Preprocessing using Dataflow

This notebook illustrates:

1. Creating datasets for Machine Learning using Dataflow

   While Pandas is fine for experimenting, for operationalization of your workflow, it is better to do preprocessing in Apache Beam. This will also help if you need to preprocess data in flight, since Apache Beam also allows for streaming.

   Each learning objective will correspond to a **#TODO** in this student lab notebook -- try to complete this notebook first and then review the solution notebook.

In [1]:
```
!sudo chown -R jupyter:jupyter /home/jupyter/training-data-analyst
```

In [2]:
```
!pip install --user google-cloud-bigquery==1.25.0
```

```
Requirement already satisfied: google-cloud-bigquery==1.25.0 in /home/jupyter/.l
ocal/lib/python3.7/site-packages (1.25.0)
Requirement already satisfied: google-api-core<2.0dev,>=1.15.0 in /opt/conda/li
b/python3.7/site-packages (from google-cloud-bigquery==1.25.0) (1.31.1)
Requirement already satisfied: google-cloud-core<2.0dev,>=1.1.0 in /opt/conda/li
b/python3.7/site-packages (from google-cloud-bigquery==1.25.0) (1.7.2)
Requirement already satisfied: protobuf>=3.6.0 in /opt/conda/lib/python3.7/site-
packages (from google-cloud-bigquery==1.25.0) (3.16.0)
Requirement already satisfied: google-auth<2.0dev,>=1.9.0 in /opt/conda/lib/pyth
on3.7/site-packages (from google-cloud-bigquery==1.25.0) (1.34.0)
Requirement already satisfied: google-resumable-media<0.6dev,>=0.5.0 in /home/ju
pyter/.local/lib/python3.7/site-packages (from google-cloud-bigquery==1.25.0)
(0.5.1)
Requirement already satisfied: six<2.0.0dev,>=1.13.0 in /opt/conda/lib/python3.
7/site-packages (from google-cloud-bigquery==1.25.0) (1.16.0)
Requirement already satisfied: setuptools>=40.3.0 in /opt/conda/lib/python3.7/si
te-packages (from google-api-core<2.0dev,>=1.15.0->google-cloud-bigquery==1.25.
0) (49.6.0.post20210108)
Requirement already satisfied: pytz in /opt/conda/lib/python3.7/site-packages (f
rom google-api-core<2.0dev,>=1.15.0->google-cloud-bigquery==1.25.0) (2021.1)
Requirement already satisfied: requests<3.0.0dev,>=2.18.0 in /opt/conda/lib/pyth
on3.7/site-packages (from google-api-core<2.0dev,>=1.15.0->google-cloud-bigquery
==1.25.0) (2.25.1)
Requirement already satisfied: googleapis-common-protos<2.0dev,>=1.6.0 in /opt/c
onda/lib/python3.7/site-packages (from google-api-core<2.0dev,>=1.15.0->google-c
loud-bigquery==1.25.0) (1.53.0)
Requirement already satisfied: packaging>=14.3 in /opt/conda/lib/python3.7/site-
packages (from google-api-core<2.0dev,>=1.15.0->google-cloud-bigquery==1.25.0)
(21.0)
Requirement already satisfied: cachetools<5.0,>=2.0.0 in /opt/conda/lib/python3.
7/site-packages (from google-auth<2.0dev,>=1.9.0->google-cloud-bigquery==1.25.0)
(4.2.2)
Requirement already satisfied: rsa<5,>=3.1.4 in /opt/conda/lib/python3.7/site-pa
ckages (from google-auth<2.0dev,>=1.9.0->google-cloud-bigquery==1.25.0) (4.7.2)
Requirement already satisfied: pyasn1-modules>=0.2.1 in /opt/conda/lib/python3.
7/site-packages (from google-auth<2.0dev,>=1.9.0->google-cloud-bigquery==1.25.0)
```

```
(0.2.7)
Requirement already satisfied: pyparsing>=2.0.2 in /opt/conda/lib/python3.7/site
-packages (from packaging>=14.3->google-api-core<2.0dev,>=1.15.0->google-cloud-b
igquery==1.25.0) (2.4.7)
Requirement already satisfied: pyasn1<0.5.0,>=0.4.6 in /opt/conda/lib/python3.7/
site-packages (from pyasn1-modules>=0.2.1->google-auth<2.0dev,>=1.9.0->google-cl
oud-bigquery==1.25.0) (0.4.8)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python3.7/site-pac
kages (from requests<3.0.0dev,>=2.18.0->google-api-core<2.0dev,>=1.15.0->google-
cloud-bigquery==1.25.0) (2.10)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /opt/conda/lib/python3.
7/site-packages (from requests<3.0.0dev,>=2.18.0->google-api-core<2.0dev,>=1.15.
0->google-cloud-bigquery==1.25.0) (1.26.6)
Requirement already satisfied: chardet<5,>=3.0.2 in /opt/conda/lib/python3.7/sit
e-packages (from requests<3.0.0dev,>=2.18.0->google-api-core<2.0dev,>=1.15.0->go
ogle-cloud-bigquery==1.25.0) (4.0.0)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.7/si
te-packages (from requests<3.0.0dev,>=2.18.0->google-api-core<2.0dev,>=1.15.0->g
oogle-cloud-bigquery==1.25.0) (2021.5.30)
```

Kindly ignore the deprecation warnings and incompatibility errors related to google-cloud-storage.

In [3]:
```
!pip install --user apache-beam[interactive]==2.24.0
```

```
Requirement already satisfied: apache-beam[interactive]==2.24.0 in /home/jupyte
r/.local/lib/python3.7/site-packages (2.24.0)
Requirement already satisfied: requests<3.0.0,>=2.24.0 in /opt/conda/lib/python
3.7/site-packages (from apache-beam[interactive]==2.24.0) (2.25.1)
Requirement already satisfied: avro-python3!=1.9.2,<1.10.0,>=1.8.1 in /opt/cond
a/lib/python3.7/site-packages (from apache-beam[interactive]==2.24.0) (1.9.2.1)
Requirement already satisfied: oauth2client<4,>=2.0.1 in /home/jupyter/.local/li
b/python3.7/site-packages (from apache-beam[interactive]==2.24.0) (3.0.0)
Requirement already satisfied: dill<0.3.2,>=0.3.1.1 in /home/jupyter/.local/lib/
python3.7/site-packages (from apache-beam[interactive]==2.24.0) (0.3.1.1)
Requirement already satisfied: numpy<2,>=1.14.3 in /opt/conda/lib/python3.7/site
-packages (from apache-beam[interactive]==2.24.0) (1.19.5)
Requirement already satisfied: crcmod<2.0,>=1.7 in /opt/conda/lib/python3.7/site
-packages (from apache-beam[interactive]==2.24.0) (1.7)
Requirement already satisfied: python-dateutil<3,>=2.8.0 in /opt/conda/lib/pytho
n3.7/site-packages (from apache-beam[interactive]==2.24.0) (2.8.2)
Requirement already satisfied: future<1.0.0,>=0.18.2 in /opt/conda/lib/python3.
7/site-packages (from apache-beam[interactive]==2.24.0) (0.18.2)
Requirement already satisfied: pyarrow<0.18.0,>=0.15.1 in /home/jupyter/.local/l
ib/python3.7/site-packages (from apache-beam[interactive]==2.24.0) (0.17.1)
Requirement already satisfied: pydot<2,>=1.2.0 in /opt/conda/lib/python3.7/site-
packages (from apache-beam[interactive]==2.24.0) (1.4.2)
Requirement already satisfied: protobuf<4,>=3.12.2 in /opt/conda/lib/python3.7/s
ite-packages (from apache-beam[interactive]==2.24.0) (3.16.0)
Requirement already satisfied: pytz>=2018.3 in /opt/conda/lib/python3.7/site-pac
kages (from apache-beam[interactive]==2.24.0) (2021.1)
Requirement already satisfied: pymongo<4.0.0,>=3.8.0 in /opt/conda/lib/python3.
7/site-packages (from apache-beam[interactive]==2.24.0) (3.12.0)
Requirement already satisfied: fastavro<0.24,>=0.21.4 in /home/jupyter/.local/li
b/python3.7/site-packages (from apache-beam[interactive]==2.24.0) (0.23.6)
Requirement already satisfied: grpcio<2,>=1.29.0 in /opt/conda/lib/python3.7/sit
e-packages (from apache-beam[interactive]==2.24.0) (1.38.1)
Requirement already satisfied: hdfs<3.0.0,>=2.1.0 in /opt/conda/lib/python3.7/si
te-packages (from apache-beam[interactive]==2.24.0) (2.6.0)
Requirement already satisfied: httplib2<0.18.0,>=0.8 in /home/jupyter/.local/li
```

```
b/python3.7/site-packages (from apache-beam[interactive]==2.24.0) (0.17.4)
Requirement already satisfied: mock<3.0.0,>=1.0.1 in /home/jupyter/.local/lib/py
thon3.7/site-packages (from apache-beam[interactive]==2.24.0) (2.0.0)
Requirement already satisfied: typing-extensions<3.8.0,>=3.7.0 in /home/jupyte
r/.local/lib/python3.7/site-packages (from apache-beam[interactive]==2.24.0) (3.
7.4.3)
Requirement already satisfied: ipykernel<6,>=5.2.0 in /opt/conda/lib/python3.7/s
ite-packages (from apache-beam[interactive]==2.24.0) (5.5.5)
Requirement already satisfied: ipython<8,>=5.8.0 in /opt/conda/lib/python3.7/sit
e-packages (from apache-beam[interactive]==2.24.0) (7.25.0)
Requirement already satisfied: timeloop<2,>=1.0.2 in /home/jupyter/.local/lib/py
thon3.7/site-packages (from apache-beam[interactive]==2.24.0) (1.0.2)
Requirement already satisfied: facets-overview<2,>=1.0.0 in /home/jupyter/.loca
l/lib/python3.7/site-packages (from apache-beam[interactive]==2.24.0) (1.0.0)
Requirement already satisfied: pandas>=0.22.0 in /opt/conda/lib/python3.7/site-p
ackages (from facets-overview<2,>=1.0.0->apache-beam[interactive]==2.24.0) (1.3.
1)
Requirement already satisfied: six>=1.5.2 in /opt/conda/lib/python3.7/site-packa
ges (from grpcio<2,>=1.29.0->apache-beam[interactive]==2.24.0) (1.16.0)
Requirement already satisfied: docopt in /opt/conda/lib/python3.7/site-packages
(from hdfs<3.0.0,>=2.1.0->apache-beam[interactive]==2.24.0) (0.6.2)
Requirement already satisfied: tornado>=4.2 in /opt/conda/lib/python3.7/site-pac
kages (from ipykernel<6,>=5.2.0->apache-beam[interactive]==2.24.0) (6.1)
Requirement already satisfied: traitlets>=4.1.0 in /opt/conda/lib/python3.7/site
-packages (from ipykernel<6,>=5.2.0->apache-beam[interactive]==2.24.0) (5.0.5)
Requirement already satisfied: jupyter-client in /opt/conda/lib/python3.7/site-p
ackages (from ipykernel<6,>=5.2.0->apache-beam[interactive]==2.24.0) (6.1.12)
Requirement already satisfied: jedi>=0.16 in /opt/conda/lib/python3.7/site-packa
ges (from ipython<8,>=5.8.0->apache-beam[interactive]==2.24.0) (0.18.0)
Requirement already satisfied: decorator in /opt/conda/lib/python3.7/site-packag
es (from ipython<8,>=5.8.0->apache-beam[interactive]==2.24.0) (5.0.9)
Requirement already satisfied: pexpect>4.3 in /opt/conda/lib/python3.7/site-pack
ages (from ipython<8,>=5.8.0->apache-beam[interactive]==2.24.0) (4.8.0)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in /
opt/conda/lib/python3.7/site-packages (from ipython<8,>=5.8.0->apache-beam[inter
active]==2.24.0) (3.0.19)
Requirement already satisfied: backcall in /opt/conda/lib/python3.7/site-package
s (from ipython<8,>=5.8.0->apache-beam[interactive]==2.24.0) (0.2.0)
Requirement already satisfied: pickleshare in /opt/conda/lib/python3.7/site-pack
ages (from ipython<8,>=5.8.0->apache-beam[interactive]==2.24.0) (0.7.5)
Requirement already satisfied: matplotlib-inline in /opt/conda/lib/python3.7/sit
e-packages (from ipython<8,>=5.8.0->apache-beam[interactive]==2.24.0) (0.1.2)
Requirement already satisfied: pygments in /opt/conda/lib/python3.7/site-package
s (from ipython<8,>=5.8.0->apache-beam[interactive]==2.24.0) (2.9.0)
Requirement already satisfied: setuptools>=18.5 in /opt/conda/lib/python3.7/site
-packages (from ipython<8,>=5.8.0->apache-beam[interactive]==2.24.0) (49.6.0.pos
t20210108)
Requirement already satisfied: parso<0.9.0,>=0.8.0 in /opt/conda/lib/python3.7/s
ite-packages (from jedi>=0.16->ipython<8,>=5.8.0->apache-beam[interactive]==2.2
4.0) (0.8.2)
Requirement already satisfied: pbr>=0.11 in /home/jupyter/.local/lib/python3.7/s
ite-packages (from mock<3.0.0,>=1.0.1->apache-beam[interactive]==2.24.0) (5.6.0)
Requirement already satisfied: rsa>=3.1.4 in /opt/conda/lib/python3.7/site-packa
ges (from oauth2client<4,>=2.0.1->apache-beam[interactive]==2.24.0) (4.7.2)
Requirement already satisfied: pyasn1-modules>=0.0.5 in /opt/conda/lib/python3.
7/site-packages (from oauth2client<4,>=2.0.1->apache-beam[interactive]==2.24.0)
(0.2.7)
Requirement already satisfied: pyasn1>=0.1.7 in /opt/conda/lib/python3.7/site-pa
ckages (from oauth2client<4,>=2.0.1->apache-beam[interactive]==2.24.0) (0.4.8)
Requirement already satisfied: ptyprocess>=0.5 in /opt/conda/lib/python3.7/site-
packages (from pexpect>4.3->ipython<8,>=5.8.0->apache-beam[interactive]==2.24.0)
```

```
(0.7.0)
Requirement already satisfied: wcwidth in /opt/conda/lib/python3.7/site-packages
(from prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0->ipython<8,>=5.8.0->apache-be
am[interactive]==2.24.0) (0.2.5)
Requirement already satisfied: pyparsing>=2.1.4 in /opt/conda/lib/python3.7/site
-packages (from pydot<2,>=1.2.0->apache-beam[interactive]==2.24.0) (2.4.7)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python3.7/site-pac
kages (from requests<3.0.0,>=2.24.0->apache-beam[interactive]==2.24.0) (2.10)
Requirement already satisfied: chardet<5,>=3.0.2 in /opt/conda/lib/python3.7/sit
e-packages (from requests<3.0.0,>=2.24.0->apache-beam[interactive]==2.24.0) (4.
0.0)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.7/si
te-packages (from requests<3.0.0,>=2.24.0->apache-beam[interactive]==2.24.0) (20
21.5.30)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /opt/conda/lib/python3.
7/site-packages (from requests<3.0.0,>=2.24.0->apache-beam[interactive]==2.24.0)
(1.26.6)
Requirement already satisfied: ipython-genutils in /opt/conda/lib/python3.7/site
-packages (from traitlets>=4.1.0->ipykernel<6,>=5.2.0->apache-beam[interactive]=
=2.24.0) (0.2.0)
Requirement already satisfied: pyzmq>=13 in /opt/conda/lib/python3.7/site-packag
es (from jupyter-client->ipykernel<6,>=5.2.0->apache-beam[interactive]==2.24.0)
(22.1.0)
Requirement already satisfied: jupyter-core>=4.6.0 in /opt/conda/lib/python3.7/s
ite-packages (from jupyter-client->ipykernel<6,>=5.2.0->apache-beam[interactive]
==2.24.0) (4.7.1)
```

