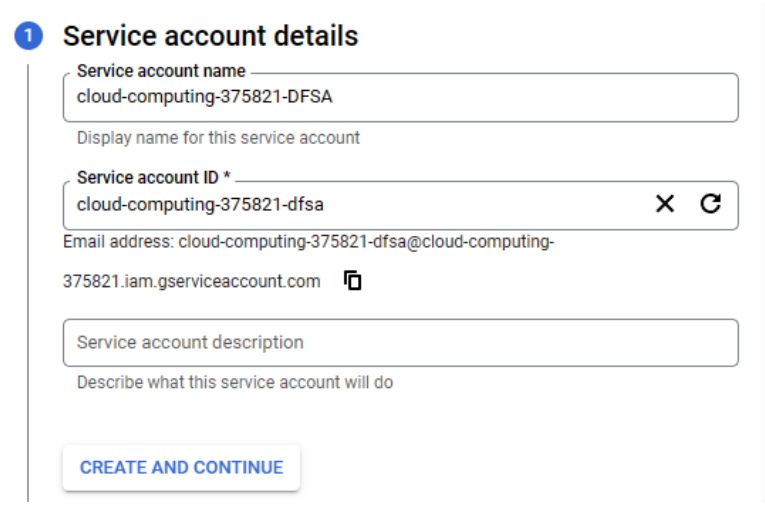
Project Milestone-- Data Processing: Dataflow- apache beam
Aaditya Rajput
100622434

Created new service account:



Not shown but also added Compute Engine Service agent and Pub/Sub Admin.

Below I installed the necessary python Libraries:

```
aadiraj092@cloudshell:~ (cloud-computing-375821)$  pip install pip --upgrade
Requirement already satisfied: pip in /usr/lib/python3/dist-packages (20.3.4)
Collecting pip
  Downloading pip-23.0-py3-none-any.whl (2.1 MB)
     |████████████████████████████████| 2.1 MB 4.4 MB/s
Installing collected packages: pip
  WARNING: The scripts pip, pip3 and pip3.9 are installed in '/home/aadiraj092/.local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed pip-23.0
aadiraj092@cloudshell:~ (cloud-computing-375821)$  pip install 'apache-beam[gcp]'
Defaulting to user installation because normal site-packages is not writeable
Collecting apache-beam[gcp]
  Downloading apache_beam-2.45.0-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (14.3 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 14.3/14.3 MB 46.0 MB/s eta 0:00:00
Collecting pymongo<4.0.0,>=3.8.0
  Downloading pymongo-3.13.0-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (515 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 515.5/515.5 kB 44.9 MB/s eta 0:00:00
Collecting orjson<4.0
  Downloading orjson-3.8.6-cp39-cp39-manylinux_2_28_x86_64.whl (140 kB)
```

Running wordcount.py and then displaying the outputs:

```
aadiraj092@cloudshell:~ (cloud-computing-375821)$ python -wordcount.py --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:2.45.0
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function annotate_downstream_side_inputs at 0x7f698d2aa5e0> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function fix_side_input_pcoll_coders at 0x7f698d2aa700> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function pack_combiners at 0x7f698d2aac10> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function lift_combiners at 0x7f698d2aaca0> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function expand_sdf at 0x7f698d2aae50> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function expand_gbk at 0x7f698d2aaee0> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function sink_flattens at 0x7f698d2ab040> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function greedily_fuse at 0x7f698d2ab0d0> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function read_to_impulse at 0x7f698d2ab160> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function impulse_to_input at 0x7f698d2ab1f0> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function sort_stages at 0x7f698d2ab430> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function add_impulse_to_dangling_transforms at 0x7f698d2ab550> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function setup_timer_mapping at 0x7f698d2ab3a0> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function populate_data_channel_coders at 0x7f698d2ab4c0> ====================
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 0x7f698d24e430> for environment ref_Environment_
fault_environment_1 (beam: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 0.01 seconds.
aadiraj092@cloudshell:~ (cloud-computing-375821)$ ls outputs*
outputs-00000-of-00001
```

