Pranjal Saloni 100653360

Setting up the GCP environment for Dataflow and running wordcount examples:
https://youtu.be/re6c_ee7uTc
Screenshots:

```
pranjal_saloni612@cloudshell:~ (projectmil1)$ python3 --version
Python 3.9.2
pranjal_saloni612@cloudshell:~ (projectmil1)$ python3 -m venv env
pranjal_saloni612@cloudshell:~ (projectmil1)$ ls
env  index.html  README-cloudshell.txt  SOFE4630U-tut3
pranjal_saloni612@cloudshell:~ (projectmil1)$ source ~/env/bin/activate
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ deactivate
pranjal_saloni612@cloudshell:~ (projectmil1)$ source ~/env/bin/activate
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ python --version
Python 3.9.2
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ pip install pip --upgrade
Requirement already satisfied: pip in ./env/lib/python3.9/site-packages (20.3.4)
Collecting pip
  Downloading pip-22.0.4-py3-none-any.whl (2.1 MB)
     |████████████████████████████████| 2.1 MB 5.1 MB/s
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-22.0.4
```

```
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ pip install 'apache-beam[gcp]'
Collecting apache-beam[gcp]
  Downloading apache_beam-2.37.0-cp39-cp39-manylinux2010_x86_64.whl (11.1 MB)
     ---------------------------------------- 11.1/11.1 MB 42.2 MB/s eta 0:00:00
Collecting pytz>=2018.3
  Downloading pytz-2022.1-py2.py3-none-any.whl (503 kB)
     ---------------------------------------- 503.5/503.5 KB 39.6 MB/s eta 0:00:00
Collecting requests<3.0.0,>=2.24.0
  Downloading requests-2.27.1-py2.py3-none-any.whl (63 kB)
     ---------------------------------------- 63.1/63.1 KB 8.9 MB/s eta 0:00:00
Collecting protobuf<4,>=3.12.2
  Downloading protobuf-3.19.4-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.1 MB)
     ---------------------------------------- 1.1/1.1 MB 59.9 MB/s eta 0:00:00
Collecting crcmod<2.0,>=1.7
  Downloading crcmod-1.7.tar.gz (89 kB)
     ---------------------------------------- 89.7/89.7 KB 12.2 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting pymongo<4.0.0,>=3.8.0
  Downloading pymongo-3.12.3-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (516 kB)
     ---------------------------------------- 516.3/516.3 KB 39.3 MB/s eta 0:00:00
Collecting typing-extensions>=3.7.0
  Downloading typing_extensions-4.1.1-py3-none-any.whl (26 kB)
Collecting pyarrow<7.0.0,>=0.15.1
  Downloading pyarrow-6.0.1-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (25.6 MB)
     ---------------------------------------- 25.6/25.6 MB 31.1 MB/s eta 0:00:00
Collecting fastavro<2,>=0.23.6
  Downloading fastavro-1.4.10-cp39-cp39-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (2.5 MB)
     ---------------------------------------- 2.5/2.5 MB 77.7 MB/s eta 0:00:00
Collecting numpy<1.22.0,>=1.14.3
  Downloading numpy-1.21.5-cp39-cp39-manylinux_2_12_x86_64.manylinux2010_x86_64.whl (15.7 MB)
     ---------------------------------------- 15.7/15.7 MB 45.1 MB/s eta 0:00:00
Collecting oauth2client<5,>=2.0.1
```

```
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ 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:oauth2client.transport:Attempting refresh to obtain initial access_token
WARNING:root:Make sure that locally built Python SDK docker image has Python 3.9 interpreter.
INFO:root:Default Python SDK image for environment is apache/beam_python3.9_sdk:2.37.0
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function annotate_downstream_side_inputs at 0x7f9ff335bdc0> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function fix_side_input_pcoll_coders at 0x7f9ff335bee0> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function pack_combiners at 0x7f9ff335d430> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function lift_combiners at 0x7f9ff335d4c0> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function expand_sdf at 0x7f9ff335d670> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function expand_gbk at 0x7f9ff335d700> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <function sink_flattens at 0x7f9ff335d820> ====================
```

```
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ ls
env  index.html  outputs-00000-of-00001  README-cloudshell.txt  SOFE4630U-tut3
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ more outputs-00000-of-00001
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
```

```
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ PROJECT=projectmil1
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ echo $PROJECT
projectmil1
```


```
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ BUCKET=gs://projectmil1-gs
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ echo $BUCKET
gs://projectmil1-gs
```

```
(env) pranjal_saloni612@cloudshell:~ (projectmill)$ python -m apache_beam.examples.word
ount --project $PROJECT --region northamerica-northeast2 --runner Dataflowrunner --temp
location $BUCKET/temp/ --output $BUCKET/result/outputs
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:oauth2client.transport:Attempting refresh to obtain initial access_token
INFO:apache_beam.runners.portability.stager:Downloading source distribution of the SDK
rom PyPi
INFO:apache_beam.runners.portability.stager:Executing command: ['/home/pranjal_saloni61
/env/bin/python', '-m', 'pip', 'download', '--dest', '/tmp/tmp_sxaw0w7', 'apache-beam==
.37.0', '--no-deps', '--no-binary', ':all:']
INFO:apache_beam.runners.portability.stager:Staging SDK sources from PyPI: dataflow_pyt
on_sdk.tar
INFO:apache_beam.runners.portability.stager:Downloading binary distribution of the SDK
rom PyPi
INFO:apache_beam.runners.portability.stager:Executing command: ['/home/pranjal_saloni61
/env/bin/python', '-m', 'pip', 'download', '--dest', '/tmp/tmp_sxaw0w7', 'apache-beam==
.37.0', '--no-deps', '--only-binary', ':all:', '--python-version', '39', '--implementat
on', 'cp', '--abi', 'cp39', '--platform', 'manylinux1_x86_64']
INFO:apache_beam.runners.portability.stager:Staging binary distribution of the SDK from
PyPI: apache_beam-2.37.0-cp39-cp39-manylinux1_x86_64.whl
WARNING:root:Make sure that locally built Python SDK docker image has Python 3.9 interp
eter.