NOTE: In the output of the above cell you can safely ignore any **WARNINGS** (in Yellow text) related to: "hdfscli", "hdfscli-avro", "pbr", "fastavro", "gen_client" and **ERRORS** (in Red text) related to the related to: "witwidget-gpu", "fairing" etc.

If you get any related errors or warnings mentioned above please rerun the above cell.

Note: Restart your kernel to use updated packages.

Make sure the Dataflow API is enabled by going to this link. Ensure that you've installed Beam by importing it and printing the version number.

In [1]:
```
import apache_beam as beam
print(beam.__version__)
```

```
2.24.0
```

In [2]:
```
import tensorflow as tf
print("TensorFlow version: ",tf.version.VERSION)
```

```
TensorFlow version:  2.5.0
```

You may receive a `UserWarning` about the Apache Beam SDK for Python 3 as not being yet fully supported. Don't worry about this.

In [3]:
```
# change these to try this notebook out
BUCKET = 'training-data-analyst'
PROJECT = 'qwiklabs-gcp-02-10a67a8d6c29'
REGION = 'us-central1'
```

In [4]:
```
import os
```

```
os.environ['BUCKET'] = BUCKET
os.environ['PROJECT'] = PROJECT
os.environ['REGION'] = REGION
```

In [5]:

```
%%bash
if ! gsutil ls | grep -q gs://${BUCKET}/; then
  gsutil mb -l ${REGION} gs://${BUCKET}
fi
```

Save the query from earlier

The data is natality data (record of births in the US). My goal is to predict the baby's weight given a number of factors about the pregnancy and the baby's mother. Later, we will want to split the data into training and eval datasets. The hash of the year-month will be used for that.

In [6]:

```
# Create SQL query using natality data after the year 2000
query = """
SELECT
  weight_pounds,
  is_male,
  mother_age,
  plurality,
  gestation_weeks,
  FARM_FINGERPRINT(CONCAT(CAST(YEAR AS STRING), CAST(month AS STRING))) AS hashm
FROM
  publicdata.samples.natality
WHERE year > 2000
"""
```

In [7]:

```
# Call BigQuery and examine in dataframe
from google.cloud import bigquery
df = bigquery.Client().query(query + " LIMIT 100").to_dataframe()
df.head()
```

Out[7]:

	weight_pounds	is_male	mother_age	plurality	gestation_weeks	hashmonth
0	7.063611	True	32	1	37.0	7108882242435606404
1	4.687028	True	30	3	33.0	-7170969733900686954
2	7.561856	True	20	1	39.0	6392072535155213407
3	7.561856	True	31	1	37.0	-2126480030009879160
4	7.312733	True	32	1	40.0	3408502330831153141

Create ML dataset using Dataflow

Let's use Cloud Dataflow to read in the BigQuery data, do some preprocessing, and write it out as CSV files.

Instead of using Beam/Dataflow, I had three other options:

- Use Cloud Dataprep to visually author a Dataflow pipeline. Cloud Dataprep also allows me to explore the data, so we could have avoided much of the handcoding of Python/Seaborn calls above as well!
- Read from BigQuery directly using TensorFlow.
- Use the BigQuery console (http://bigquery.cloud.google.com) to run a Query and save the result as a CSV file. For larger datasets, you may have to select the option to "allow large results" and save the result into a CSV file on Google Cloud Storage.

