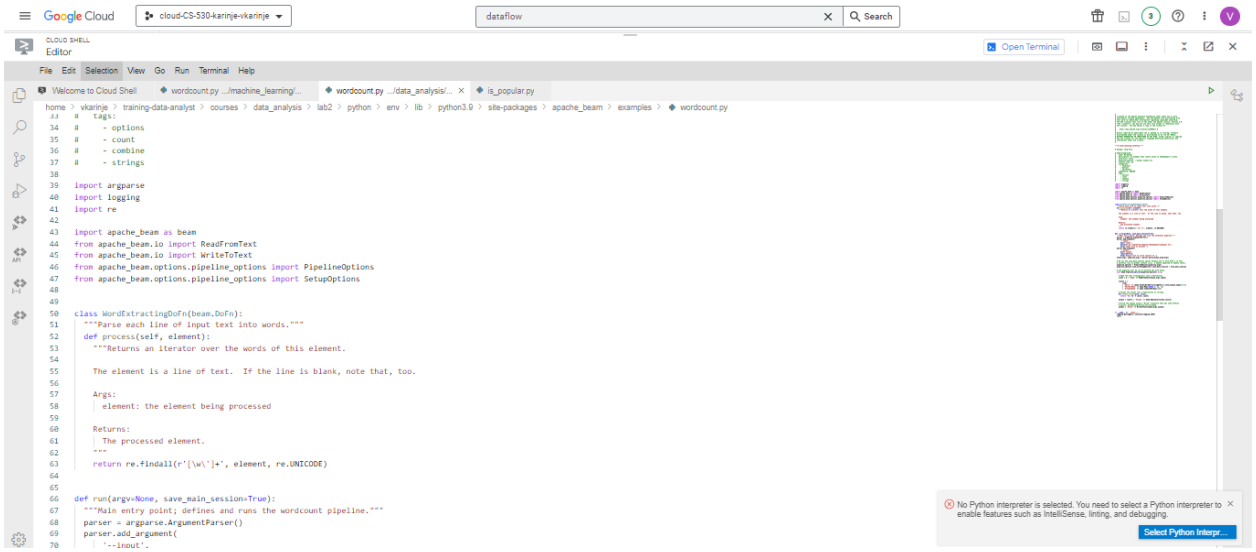
- Take a screenshot of its contents

```
(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.

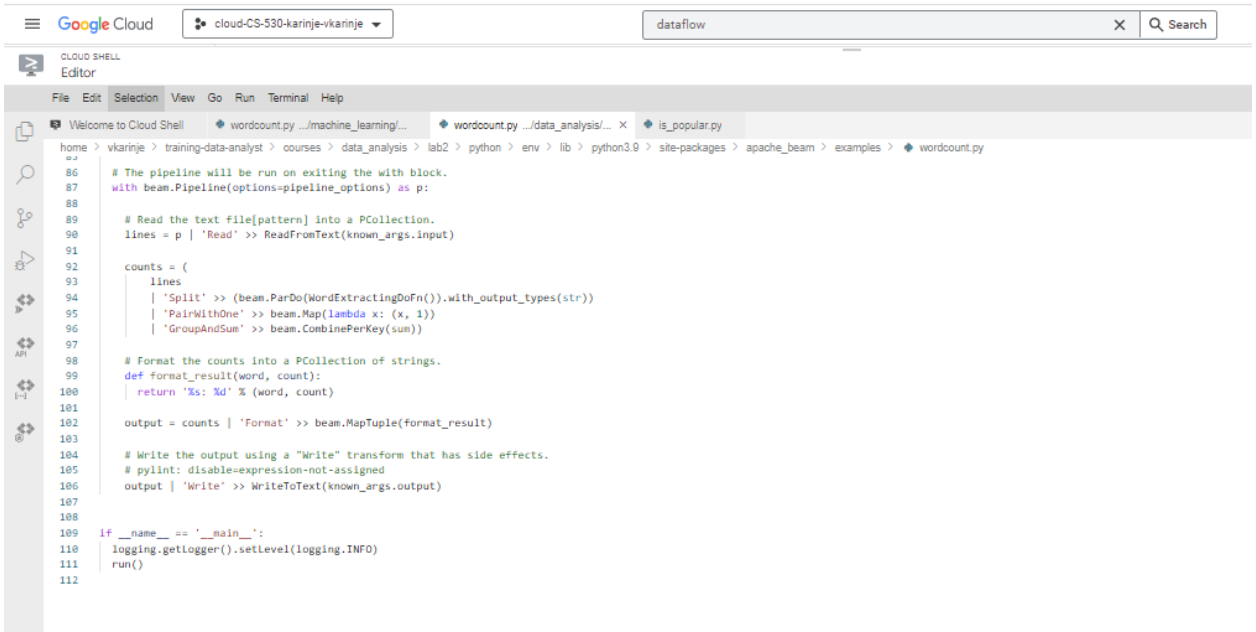
Ans: The data in this output file corresponds to the most used packages in the java files present in the directory `../javahelp/src/main/java/com/google/cloud/training/dataanalyst/javahelp/'`

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



- 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.