Created a bucket below:

cloud-computing-375821-bucket

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	None

OBJECTS | CONFIGURATION | PERMISSIONS | PROTECTION | LIFECYCLE | OBSERVABILITY NEW

Buckets > cloud-computing-375821-bucket

Ran the pipeline using dataflow and as you can see it says JOB_STATE_DONE.

```
inalizeWrite
INFO:apache_beam.runners.dataflow.dataflow_runner:2023-02-17T02:07:38.408Z: JOB_MESSAGE_DEBUG: Value "Write/Write/WriteImpl/Finaliz
.out" materialized.
INFO:apache_beam.runners.dataflow.dataflow_runner:2023-02-17T02:07:38.435Z: JOB_MESSAGE_BASIC: Executing operation Write/Write/Writ
INFO:apache_beam.runners.dataflow.dataflow_runner:2023-02-17T02:07:40.865Z: JOB_MESSAGE_BASIC: Finished operation Write/Write/Write
INFO:apache_beam.runners.dataflow.dataflow_runner:2023-02-17T02:07:40.884Z: JOB_MESSAGE_DEBUG: Executing success step success32
INFO:apache_beam.runners.dataflow.dataflow_runner:2023-02-17T02:07:40.920Z: JOB_MESSAGE_DETAILED: Cleaning up.
INFO:apache_beam.runners.dataflow.dataflow_runner:2023-02-17T02:07:41.218Z: JOB_MESSAGE_DEBUG: Starting worker pool teardown.
INFO:apache_beam.runners.dataflow.dataflow_runner:2023-02-17T02:07:41.227Z: JOB_MESSAGE_BASIC: Stopping worker pool...
INFO:apache_beam.runners.dataflow.dataflow_runner:2023-02-17T02:08:23.866Z: JOB_MESSAGE_DETAILED: Autoscaling: Resized worker pool
INFO:apache_beam.runners.dataflow.dataflow_runner:2023-02-17T02:08:23.884Z: JOB_MESSAGE_BASIC: Worker pool stopped.
INFO:apache_beam.runners.dataflow.dataflow_runner:2023-02-17T02:08:23.894Z: JOB_MESSAGE_DEBUG: Tearing down pending resources...
INFO:apache_beam.runners.dataflow.dataflow_runner:Job 2023-02-16_18_03_40-15509227898021680147 is in state JOB_STATE_DONE
aadiraj092@cloudshell:~ (cloud-computing-375821)$ $S
```

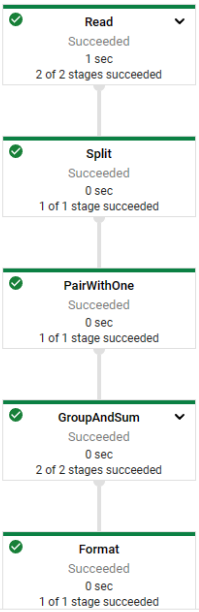
Dataflow Jobs:

Jobs | CREATE JOB FROM TEMPLATE | ENABLE SORTING | REFRESH | LEARN

Running | Filter | Filter jobs

Name	Type	End time	Elapsed time	Start time	Status	SDK version	ID	Region	Insights
beamapp-aadiraj092-0217020337-994619-7i1whdnd	Batch	Feb 16, 2023, 9:08:28 PM	4 min 47 sec	Feb 16, 2023, 9:03:41 PM	Succeeded	2.45.0	2023-02-16_18_03_40-15509227898021680147	northamerica-northeast2	1 INSIGHT

Job steps view: Graph view



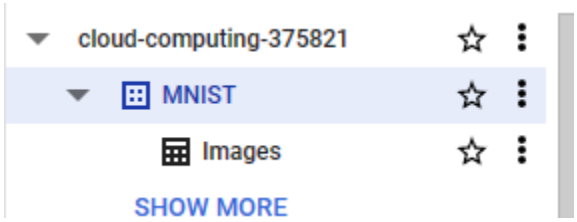
Clone github:

```
aadiraj092@cloudshell:~ (cloud-computing-375821)$ git clone https://github.com/GeorgeDaoud3/SOFE4630U-MS2.git
Cloning into 'SOFE4630U-MS2'...
remote: Enumerating objects: 175, done.
remote: Counting objects: 100% (175/175), done.
remote: Compressing objects: 100% (158/158), done.
remote: Total 175 (delta 35), reused 117 (delta 13), pack-reused 0
Receiving objects: 100% (175/175), 18.50 MiB | 21.40 MiB/s, done.
Resolving deltas: 100% (35/35), done.
aadiraj092@cloudshell:~ (cloud-computing-375821)$
```

Job Status Done:

```
INFO:apache_beam.runners.dataflow.dataflow_runner:2023-02-17T02:49:02.674Z: JOB_MESSAGE_DEBUG: Tearing down pending resources...
INFO:apache_beam.runners.dataflow.dataflow_runner:Job 2023-02-16_18_44_35-5528749527164652543 is in state JOB_STATE_DONE
aadiraj092@cloudshell:~/SOFE4630U-MS2/wordcount (cloud-computing-375821)$
```

MNIST Dataset:



After: