Cloud Computing
(SOFE 4630U)

Project Milestone 3: Data Processing: Dataflow/Apache Beam

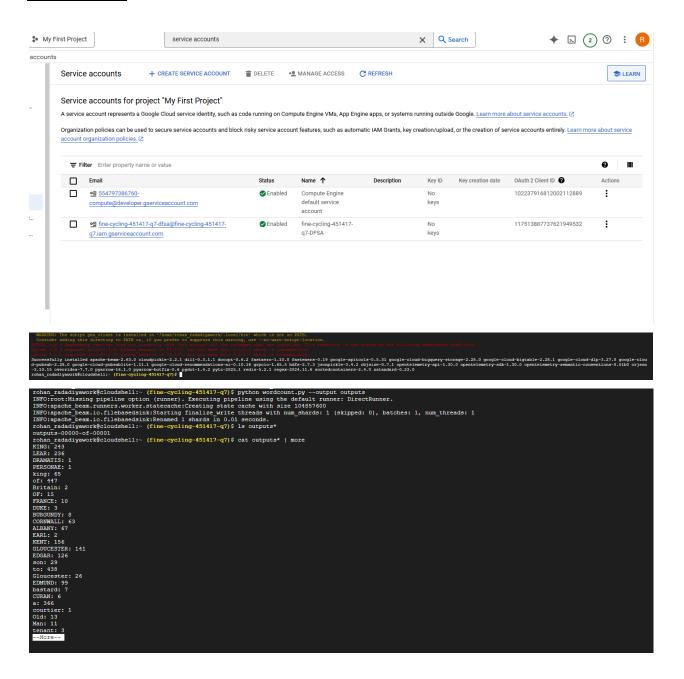| Name	Student ID
Rohan Radadiya	100704614

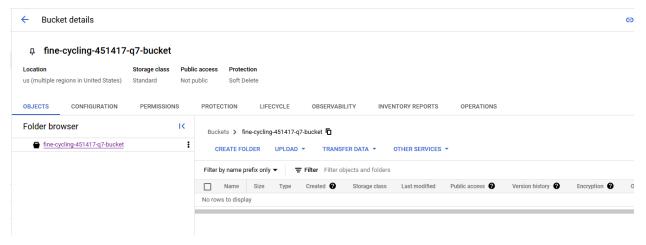**Due Date:** February 24, 2025

GitHub Link:

https://github.com/rohanradadiya/Cloud-Computing-Project-MS-3

Google Drive for Videos:

https://drive.google.com/drive/folders/1HcPcpJIkzAwhymcY-gjWLjAgn-H_Qgwy?usp=sharing
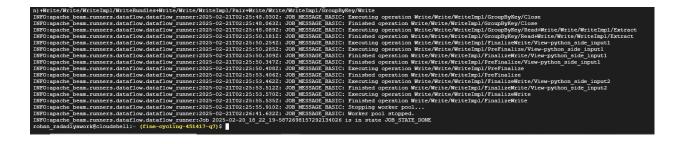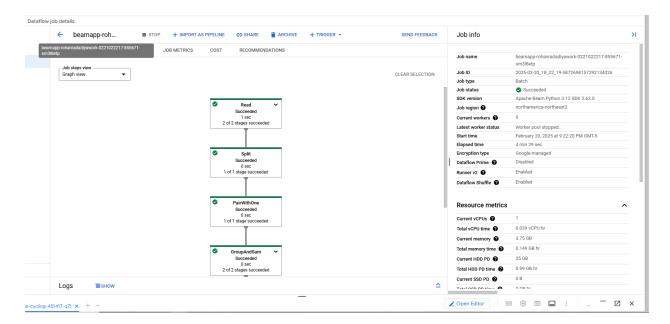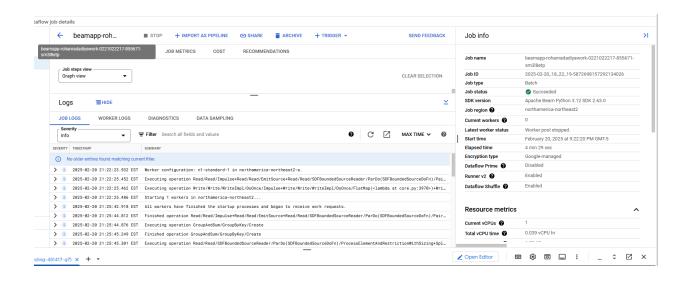
Screenshots:


```
rohan_radadiyawork@cloudshell:~ (fine-cycling-451417-q7)$ python wordcount.py --output outputs
INFO:root:Missing pipeline option (runner). Executing pipeline using the default runner: DirectRunner.
INFO:apache_beam.runners.worker.statecache:Creating state cache with size 104857600
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 0.01 seconds.
rohan_radadiyawork@cloudshell:~ (fine-cycling-451417-q7)$ ls outputs*
outputs-00000-of-00001
rohan_radadiyawork@cloudshell:~ (fine-cycling-451417-q7)$ cat outputs* | more
KING: 243
LEAR: 236
DRAMATIS: 1
PERSONAE: 1
king: 65
of: 447
Britain: 2
OF: 15
FRANCE: 10
DUKE: 3
BURGUNDY: 8
CORNWALL: 63
ALBANY: 67
EARL: 2
KENT: 156
GLOUCESTER: 141
EDGAR: 126
son: 29
to: 438
Gloucester: 26
EDMUND: 99
bastard: 7
CURAN: 6
a: 366
courtier: 1
Old: 13
Man: 11
tenant: 3
--More--
```

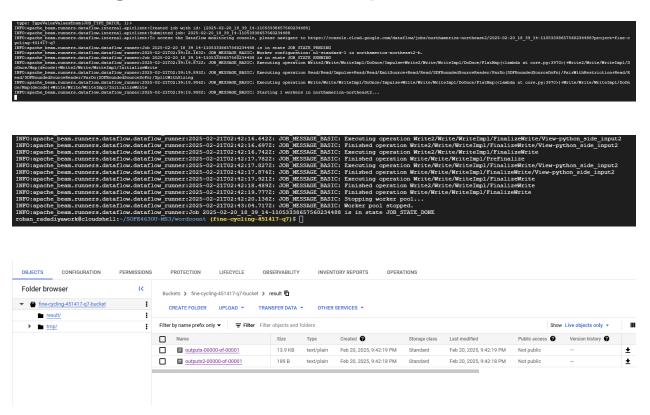**Bucket created:**

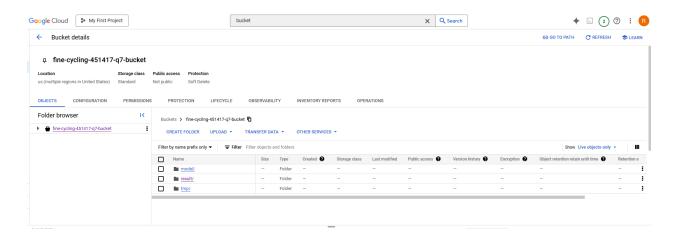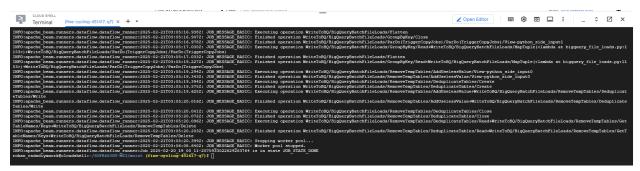**Environment variable for bucket created:**

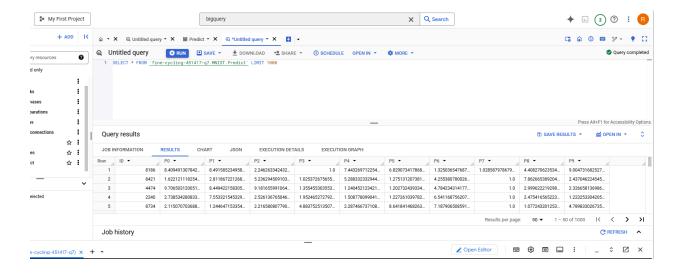# 2. Running the wordcount2 Example

3. Batch Processing of the MNIST Dataset

Model folder uploaded:

Running Dataflow after uploading the /mnist/model folder:

4. Stream Processing the MNIST Dataset

Topic ID ↑	Encryption key	Topic name	Retention	Ingestion source
☐ mnist_image	Google-managed	projects/fine-cycling-451417-q7/topics/mnist_image	–	–
☐ mnist_predict	Google-managed	projects/fine-cycling-451417-q7/topics/mnist_predict	–	–