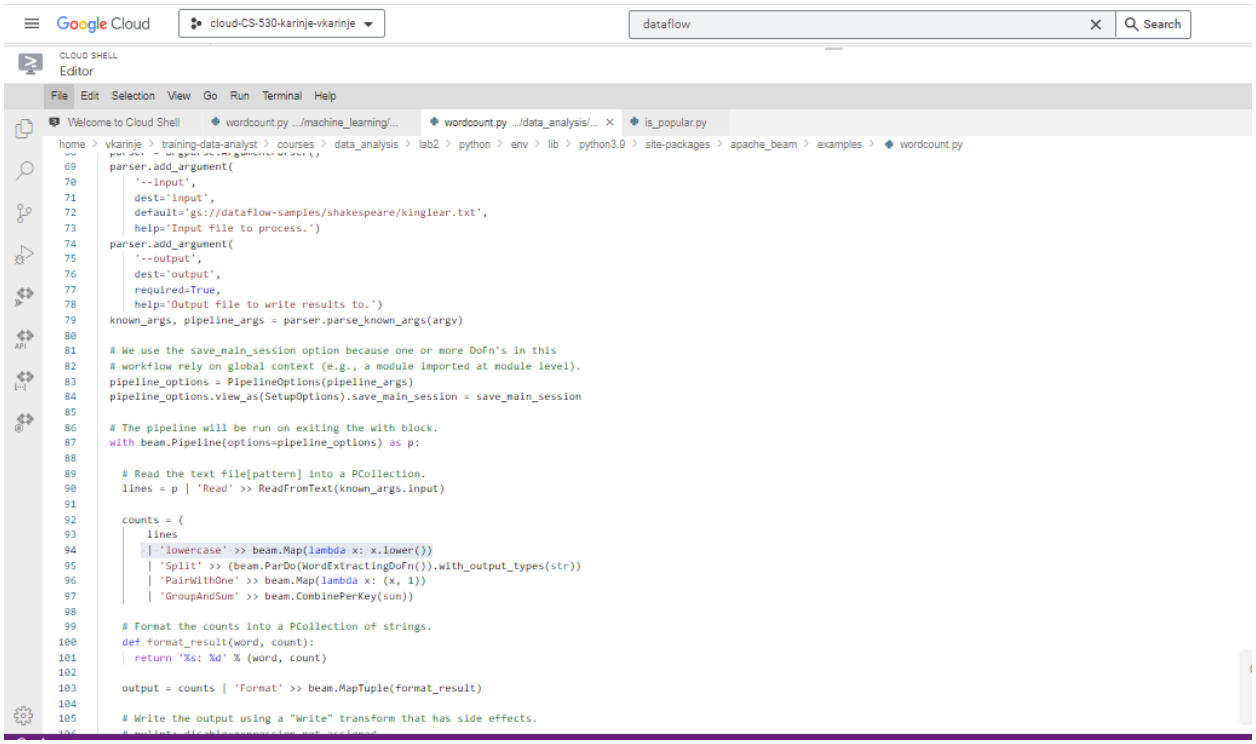
```
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje)$ edit env/lib/python3.9/site-packages/apache_beam/examples/wordcount.py
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje)$ python -m apache_beam.examples.wordcount \
  --output outputs
INFO:root:Missing pipeline option (runner). Executing pipeline using the default runner: DirectRunner.
INFO:apache_beam.internal.gcp.auth:Setting socket default timeout to 60 seconds.
INFO:apache_beam.internal.gcp.auth:socket default timeout is 60.0 seconds.
INFO:root:Default Python SDK image for environment is apache/beam_python3.9_sdk:2.42.0
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function annotate_downstream_side_inputs at 0x7f59f2b67b80> ==================
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function fix_side_input_pcoll_coders at 0x7f59f2b67ca0> ==================
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function pack_combiners at 0x7f59f2b681f0> ==================
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function lift_combiners at 0x7f59f2b68280> ==================
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function expand_sdf at 0x7f59f2b68430> ==================
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function expand_gbk at 0x7f59f2b684c0> ==================
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function sink_flattens at 0x7f59f2b685e0> ==================
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function greedily_fuse at 0x7f59f2b68670> ==================
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function read_to_impulse at 0x7f59f2b68700> ==================
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function impulse_to_input at 0x7f59f2b68790> ==================
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function sort_stages at 0x7f59f2b689d0> ==================
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function add_impulse_to_dangling_transforms at 0x7f59f2b68af0> ==================
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function setup_timer_mapping at 0x7f59f2b68940> ==================
INFO:apache_beam.runners.portability.fn_api_runner.translations:================== <function populate_data_channel_coders at 0x7f59f2b68a60> ==================
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.EmbeddedWorkerHandler object at 0x7f59f297cd00> for environment ref_Environment_default_environment_1 (be
am:env:embedded_python:v1, b'')
INFO:apache_beam.io.filebasedsink:Starting finalize_write threads with num_shards: 1 (skipped: 0), batches: 1, num_threads: 1
INFO:apache_beam.io.filebasedsink:Renamed 1 shards in 0.00 seconds.
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje)$ wan wc
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje)$ wc -w outputs-00000-of-00001
9568 outputs-00000-of-00001
(env) vkarinje@cloudshell:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python (cloud-cs-530-karinje-vkarinje)$||
```

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

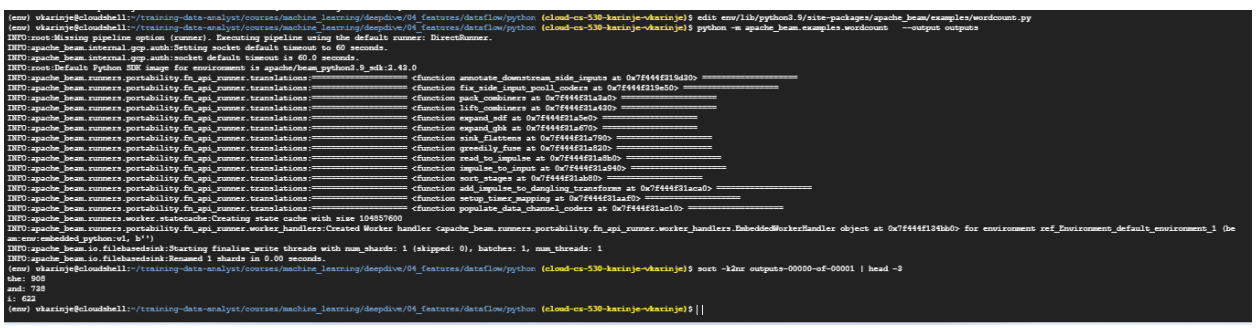
The top 3 words in King Lear are the, I & and
the-786
I- 622
and- 594
 Find a place in the pipeline that you can insert a stage that transforms all of the characters it receives into lowercase.



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:



  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/acat.p2-791612085972-647a32d2-e9a7-4678-9589-13ac58291e04" 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:
```

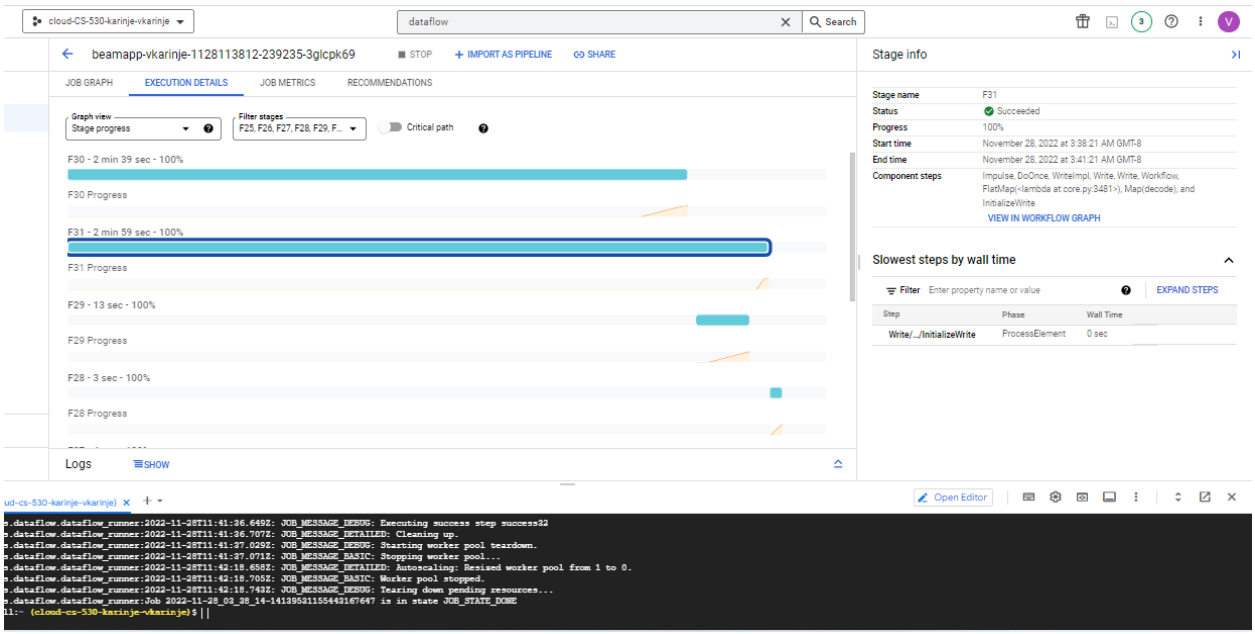
```
(env) vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ gcloud iam service-accounts keys create df-lab.json --iam-account df-lab@${GOOGLE_CLOUD_PROJECT}.iam.gserviceaccount.com
created key [0c57fa0ac575b3099da3ce2c18316a9f509515da] of type [json] as [df-lab.json] for [df-lab@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com]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

Run code using Dataflow runner

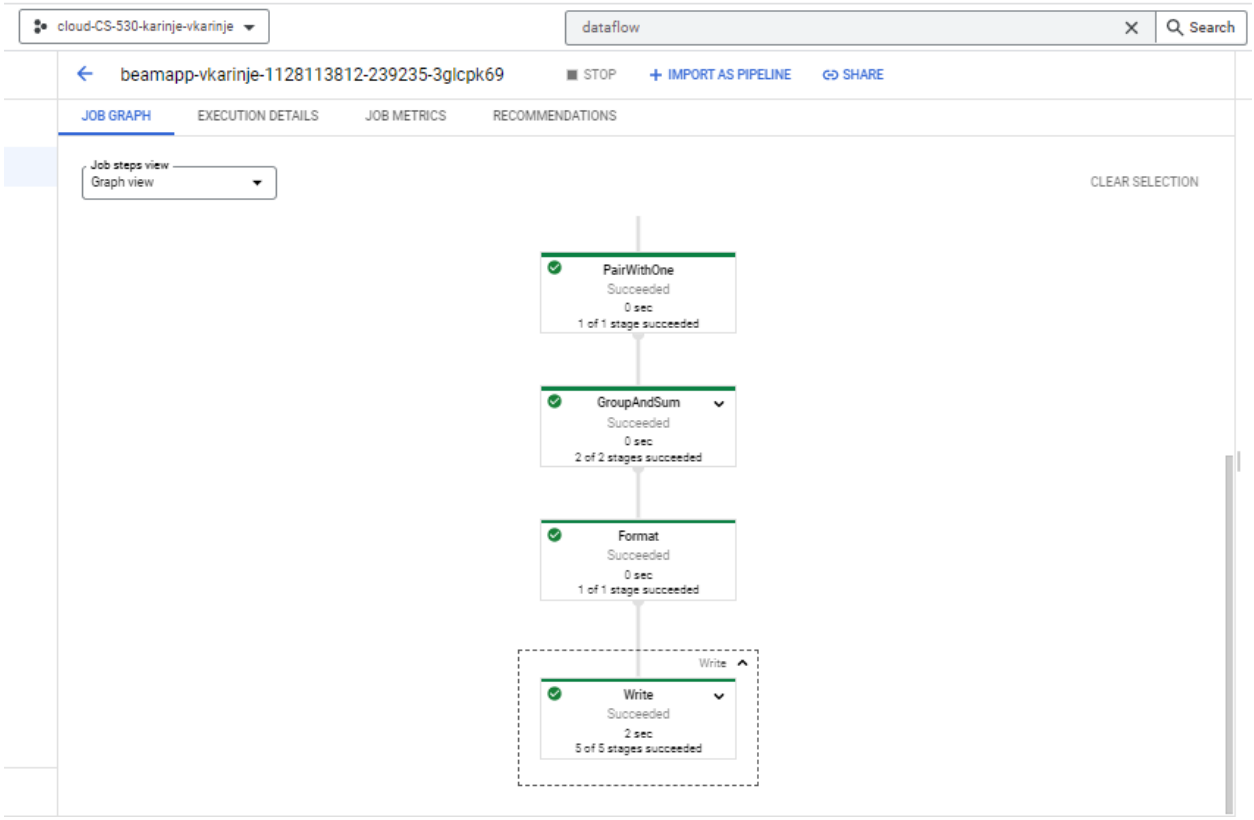
Include the following in your lab notebook:

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

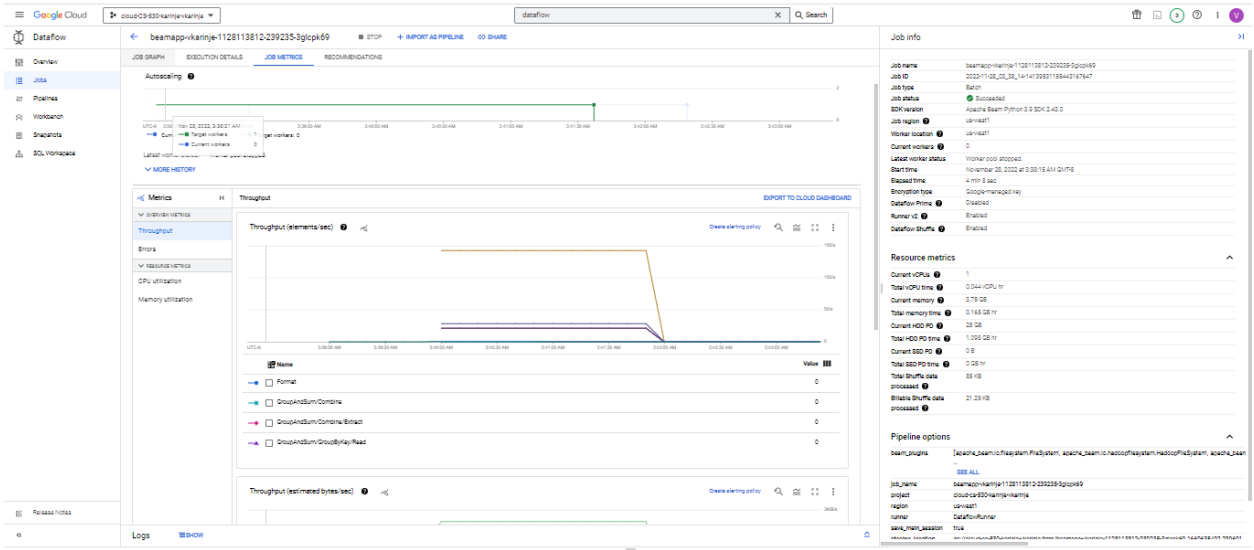
Ans: F31



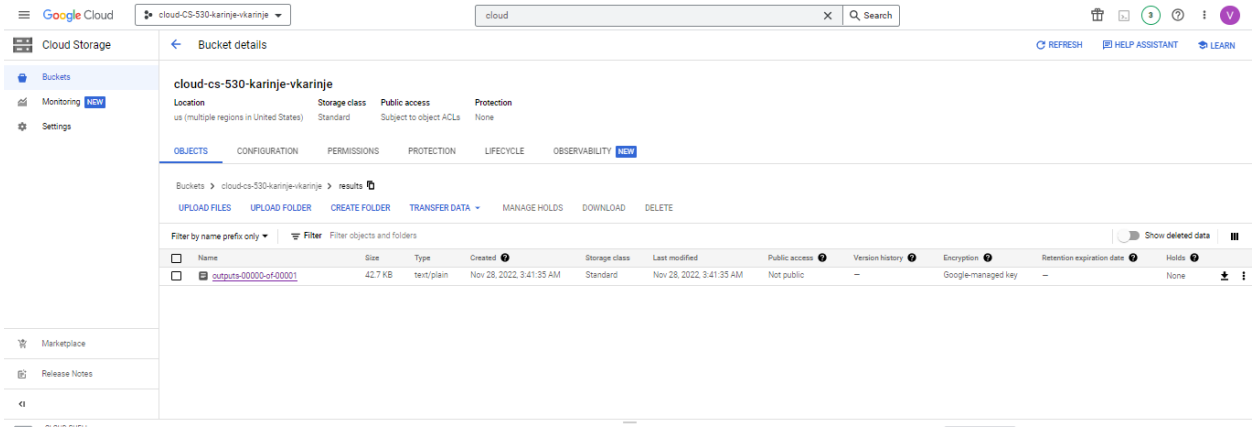
Write has taken the longest time to complete



- 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?



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

Clean up

```
  - serviceAccount:guestbook@cloud-cs-530-karinje-vkarinje.iam.gserviceaccount.com
  role: roles/storage.objectViewer
etag: BwXuhpd2i14=
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)$ gsutil -m rm -r gs://${BUCKET}
Removing gs://cloud-cs-530-karinje-vkarinje/results/outputs-00000-of-00001#1669635694987227...
Removing gs://cloud-cs-530-karinje-vkarinje/tmp/beamapp-vkarinje-1128113812-239235-3glcpk69.1669635492.239601/pickled_main_session#1669635492574421...
Removing gs://cloud-cs-530-karinje-vkarinje/tmp/beamapp-vkarinje-1128113812-239235-3glcpk69.1669635492.239601/apache_beam-2.43.0-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl#1669635493673368...
Removing gs://cloud-cs-530-karinje-vkarinje/tmp/beamapp-vkarinje-1128113812-239235-3glcpk69.1669635492.239601/dataflow_python_sdk.tar#1669635492942016...
Removing gs://cloud-cs-530-karinje-vkarinje/tmp/beamapp-vkarinje-1128113812-239235-3glcpk69.1669635492.239601/tmp-f416e2f582270cc8-00000-of-00001.sdfmeta#1669635659871883...
Removing gs://cloud-cs-530-karinje-vkarinje/tmp/beamapp-vkarinje-1128113812-239235-3glcpk69.1669635492.239601/pipeline.pb#1669635493870018...
/ [6/6 objects] 100% Done
Operation completed over 6 objects.
Removing gs://cloud-cs-530-karinje-vkarinje/...
(env) vkarinje@cloudshell:~ (cloud-cs-530-karinje-vkarinje)$ ||
```

09.2g: BigQuery, JupyterLab