However, in this case, I want to do some preprocessing, modifying data so that we can simulate what is known if no ultrasound has been performed. If I didn't need preprocessing, I could have used the web console. Also, I prefer to script it out rather than run queries on the user interface, so I am using Cloud Dataflow for the preprocessing.

Note that after you launch this, the actual processing is happening on the cloud. Go to the GCP web console to the Dataflow section and monitor the running job. It took about 20 minutes for me.

If you wish to continue without doing this step, you can copy my preprocessed output:

```
gsutil -m cp -r gs://cloud-training-demos/babyweight/preproc
gs://your-bucket/
```

Lab Task #1: Creating datasets for ML using Dataflow

In [31]:

```
 # TODO 1
import datetime, os
def to_csv(rowdict):
    #Pull columns from BQ and create a line
    import hashlib
    import copy
    CSV_COLUMNS = 'weight_pounds, is_male, mother_age, Plurality, gestation_week
    # Create synthetic data where we assume that no ultrasound has been performe
    # and so we don't know sex of the baby. Let's assume that we can tell the di
    # between single and multiple, but that the errors rates in determining exac
    # is difficult in the absence of an ultrasound.
    no_ultrasound = copy.deepcopy(rowdict)
    w_ultrasound = copy.deepcopy(rowdict)

    no_ultrasound['is_male'] = 'Unknown'
    if rowdict['Plurality']>1:
        no_ultrasound['Plurality'] = 'Multiple(2+)'
    else:
        no_ultrasound['Plurality'] = 'Single(1)'

    #Chnage plurality to string
    w_ultrasound['Plurality'] = ['Single(1)', 'Twins(2)', 'Triplets(3)', 'Quadri

    # Write out two rows for each input row, one with ultrasound and one without
    for result in [no_ultrasound, w_ultrasound]:
        data = ','.join([str(result[k]) if k in result else 'None' for k in CSV_
        key = hashlib.sha224(data.encode('utf-8')).hexdigest()  # hash the colum
        yield str('{},{}'.format(data, key))

def preprocess(in_test_mode):
    import shutil, os, subprocess
```

```
job_name = 'preprocess-babyweight-features' + '-' + datetime.datetime.now().
if in_test_mode:
    print('Launching local job ... hang on')
    OUTPUT_DIR = './preproc'
    shutil.rmtree(OUTPUT_DIR, ignore_errors=True)
    os.makedirs(OUTPUT_DIR)
else:
    print('Launching Dataflow job {} ... hang on'.format(job_name))
    OUTPUT_DIR = 'gs://{0}/babyweight/preproc/'.format(BUCKET)
    try:
        subprocess.check_call('gsutil -m rm -r {}'.format(OUTPUT_DIR).split(
    except:
        pass

options = {
    'staging_location': os.path.join(OUTPUT_DIR, 'tmp', 'staging'),
    'temp_location': os.path.join(OUTPUT_DIR, 'tmp'),
    'job_name': job_name,
    'project': PROJECT,
    'region': REGION,
    'teardown_policy': 'TEARDOWN_ALWAYS',
    'no_save_main_session': True,
    'num_workers': 4,
    'max_num_workers': 5
}
opts = beam.pipeline.PipelineOptions(flags = [], **options)
if in_test_mode:
    RUNNER = 'DirectRunner'
else:
    RUNNER = 'DataflowRunner'
p = beam.Pipeline(RUNNER, options = opts)
query = """
SELECT
  weight_pounds,
  is_male,
  mother_age,
  plurality,
  gestation_weeks,
  FARM_FINGERPRINT(CONCAT(CAST(YEAR AS STRING), CAST(month AS STRING))) AS h
FROM
  publicdata.samples.natality
WHERE year > 2000
AND weight_pounds > 0
AND mother_age > 0
AND plurality > 0
AND gestation_weeks > 0
AND month > 0
    """

if in_test_mode:
    query = query + ' LIMIT 100'

for step in ['train', 'eval']:
    if step == 'train':
        selquery = 'SELECT * FROM ({}) WHERE ABS(MOD(hashmonth, 4)) < 3'.for
    else:
        selquery = 'SELECT * FROM ({}) WHERE ABS(MOD(hashmonth, 4)) = 3'.for

    (p
        | '{}_read'.format(step) >> beam.io.Read(beam.io.BigQuerySource(quer
        | '{}_csv'.format(step) >> beam.FlatMap(to_csv)
```

```
            | '{}_out'.format(step) >> beam.io.Write(beam.io.WriteToText(os.path
        )

    job = p.run()
    if in_test_mode:
        job.wait_until_finish()
        print("Done!")

preprocess(in_test_mode = False)
```

```
Launching Dataflow job preprocess-babyweight-features-210831-125149 ... hang on


WARNING:root:Make sure that locally built Python SDK docker image has Python 3.7
interpreter.
WARNING:apache_beam.options.pipeline_options:Discarding invalid overrides: {'tea
rdown_policy': 'TEARDOWN_ALWAYS', 'no_save_main_session': True}
WARNING:apache_beam.options.pipeline_options:Discarding invalid overrides: {'tea
rdown_policy': 'TEARDOWN_ALWAYS', 'no_save_main_session': True}
```