INFO:root:Default Python SDK image for environment is apache/beam_python3.9_sdk:2.37.0
```

```
prefer: 1
Beat: 1
pight: 1
beaten: 2
summit: 1
grossly: 1
striving: 1
Fairest: 1
meats: 1
glove: 2
notice: 2
encounter: 1
bold: 4
Messenger: 10
knaves: 3
passion: 4
zwaggered: 1
meeting: 2
garb: 1
headlong: 1
cage: 1
needless: 1
patron: 2
spaniel: 1
FRANCE: 10
condemn'd: 1
corky: 1
dissuaded: 1
smile: 2
buzz: 1
Wherefore: 5
egg: 4
despised: 2
football: 1
gracious: 1
```

```
pranjal_saloni612@cloudshell:~ (projectmill)$ find ~/env -name 'wordcount.py'
/home/pranjal_saloni612/env/lib/python3.9/site-packages/apache_beam/examples/dataframe/wordcount.py
/home/pranjal_saloni612/env/lib/python3.9/site-packages/apache_beam/examples/wordcount.py
pranjal_saloni612@cloudshell:~ (projectmill)$ ls
env  index.html  outputs-00000-of-00001  README-cloudshell.txt  SOFE4630U-tut3
pranjal_saloni612@cloudshell:~ (projectmill)$ cp /home/pranjal_saloni612/env/lib/python3pranjal_saloni612@cloudshell:~ (projectmill)$ cp /home/pranjal_saloni612/env/lib/python3.9/site-packages
apache_beam/examples/wordcount.py
cp: missing destination file operand after '/home/pranjal_saloni612/env/lib/python3.9/site-packages/apache_beam/examples/wordcount.py'
Try 'cp --help' for more information.
pranjal_saloni612@cloudshell:~ (projectmill)$ cp /home/pranjal_saloni612/env/lib/python3.9/site-packages/apache_beam/examples/wordcount.py ~/wordcount.py
pranjal_saloni612@cloudshell:~ (projectmill)$ ls
env         outputs-00000-of-00001  SOFE4630U-tut3
index.html  README-cloudshell.txt   wordcount.py
pranjal_saloni612@cloudshell:~ (projectmill)$ source ~/env/bin/activate
(env) pranjal_saloni612@cloudshell:~ (projectmill)$ cp ~/wordcount2.py --output outputs
cp: unrecognized option '--output'
Try 'cp --help' for more information.
(env) pranjal_saloni612@cloudshell:~ (projectmill)$
```

```
l-WriteBundles_50))+(ref_AppliedPTransform_Write2-Write-WriteImpl-Pair_51))+(Write2/Write/WriteImpl/GroupByKey/Write)
INFO:apache_beam.runners.portability.fn_api_runner.fn_runner:Running ((Write2/Write/WriteImpl/GroupByKey/Read)+(ref_AppliedPTransform_Write2-Write-WriteImp
l-Extract_53))+(ref_PCollection_PCollection_35/Write)
INFO:apache_beam.runners.portability.fn_api_runner.fn_runner:Running ((ref_PCollection_PCollection_29/Read)+(ref_AppliedPTransform_Write2-Write-WriteImpl-P
reFinalize_54))+(ref_PCollection_PCollection_36/Write)
INFO:apache_beam.runners.portability.fn_api_runner.fn_runner:Running (ref_PCollection_PCollection_29/Read)+(ref_AppliedPTransform_Write2-Write-WriteImpl-Fi
nalizeWrite_55)
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.
INFO:apache_beam.runners.portability.fn_api_runner.fn_runner:Running (((((ref_AppliedPTransform_Write-Write-WriteImpl-DoOnce-Impulse_28)+(ref_AppliedPTrans
form_Write-Write-WriteImpl-DoOnce-FlatMap-lambda-at-core-py-3228-_29))+(ref_AppliedPTransform_Write-Write-WriteImpl-DoOnce-Map-decode-_31))+(ref_AppliedPTr
ansform_Write-Write-WriteImpl-InitializeWrite_32))+(ref_PCollection_PCollection_18/Write))+(ref_PCollection_PCollection_19/Write)
INFO:apache_beam.runners.portability.fn_api_runner.fn_runner:Running (((((((GroupAndSum/Group/Read)+(GroupAndSum/Merge))+(GroupAndSum/ExtractOutputs))+(ref
_AppliedPTransform_Format_22))+(ref_AppliedPTransform_Write-Write-WriteImpl-WindowInto-WindowIntoFn-_33))+(ref_AppliedPTransform_Write-Write-WriteImpl-Writ
eBundles_34))+(ref_AppliedPTransform_Write-Write-WriteImpl-Pair_35))+(Write/Write/WriteImpl/GroupByKey/Write)
INFO:apache_beam.runners.portability.fn_api_runner.fn_runner:Running ((Write/Write/WriteImpl/GroupByKey/Read)+(ref_AppliedPTransform_Write-Write-WriteImpl-
Extract_37))+(ref_PCollection_PCollection_24/Write)
INFO:apache_beam.runners.portability.fn_api_runner.fn_runner:Running ((ref_PCollection_PCollection_18/Read)+(ref_AppliedPTransform_Write-Write-WriteImpl-Pr
eFinalize_38))+(ref_PCollection_PCollection_25/Write)
INFO:apache_beam.runners.portability.fn_api_runner.fn_runner:Running (ref_PCollection_PCollection_18/Read)+(ref_AppliedPTransform_Write-Write-WriteImpl-Fin
alizeWrite_39)
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) pranjal_saloni612@cloudshell:~ (projectmil1)$ gcloud config list project
[core]
project = projectmil1

Your active configuration is: [cloudshell-15651]
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ gcloud config list project --format
'value(core.prject)'

(env) pranjal_saloni612@cloudshell:~ (projectmil1)$
```

```
(env) goergedaoud2@cloudshell:~ (nomadic-bedrock-342622)$ gcloud config list project --format 'value(core.project)'
nomadic-bedrock-342622
(env) goergedaoud2@cloudshell:~ (nomadic-bedrock-342622)$ PROJECT=$(gcloud config list project --format 'value(core.project)')
(env) goergedaoud2@cloudshell:~ (nomadic-bedrock-342622)$ echo $PROJECT
nomadic-bedrock-342622
(env) goergedaoud2@cloudshell:~ (nomadic-bedrock-342622)$ BUCKET=gs://$PROJECT-gs
echo $BUCKET
gs://nomadic-bedrock-342622-gs
(env) goergedaoud2@cloudshell:~ (nomadic-bedrock-342622)$ python wordcount2.py \
    --region northamerica-northeast2 \
    --runner DataflowRunner \
    --project $PROJECT \
    --temp_location $BUCKET/tmp/ \
    --output $BUCKET/result/outputs \
    --output2 $BUCKET/result/outputs2 \
    --experiment use_unsupported_python_version
```

```
gs://dataflow-samples/shakespeare/macbeth.txt
gs://dataflow-samples/shakespeare/measureforemeasure.txt
gs://dataflow-samples/shakespeare/merchantofvenice.txt
gs://dataflow-samples/shakespeare/merrywivesofwindsor.txt
gs://dataflow-samples/shakespeare/midsummernightsdream.txt
gs://dataflow-samples/shakespeare/muchadoaboutnothing.txt
gs://dataflow-samples/shakespeare/othello.txt
gs://dataflow-samples/shakespeare/periclesprinceoftyre.txt
gs://dataflow-samples/shakespeare/rapeoflucrece.txt
gs://dataflow-samples/shakespeare/romeoandjuliet.txt
gs://dataflow-samples/shakespeare/sonnets.txt
gs://dataflow-samples/shakespeare/tamingoftheshrew.txt
gs://dataflow-samples/shakespeare/tempest.txt
gs://dataflow-samples/shakespeare/timonofathens.txt
gs://dataflow-samples/shakespeare/titusandronicus.txt
gs://dataflow-samples/shakespeare/troilusandcressida.txt
gs://dataflow-samples/shakespeare/twelfthnight.txt
gs://dataflow-samples/shakespeare/twogentlemenofverona.txt
gs://dataflow-samples/shakespeare/various.txt
gs://dataflow-samples/shakespeare/venusandadonis.txt
gs://dataflow-samples/shakespeare/winterstale.txt
goergedaoud2@cloudshell:~ (nomadic-bedrock-342622)$ gsutil ls gs://nomadic-bedrock-342622-gs/result/outputs-00000-of-00003
gs://nomadic-bedrock-342622-gs/result/outputs-00000-of-00003
goergedaoud2@cloudshell:~ (nomadic-bedrock-342622)$ gsutil ls gs://nomadic-bedrock-342622-gs/result/outputs-00000-of-00003
```

Various Dataflow examples for Batch and stream processing for the mnist dataset for various source and destination types; text file, MySQL database, and Kafka topics.