"mnist_image" and "mnist_predict" topics created:

```
location: 'northamerica-northeast2'
name: 'beamapp-rohanradadiyawork-0221032142-299956-in71akj0'
projectId: 'fine-cycling-451417-q7'
stageStates: []
startTime: '2025-02-21T03:21:44.100748Z'
steps: []
tempFiles: []
type: TypeValueValuesEnum(JOB_TYPE_STREAMING, 2)>
INFO:apache_beam.runners.dataflow.internal.apiclient:Created job with id: [2025-02-20_19_21_43-13316342389863091699]
INFO:apache_beam.runners.dataflow.internal.apiclient:Submitted job: 2025-02-20_19_21_43-13316342389863091699
INFO:apache_beam.runners.dataflow.internal.apiclient:To access the Dataflow monitoring console, please navigate to https://console.cloud.google.com/dataflow/jobs/northamerica-northeast2/2025-02-20_19_21_43-13316342389863091699?project=fine-c
ycling-451417-q7
INFO:apache_beam.runners.dataflow.dataflow_runner:Job 2025-02-20_19_21_43-13316342389863091699 is in state JOB_STATE_PENDING
WARNING:apache_beam.runners.dataflow.dataflow_runner:2025-02-21T03:21:45.1952: JOB_MESSAGE_WARNING: Autoscaling is enabled for Dataflow Streaming Engine. Workers will scale between 1 and 100 unless maxNumWorkers is specified.
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-02-21T03:21:49.413Z: JOB_MESSAGE_BASIC: Worker configuration: n1-standard-2 in northamerica-northeast2-b.
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-02-21T03:21:51.036Z: JOB_MESSAGE_BASIC: Running job using Streaming Engine
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-02-21T03:21:51.218Z: JOB_MESSAGE_BASIC: Starting 1 workers in northamerica-northeast2...
INFO:apache_beam.runners.dataflow.dataflow_runner:Job 2025-02-20_19_21_43-13316342389863091699 is in state JOB_STATE_RUNNING
INFO:apache_beam.runners.dataflow.dataflow_runner:2025-02-21T03:25:27.092Z: JOB_MESSAGE_BASIC: All workers have finished the startup processes and began to receive work requests.
[]
```
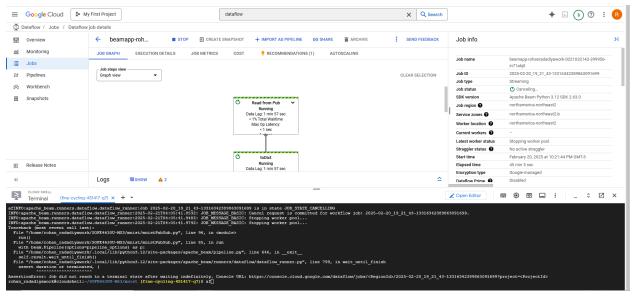
Running the Data flow job that reads JSON objects from the mnist_image topic, applying it to the ML model, and sending the prediction results via the mnist_predict topic:

```
25
26    df=pd.read_csv('mnist.csv')

PROBLEMS 1    OUTPUT    DEBUG CONSOLE    TERMINAL    PORTS

[notice] To update, run: python.exe -m pip install --upgrade pip
PS C:\Users\rohan\Downloads\mnist\data> python producerMnistPubSup.py
Image with key 0 is sent
Image with key 1 is sent
Image with key 2 is sent
Image with key 3 is sent
Image with key 4 is sent
Image with key 5 is sent
Image with key 6 is sent
Image with key 7 is sent
Image with key 8 is sent
Image with key 9 is sent
Image with key 10 is sent
Image with key 11 is sent
Image with key 12 is sent
Image with key 13 is sent
Image with key 14 is sent
Image with key 15 is sent
Image with key 16 is sent
Image with key 17 is sent
Image with key 18 is sent
Image with key 19 is sent
Image with key 20 is sent
Image with key 21 is sent
Image with key 22 is sent
Image with key 23 is sent
Image with key 24 is sent
Image with key 25 is sent
Image with key 26 is sent
Image with key 27 is sent
Image with key 28 is sent
Image with key 29 is sent
```

After updating producerMnistPubSup.py script and running it:

After updating consumerMnistPubSup.py script and running it:

Dataflow job is stopped manually:

Design:

```
Google Cloud SDK Shell                                                    —   □   ×

---
C:\Users\rohan\AppData\Local\Google\Cloud SDK>gsutil version
gsutil version: 5.33

C:\Users\rohan\AppData\Local\Google\Cloud SDK>cd C:\Users\rohan\Documents\pedestrian_detection_pipeline

C:\Users\rohan\Documents\pedestrian_detection_pipeline>gsutil -m cp -r . gs://fine-cycling-451417-q7-bucket/
Copying file://.\models\depth_model.pth [Content-Type=application/octet-stream]...
Copying file://.\models\yolov8n.pt [Content-Type=application/octet-stream]...
Copying file://.\src\object_detection.py [Content-Type=text/x-python]...
Copying file://.\src\pipeline.py [Content-Type=text/x-python]...
Copying file://.\Dockerfile [Content-Type=application/octet-stream]...
Copying file://.\pubsub_publisher.py [Content-Type=text/x-python]...
Copying file://.\data\Labels.csv [Content-Type=application/vnd.ms-excel]...
Copying file://.\pedestrian_detection_pipeline_requirements.txt [Content-Type=text/plain]...
Copying file://.\src\depth_estimation.py [Content-Type=text/x-python]...
Copying file://.\data\Dataset_Occluded_Pedestrian\A_085.png [Content-Type=image/png]...
Copying file://.\data\Dataset_Occluded_Pedestrian\A_020.png [Content-Type=image/png]...
Copying file://.\data\images\depth.jpg [Content-Type=image/jpeg]...
Copying file://.\pubsub_subscriber.py [Content-Type=text/x-python]...
Copying file://.\data\Dataset_Occluded_Pedestrian\C_020.png [Content-Type=image/png]...
Copying file://.\data\images\problem.jpg [Content-Type=image/jpeg]...
Copying file://.\data\images\ped_depth.jpg [Content-Type=image/jpeg]...
Copying file://.\data\images\yolo.jpg [Content-Type=image/jpeg]...
/ [17/17 files][  3.2 MiB/  3.2 MiB] 100% Done
Operation completed over 17 objects/3.2 MiB.

C:\Users\rohan\Documents\pedestrian_detection_pipeline>
C:\Users\rohan\Documents\pedestrian_detection_pipeline>
```

```
[notice] A new release of pip is available: 25.0 -> 25.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip

C:\Users\rohan\Documents\pedestrian_detection_pipeline>
C:\Users\rohan\Documents\pedestrian_detection_pipeline>set PYTHONPATH=C:\Users\rohan\Documents\pedestrian_detection_pip
eline\src

C:\Users\rohan\Documents\pedestrian_detection_pipeline>
C:\Users\rohan\Documents\pedestrian_detection_pipeline>python src/pipeline.py --runner DataflowRunner --project fine-cy
cling-451417-q7 --staging_location gs://fine-cycling-451417-q7-bucket/staging --temp_location gs://fine-cycling-451417-
q7-bucket/temp --region northamerica-northeast2 --template_location gs://fine-cycling-451417-q7-bucket/templates/pedest
rian_detection_template
Traceback (most recent call last):
  File "C:\Users\rohan\Documents\pedestrian_detection_pipeline\src\pipeline.py", line 5, in <module>
    from src.object_detection import detect_pedestrians
ModuleNotFoundError: No module named 'src'

C:\Users\rohan\Documents\pedestrian_detection_pipeline>
C:\Users\rohan\Documents\pedestrian_detection_pipeline>set PYTHONPATH=C:\Users\rohan\Documents\pedestrian_detection_pip
eline

C:\Users\rohan\Documents\pedestrian_detection_pipeline>python src/pipeline.py --runner DataflowRunner --project fine-cy
cling-451417-q7 --staging_location gs://fine-cycling-451417-q7-bucket/staging --temp_location gs://fine-cycling-451417-
q7-bucket/temp --region northamerica-northeast2 --template_location gs://fine-cycling-451417-q7-bucket/templates/pedest
rian_detection_template
```

```
C:\Users\rohan\Documents\pedestrian_detection_pipeline>
C:\Users\rohan\Documents\pedestrian_detection_pipeline>curl -L -o models/yolov8n.pt https://github.com/ultralytics/yolo
v8/releases/latest/download/yolov8n.pt
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100     9  100     9    0     0     28      0 --:--:-- --:--:-- --:--:--    28

C:\Users\rohan\Documents\pedestrian_detection_pipeline>python src/pipeline.py --runner DataflowRunner --project fine-cy
cling-451417-q7 --staging_location gs://fine-cycling-451417-q7-bucket/staging --temp_location gs://fine-cycling-451417-
q7-bucket/temp --region northamerica-northeast2 --template_location gs://fine-cycling-451417-q7-bucket/templates/pedest
rian_detection_template
```

As seen from the screenshots above, the necessary setup was done in order to be able to run the dataflow job. The correct object detection model was also installed and attempted, however there were issues in being able to detect it. The yolov8n.pt was downloaded and set up, but the issue related to being able to connect that with the project and then run the dataflow job. There were also some issues in getting the job to appear on the Google Cloud, but with some more debugging and work, the dataflow job would be able to run successfully for this milestone. However, the screenshots for both parts of the milestone along with the videos are attached to the submission.