The above step will take 20+ minutes. Go to the GCP web console, navigate to the Dataflow section and **wait for the job to finish** before you run the following step.

Please re-run the above cell if you get a **failed status** of the job in the dataflow UI console.

In [32]:

```
%%bash
gsutil ls gs://${BUCKET}/babyweight/preproc/*-00000*
```

```
CommandException: One or more URLs matched no objects.
---------------------------------------------------------------------------
CalledProcessError                        Traceback (most recent call last)
<ipython-input-32-5265f5c54dcd> in <module>
----> 1 get_ipython().run_cell_magic('bash', '', 'gsutil ls gs://${BUCKET}/babyw
eight/preproc/*-00000*\n')

/opt/conda/lib/python3.7/site-packages/IPython/core/interactiveshell.py in run_c
ell_magic(self, magic_name, line, cell)
   2401             with self.builtin_trap:
   2402                 args = (magic_arg_s, cell)
-> 2403                 result = fn(*args, **kwargs)
   2404             return result
   2405

/opt/conda/lib/python3.7/site-packages/IPython/core/magics/script.py in named_sc
ript_magic(line, cell)
    140             else:
    141                 line = script
--> 142             return self.shebang(line, cell)
    143
    144         # write a basic docstring:

/opt/conda/lib/python3.7/site-packages/decorator.py in fun(*args, **kw)
    230             if not kwsyntax:
    231                 args, kw = fix(args, kw, sig)
--> 232             return caller(func, *(extras + args), **kw)
    233     fun.__name__ = func.__name__
    234     fun.__doc__ = func.__doc__

/opt/conda/lib/python3.7/site-packages/IPython/core/magic.py in <lambda>(f, *a,
 **k)
```

```
    185     # but it's overkill for just that one bit of state.
    186     def magic_deco(arg):
--> 187         call = lambda f, *a, **k: f(*a, **k)
    188 
    189         if callable(arg):

/opt/conda/lib/python3.7/site-packages/IPython/core/magics/script.py in shebang
(self, line, cell)
    243             sys.stderr.flush()
    244         if args.raise_error and p.returncode!=0:
--> 245             raise CalledProcessError(p.returncode, cell, output=out, std
err=err)
    246 
    247     def _run_script(self, p, cell, to_close):

CalledProcessError: Command 'b'gsutil ls gs://${BUCKET}/babyweight/preproc/*-000
00*\n'' returned non-zero exit status 1.
```

Copyright 2020 Google Inc. Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at http://www.apache.org/licenses/LICENSE-2.0 Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License