https://youtu.be/9ZDj9KDGtEs
Screenshots:

```
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ git clone https://github.com/goerge
aoud/SOFE4630U-tut4.git
Cloning into 'SOFE4630U-tut4'...
remote: Enumerating objects: 122, done.
remote: Counting objects: 100% (122/122), done.
remote: Compressing objects: 100% (85/85), done.
remote: Total 122 (delta 58), reused 96 (delta 37), pack-reused 0
Receiving objects: 100% (122/122), 14.85 MiB | 22.16 MiB/s, done.
Resolving deltas: 100% (58/58), done.
```

```
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ ls
env          outputs-00000-of-00001  SOFE4630U-tut3  wordcount.py
index.html   README-cloudshell.txt   SOFE4630U-tut4
(env) pranjal_saloni612@cloudshell:~ (projectmil1)$ cd SOFE4630U-tut4
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4 (projectmil1)$ ls
mnist  wordcount
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4 (projectmil1)$ cd mnist
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmil1)$ ls
data        mysql-app.yaml  predictV1.py  predictV3.py  setup.py
mnist.txt   mysql-pvc.yaml  predictV2.py  predictV4.py
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmil1)$ source ~/env/b
n/activate                                                              pip install te
sorflow-cpu==2.8.0ni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmil1)$ pip install te
Collectingutensorflow-cpu==2.8.0
  Downloading tensorflow_cpu-2.8.0-cp39-cp39-manylinux2010_x86_64.whl (190.6 MB)
     ---------------------------------------- 190.6/190.6 MB 5.0 MB/s eta 0:00:00
Requirement already satisfied: protobuf>=3.9.2 in /home/pranjal_saloni612/env/lib/pytho
3.9/site-packages (from tensorflow-cpu==2.8.0) (3.19.4)
Collecting google-pasta>=0.1.1
  Downloading google_pasta-0.2.0-py3-none-any.whl (57 kB)
     ---------------------------------------- 57.5/57.5 KB 9.7 MB/s eta 0:00:00
Collecting astunparse>=1.6.0
  Downloading astunparse-1.6.3-py2.py3-none-any.whl (12 kB)
```

```
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmil1)$ df
Filesystem                         1K-blocks     Used Available Use% Mounted on
overlay                             62742040 48102400  14623256  77% /
tmpfs                                  65536        0     65536   0% /dev
tmpfs                                8196724        0   8196724   0% /sys/fs/cgroup
/dev/sda1                           62742040 48102400  14623256  77% /root
/dev/disk/by-id/google-home-part1    5028480  1552972   3197032  33% /home
/dev/root                            2003760  1063960    939800  54% /usr/lib/modules
shm                                    65536        0     65536   0% /dev/shm
tmpfs                                3278692      904   3277788   1% /google/host/var/ru
```

```
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmil1)$ ls -la
total 56
drwxr-xr-x 3 pranjal_saloni612 pranjal_saloni612 4096 Mar 29 04:27 .
drwxr-xr-x 5 pranjal_saloni612 pranjal_saloni612 4096 Mar 29 04:27 ..
drwxr-xr-x 3 pranjal_saloni612 pranjal_saloni612 4096 Mar 29 04:27 data
-rw-r--r-- 1 pranjal_saloni612 pranjal_saloni612 3940 Mar 29 04:27 mnist.txt
-rw-r--r-- 1 pranjal_saloni612 pranjal_saloni612 1053 Mar 29 04:27 mysql-app.yaml
-rw-r--r-- 1 pranjal_saloni612 pranjal_saloni612  166 Mar 29 04:27 mysql-pvc.yaml
-rw-r--r-- 1 pranjal_saloni612 pranjal_saloni612 3575 Mar 29 04:27 predictV1.py
-rw-r--r-- 1 pranjal_saloni612 pranjal_saloni612 4706 Mar 29 04:27 predictV2.py
```

```
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmil1)$ python ./predi
python ./predictV1.py \
>  --staging_location ./staging \
    --temp_location ./temp \
    --model ./data \
    --source text \
        --setup_file ./setup.py \
    --input ./data/images.txt \
   --staging_location ./staging \
    --temp_location ./temp \
    --model ./data \
    --source text \
        --setup_file ./setup.py \
    --input ./data/images.txt \
    --output ./predict
INFO:root:Missing pipeline option (runner). Executing pipeline using the default runner
irectRunner.
WARNING:root:Make sure that locally built Python SDK docker image has Python 3.9 interp
eter.
INFO:root:Default Python SDK image for environment is apache/beam_python3.9_sdk:2.37.0
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <f
nction annotate_downstream_side_inputs at 0x7f2b4e3311f0> ====================
INFO:apache_beam.runners.portability.fn_api_runner.translations:==================== <f
```

```
INFO:apache_beam.io.filebasedsink:Renamed 1 shards in 0.00 seconds.
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmil1)$ ls
data        mysql-app.yaml  predict-00000-of-00001  predictV2.py  predictV4.py
mnist.txt   mysql-pvc.yaml  predictV1.py            predictV3.py  setup.py
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmil1)$
```

```
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill)$ PROJECT=$(gclo
d config list project --format "value(core.project)")
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill)$ echo PROJECT
PROJECT
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill)$ echo $PROJECT
projectmill
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill)$ PROJECT=$(gclo
d config list project --format "value(core.project)")
echo $PROJECT
BUCKET=gs://$PROJECT-gs
echo $BUCKET
projectmill
gs://projectmill-gs
```

```
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill)$ gsutil cp data
export* $BUCKET/model/
Copying file://data/export.data-00000-of-00001 [Content-Type=application/octet-stream].
.
Copying file://data/export.index [Content-Type=application/octet-stream]...
Copying file://data/export.meta [Content-Type=application/octet-stream]...
- [3 files][ 12.5 MiB/ 12.5 MiB]
Operation completed over 3 objects/12.5 MiB.
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill)$ gsutil ls $PRO
ECT
CommandException: "ls" command does not support "file://" URLs. Did you mean to use a g
:// URL?
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill)$
```

```
(env) goergedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ gsutil ls  $BUCKET
ServiceException: 401 Anonymous caller does not have storage.objects.list access to the Google Cloud Storage bucket.
(env) goergedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ gsutil ls  gs://nomadic-bedrock-342622-gs
ServiceException: 401 Anonymous caller does not have storage.objects.list access to the Google Cloud Storage bucket.
(env) goergedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ gsutil ls  gs://nomadic-bedrock-342622-gs
ServiceException: 401 Anonymous caller does not have storage.objects.list access to the Google Cloud Storage bucket.
(env) goergedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ gsutil ls  gs://nomadic-bedrock-342622-gs
gs://nomadic-bedrock-342622-gs/result/
gs://nomadic-bedrock-342622-gs/tmp/
(env) goergedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ gsutil cp ./data/export* $BUCKET/model/
Copying file://./data/export.data-00000-of-00001 [Content-Type=application/octet-stream]...
Copying file://./data/export.index [Content-Type=application/octet-stream]...
Copying file://./data/export.meta [Content-Type=application/octet-stream]...
- [3 files][ 12.5 MiB/ 12.5 MiB]
Operation completed over 3 objects/12.5 MiB.
(env) goergedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ gsutil cp ./data/images.txt $BUCKET/input/
Copying file://./data/images.txt [Content-Type=text/plain]...
- [1 files][ 45.2 MiB/ 45.2 MiB]
Operation completed over 1 objects/45.2 MiB.
```

```
nomadic-bedrock-342622
gs://nomadic-bedrock-342622-gs
(env) goergedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ gsutil ls  $BUCKET
ServiceException: 401 Anonymous caller does not have storage.objects.list access to the Google Cloud Storage bucket.
(env) goergedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ gsutil ls  gs://nomadic-bedrock-342622-gs
ServiceException: 401 Anonymous caller does not have storage.objects.list access to the Google Cloud Storage bucket.
(env) goergedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ gsutil ls  gs://nomadic-bedrock-342622-gs
ServiceException: 401 Anonymous caller does not have storage.objects.list access to the Google Cloud Storage bucket.
(env) goergedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ gsutil ls  gs://nomadic-bedrock-342622-gs
gs://nomadic-bedrock-342622-gs/result/
gs://nomadic-bedrock-342622-gs/tmp/
(env) goergedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ gsutil cp ./data/export* $BUCKET/model/
Copying file://./data/export.data-00000-of-00001 [Content-Type=application/octet-stream]...
Copying file://./data/export.index [Content-Type=application/octet-stream]...
Copying file://./data/export.meta [Content-Type=application/octet-stream]...
- [3 files][ 12.5 MiB/ 12.5 MiB]
Operation completed over 3 objects/12.5 MiB.
(env) goergedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ gsutil cp ./data/images.txt $BUCKET/input/
Copying file://./data/images.txt [Content-Type=text/plain]...
- [1 files][ 45.2 MiB/ 45.2 MiB]
Operation completed over 1 objects/45.2 MiB.
(env) goergedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ python ./predictV1.py \
    --runner DataflowRunner \
    --project $PROJECT \
    --staging_location $BUCKET/staging \
    --temp_location $BUCKET/temp \
    --model $BUCKET/model \
    --source text \
        --setup_file ./setup.py \
    --input $BUCKET/input/images.txt \
    --output $BUCKET/output/predict \
```

```
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill1)$ source ~/env/bin/activate
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill1)$ pip install beam-nuggets
Collecting beam-nuggets
  Downloading beam_nuggets-0.18.1-py3-none-any.whl (25 kB)
Collecting SQLAlchemy<2.0.0,>=1.4.0
  Downloading SQLAlchemy-1.4.32-cp39-cp39-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.6 MB)
     ---------------------------------------- 1.6/1.6 MB 23.1 MB/s eta 0:00:00
Collecting sqlalchemy-utils<1.0.0,>=0.37.0
  Downloading SQLAlchemy_Utils-0.38.2-py3-none-any.whl (100 kB)
     ---------------------------------------- 100.2/100.2 KB 13.5 MB/s eta 0:00:00
Collecting PyMySQL<2.0.0,>=0.9.3
  Downloading PyMySQL-1.0.2-py3-none-any.whl (43 kB)
     ---------------------------------------- 43.8/43.8 KB 6.7 MB/s eta 0:00:00
Collecting kafka-python>=2.0.1
  Downloading kafka_python-2.0.2-py2.py3-none-any.whl (246 kB)
     ---------------------------------------- 246.5/246.5 KB 28.4 MB/s eta 0:00:00
Requirement already satisfied: apache-beam<3.0.0,>=2.8.0 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from beam-nuggets) (2.37.0)
Collecting pg8000<2.0.0,>=1.16.6
  Downloading pg8000-1.24.1-py3-none-any.whl (33 kB)
Requirement already satisfied: dill<0.3.2,>=0.3.1.1 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from apache-beam<3.0.0,>=2.8.0->beam-nuggets) (0.3.1.1)
Requirement already satisfied: pydot<2,>=1.2.0 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from apache-beam<3.0.0,>=2.8.0->beam-nuggets) (1.4.2)
Requirement already satisfied: pymongo<4.0.0,>=3.8.0 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from apache-beam<3.0.0,>=2.8.0->beam-nuggets) (3.12.3)
Requirement already satisfied: python-dateutil<3,>=2.8.0 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from apache-beam<3.0.0,>=2.8.0->beam-nuggets) (2.8.2)
Requirement already satisfied: fastavro<2,>=0.23.6 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from apache-beam<3.0.0,>=2.8.0->beam-nuggets) (1.4.10)
Requirement already satisfied: crcmod<2.0,>=1.7 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from apache-beam<3.0.0,>=2.8.0->beam-nuggets) (1.7)
Requirement already satisfied: httplib2<0.20.0,>=0.8 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from apache-beam<3.0.0,>=2.8.0->beam-nuggets) (0.19.1)
Requirement already satisfied: proto-plus<2,>=1.7.1 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from apache-beam<3.0.0,>=2.8.0->beam-nuggets) (1.20.3)
Requirement already satisfied: protobuf<4,>=3.12.2 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from apache-beam<3.0.0,>=2.8.0->beam-nuggets) (3.19.4)
Requirement already satisfied: orjson<4.0 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from apache-beam<3.0.0,>=2.8.0->beam-nuggets) (3.6.7)
Requirement already satisfied: hdfs<3.0.0,>=2.1.0 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from apache-beam<3.0.0,>=2.8.0->beam-nuggets) (2.7.0)
Requirement already satisfied: numpy<1.22.0,>=1.14.3 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from apache-beam<3.0.0,>=2.8.0->beam-nuggets) (1.21.5)
```

```
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /home/pranjal_saloni612/env/lib/python3.9/site-packages (from requests<3.0.0,>=2.24.0->apache-beam<3.0.0,>=2.8.0->beam-nuggets) (1.26.9)
Collecting asn1crypto>=1.4.0
  Downloading asn1crypto-1.5.1-py2.py3-none-any.whl (105 kB)
     ---------------------------------------- 105.0/105.0 KB 15.8 MB/s eta 0:00:00
Installing collected packages: kafka-python, asn1crypto, scramp, PyMySQL, greenlet, SQLAlchemy, pg8000, sqlalchemy-utils, beam-nuggets
Successfully installed PyMySQL-1.0.2 SQLAlchemy-1.4.32 asn1crypto-1.5.1 beam-nuggets-0.18.1 greenlet-1.1.2 kafka-python-2.0.2 pg8000-1.24.1 scramp-1.4.1 sqlalchemy-utils-0.38.2
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill1)$ cd SOFE4630U-tut4
-bash: cd: SOFE4630U-tut4: No such file or directory
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill1)$ cd mnist
-bash: cd: mnist: No such file or directory
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill1)$ ls
data       mysql-app.yaml  predict-00000-of-00001  predictV2.py  predictV4.py
mnist.txt  mysql-pvc.yaml  predictV1.py            predictV3.py  setup.py
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill1)$ gcloud config set compute/zone northamerica-northeast2-a                          gcloud config set compute/zone 
orthamerica-northeast2-a
Updated property [compute/zone].
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill1)$ gcloud container clusters create gk-cluster --num-nodes=3                          gcloud container clusters creat
 gk-cluster --num-nodes=3
Default change: VPC-native is the default mode during cluster creation for versions greater than 1.21.0-gke.1500. To create advanced routes based clusters, please pass the `--no-enable-ip-alia
` flag
Note: Your Pod address range (`--cluster-ipv4-cidr`) can accommodate at most 1008 node(s).
Creating cluster gk-cluster in northamerica-northeast2-a... Cluster is being configured.Creatinggcluster gk-cluster in northamerica-northeast2-a... Cluster is being configured...working..
.
Creating cluster gk-cluster in northamerica-northeast2-a... Cluster is being configured.Creatinggcluster gk-cluster in northamerica-northeast2-a... Cluster is being configured.Creatinggcluster
gk-cluster in northamerica-northeast2-a... Cluster is being configured.Creatinggcluster gk-cluster in northamerica-northeast2-a... Cluster is being configured...working..
.
Creating cluster gk-cluster in northamerica-northeast2-a... Cluster is being configured.Creatinggcluster gk-cluster in northamerica-northeast2-a... Cluster is being configured.Creatinggcluster
gk-cluster in northamerica-northeast2-a... Cluster is being configured.Creatinggcluster gk-cluster in northamerica-northeast2-a... Cluster is being configured...working..
.
Creating cluster gk-cluster in northamerica-northeast2-a... Cluster is being configured.Creatinggcluster gk-cluster in northamerica-northeast2-a... Cluster is being configured.Creatinggcluster
gk-cluster in northamerica-northeast2-a... Cluster is being configured.Creatinggcluster gk-cluster in northamerica-northeast2-a... Cluster is being configured...working..
.
Creating cluster gk-cluster in northamerica-northeast2-a... Cluster is being configured.       g                                                  Creating cluster
mnist.txt  mysql-pvc.yaml  predictV1.py            predictV3.py  setup.py
```

```
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill)$ gcloud config set compute/zone northamerica-northeast2-a
Updated property [compute/zone].
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill)$ gcloud container clusters create gk-cluster --num-nodes=3
Default change: VPC-native is the default mode during cluster creation for versions greater than 1.21.0-gke.1500. To create advanced routes based clusters, please pass the `--no-enable-ip-alia
s` flag
Note: Your Pod address range (`--cluster-ipv4-cidr`) can accommodate at most 1008 node(s).
Creating cluster gk-cluster in northamerica-northeast2-a... Cluster is being configured...working..
.
Creating cluster gk-cluster in northamerica-northeast2-a... Cluster is being configured...working..
.
Creating cluster gk-cluster in northamerica-northeast2-a... Cluster is being configured...working..
.
Creating cluster gk-cluster in northamerica-northeast2-a... Cluster is being configured...working..
.
Creating cluster gk-cluster in northamerica-northeast2-a... Cluster is being configured...working..
.
Creating cluster gk-cluster in northamerica-northeast2-a... Cluster is being configured...working..
.
Creating cluster gk-cluster in northamerica-northeast2-a... Cluster is being health-checked...worki
ng...
Creating cluster gk-cluster in northamerica-northeast2-a... Cluster is being health-checked (master
 is healthy)...done.
Created [https://container.googleapis.com/v1/projects/projectmill/zones/northamerica-northeast2-a/clusters/gk-cluster].
To inspect the contents of your cluster, go to: https://console.cloud.google.com/kubernetes/workload_/gcloud/northamerica-northeast2-a/gk-cluster?project=projectmill
kubeconfig entry generated for gk-cluster.
NAME: gk-cluster
LOCATION: northamerica-northeast2-a
MASTER_VERSION: 1.21.9-gke.1002
MASTER_IP: 34.130.252.229
MACHINE_TYPE: e2-medium
NODE_VERSION: 1.21.9-gke.1002
NUM_NODES: 3
STATUS: RUNNING
(env) pranjal_saloni612@cloudshell:~/SOFE4630U-tut4/mnist (projectmill)$
```

```
NODE_VERSION: 1.21.6-gke.1500
NUM_NODES: 3
STATUS: RUNNING
(env) georgedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ gcloud container clusters get-credentials gk-clust
er
Fetching cluster endpoint and auth data.
kubeconfig entry generated for gk-cluster.
(env) georgedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ kubectl apply -f mysql-pvc.yaml
persistentvolumeclaim/mysql-volumeclaim created
(env) georgedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ kubectl apply -f mysql-app.yaml
service/mysql created
deployment.apps/mysql created
(env) georgedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ kubectl get services
NAME         TYPE           CLUSTER-IP     EXTERNAL-IP   PORT(S)          AGE
kubernetes   ClusterIP      10.112.0.1     <none>        443/TCP          2m38s
mysql        LoadBalancer   10.112.10.46   <pending>     3306:31377/TCP   10s
(env) georgedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ kubectl get services
NAME         TYPE           CLUSTER-IP     EXTERNAL-IP   PORT(S)          AGE
kubernetes   ClusterIP      10.112.0.1     <none>        443/TCP          2m43s
mysql        LoadBalancer   10.112.10.46   <pending>     3306:31377/TCP   15s
(env) georgedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ kubectl get services
NAME         TYPE           CLUSTER-IP     EXTERNAL-IP   PORT(S)          AGE
kubernetes   ClusterIP      10.112.0.1     <none>        443/TCP          2m48s
mysql        LoadBalancer   10.112.10.46   <pending>     3306:31377/TCP   20s
```

```
(env) georgedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ kubectl get services
NAME         TYPE           CLUSTER-IP     EXTERNAL-IP   PORT(S)          AGE
kubernetes   ClusterIP      10.112.0.1     <none>        443/TCP          2m56s
mysql        LoadBalancer   10.112.10.46   <pending>     3306:31377/TCP   28s
(env) georgedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ kubectl get services
NAME         TYPE           CLUSTER-IP     EXTERNAL-IP   PORT(S)          AGE
kubernetes   ClusterIP      10.112.0.1     <none>        443/TCP          3m6s
mysql        LoadBalancer   10.112.10.46   <pending>     3306:31377/TCP   38s
(env) georgedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ kubectl get services
NAME         TYPE           CLUSTER-IP     EXTERNAL-IP      PORT(S)          AGE
kubernetes   ClusterIP      10.112.0.1     <none>           443/TCP          3m13s
mysql        LoadBalancer   10.112.10.46   34.130.188.137   3306:31377/TCP   45s
(env) georgedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ kubectl get service mysql
NAME    TYPE           CLUSTER-IP     EXTERNAL-IP      PORT(S)          AGE
mysql   LoadBalancer   10.112.10.46   34.130.188.137   3306:31377/TCP   67s
(env) georgedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ kubectl get service mysql
NAME    TYPE           CLUSTER-IP     EXTERNAL-IP      PORT(S)          AGE
mysql   LoadBalancer   10.112.10.46   34.130.188.137   3306:31377/TCP   103s
(env) georgedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ kubectl get services mysql -o jsonpath='{.status.l
oadBalancer.ingress[0].ip}'
(env) georgedaoud2@cloudshell:~/SOFE4630U-tut4/mnist (nomadic-bedrock-342622)$ MYSQLIP=$(kubectl get services mysql -o jsonpath='
{.status.loadBalancer.ingress[0].ip}')ress[0].ip}')
echo $MYSQLIP
34.130.188.137
```

Q&A:

Google Cloud has another processing service called DataProc. Name another processing service that is usually used in the cloud environment (not necessarily GCP). Compare between it and both Dataflow and DataProc. Your comparison may include but is not limited to the major differences, advantages, disadvantages, and limitations.

Another processing service used in cloud environments is Google Cloud Dataprep.

	Dataflow	Dataproc	Dataprep
Main use	Exeutes Apache Beam pipelines within the GCP environment.	Highly scalable service that runs Apache Spark, Apache Flink, and various other open source tools and frameworks.	Visually explores, cleans and thraforms structured and unstructured data for analytics and reporting.
Major differences	Follows batch and stream processing of data. Can only observe and stream processing data.	Dataproc is a Google Cloud product with Data Science/ML service for Spark and Hadoop.	Dataprep is UI-driven, scales on-demand and fully automated.
When to use?	Dataflow provides a clear separation between processing logic and the underlying execution engine.	Dataproc should be used if the processing has any dependencies to tools in the Hadoop ecosystem.	Dataprep is used only as a medium of processing data for further use, such as in BigQuery.
Advantages	Helps understand the functioning and the limits of a system.	Utilizes Spark and Hadoop; help create and manage clusters quickly.	Allows quick exploration of new datasets.
Disadvantages/Limitations	Limited job computation.	No choice of selecting a specific version of Hadoop/hive/spark stack.	Limited access to APIs.

Suggest a practical application using both stream and batch processing that can be applied to a given dataset. It's expected to use the dataset uploaded in the third milestone but you can use any other dataset. If you decide to use another dataset, It should maintain both variety and huge volume. Your report should include but not limited to:

- The application.
- Its impact.
- The used dataset (size, schema/structure).

- A graph showing the proposed pipeline(s).
- List of other tools (AI, clustering,…) needed to implement that application.

An example of batch processing is monthly bill generations. Batch processing is a technique for automating and processing multiple transactions as a single group. While batch processing can be carried out at any time, it's particularly suited to end-of-cycle processing, such as for processing a bank's reports at the end of a day or generating monthly or biweekly payrolls. Repeated jobs are done fast in batch systems without user interaction. ATM transactions use stream processing as data needs to be collected and processed in real time. BigQuery is a tool that can be implemented for these applications. BigQuery is a fully-managed, serverless data warehouse that enables scalable analysis over petabytes of data.