Lab: TfTransform

Learning Objectives

1. Preproccess data and engineer new features using TfTransform
2. Create and deploy Apache Beam pipeline
3. Use processed data to train taxifare model locally then serve a prediction

Introduction

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 allows for streaming. In this lab we will pull data from BigQuery then use Apache Beam TfTransform to process the data.

Only specific combinations of TensorFlow/Beam are supported by tf.transform so make sure to get a combo that works. In this lab we will be using:

- TFT 0.15.0
- TF 2.0
- Apache Beam [GCP] 2.16.0

NOTE: In the output of the next cell you may ignore any WARNINGS or ERRORS related to the following: "witwidget-gpu", "fairing", "pbr, "hdfscli", "hdfscli-avro", "fastavro", "plasma_store", and/or "gen_client".

In [1]:

```
!pip install --user apache-beam[gcp]==2.16.0
!pip install --user tensorflow-transform==0.15.0
```

```
Collecting apache-beam[gcp]==2.16.0
  Downloading apache_beam-2.16.0-cp37-cp37m-manylinux1_x86_64.whl
(3.0 MB)
     |████████████████████████████████| 3.0 MB 4.6 MB/s eta 0:00:01
Collecting mock<3.0.0,>=1.0.1
  Downloading mock-2.0.0-py2.py3-none-any.whl (56 kB)
     |████████████████████████████████| 56 kB 6.5 MB/s  eta 0:00:01
Requirement already satisfied: pytz>=2018.3 in /opt/conda/lib/python
3.7/site-packages (from apache-beam[gcp]==2.16.0) (2019.3)
Collecting avro-python3<2.0.0,>=1.8.1; python_version >= "3.0"
  Downloading avro-python3-1.9.2.1.tar.gz (37 kB)
Requirement already satisfied: grpcio<2,>=1.12.1 in /opt/conda/lib/p
ython3.7/site-packages (from apache-beam[gcp]==2.16.0) (1.27.2)
Collecting pymongo<4.0.0,>=3.8.0
  Downloading pymongo-3.10.1-cp37-cp37m-manylinux2014_x86_64.whl (46
2 kB)
     |████████████████████████████████| 462 kB 50.7 MB/s eta 0:00:01
Collecting crcmod<2.0,>=1.7
  Downloading crcmod-1.7.tar.gz (89 kB)
     |████████████████████████████████| 89 kB 9.2 MB/s  eta 0:00:01
Requirement already satisfied: protobuf<4,>=3.5.0.post1 in /opt/cond
a/lib/python3.7/site-packages (from apache-beam[gcp]==2.16.0) (3.11.
4)
Collecting dill<0.3.1,>=0.3.0
  Downloading dill-0.3.0.tar.gz (151 kB)
     |████████████████████████████████| 151 kB 43.7 MB/s eta 0:00:01
Collecting pyyaml<4.0.0,>=3.12
  Downloading PyYAML-3.13.tar.gz (270 kB)
     |████████████████████████████████| 270 kB 49.5 MB/s eta 0:00:01
Requirement already satisfied: pydot<2,>=1.2.0 in /opt/conda/lib/pyt
hon3.7/site-packages (from apache-beam[gcp]==2.16.0) (1.4.1)
Requirement already satisfied: future<1.0.0,>=0.16.0 in /opt/conda/l
ib/python3.7/site-packages (from apache-beam[gcp]==2.16.0) (0.18.2)
Collecting oauth2client<4,>=2.0.1
  Downloading oauth2client-3.0.0.tar.gz (77 kB)
     |████████████████████████████████| 77 kB 4.9 MB/s  eta 0:00:01
Collecting hdfs<3.0.0,>=2.1.0
  Downloading hdfs-2.5.8.tar.gz (41 kB)
     |████████████████████████████████| 41 kB 833 kB/s  eta 0:00:01
Collecting httplib2<=0.12.0,>=0.8
  Downloading httplib2-0.12.0.tar.gz (218 kB)
     |████████████████████████████████| 218 kB 48.5 MB/s eta 0:00:01
Collecting fastavro<0.22,>=0.21.4
  Downloading fastavro-0.21.24-cp37-cp37m-manylinux1_x86_64.whl (1.2
MB)
     |████████████████████████████████| 1.2 MB 56.6 MB/s eta 0:00:01
Requirement already satisfied: python-dateutil<3,>=2.8.0 in /opt/con
da/lib/python3.7/site-packages (from apache-beam[gcp]==2.16.0) (2.8.
1)
Collecting pyarrow<0.15.0,>=0.11.1; python_version >= "3.0" or platf
orm_system != "Windows"
  Downloading pyarrow-0.14.1-cp37-cp37m-manylinux2010_x86_64.whl (5
8.1 MB)
     |████████████████████████████████| 58.1 MB 6.0 kB/s  eta 0:00:0
1
Collecting google-cloud-bigquery<1.18.0,>=1.6.0; extra == "gcp"
  Downloading google_cloud_bigquery-1.17.1-py2.py3-none-any.whl (142
kB)
     |████████████████████████████████| 142 kB 59.8 MB/s eta 0:00:01
Collecting google-cloud-datastore<1.8.0,>=1.7.1; extra == "gcp"
  Downloading google_cloud_datastore-1.7.4-py2.py3-none-any.whl (82
kB)
     |████████████████████████████████| 82 kB 1.3 MB/s  eta 0:00:01
Collecting google-cloud-pubsub<1.1.0,>=0.39.0; extra == "gcp"
  Downloading google_cloud_pubsub-1.0.2-py2.py3-none-any.whl (118 k
B)
     |████████████████████████████████| 118 kB 67.6 MB/s eta 0:00:01
Requirement already satisfied: cachetools<4,>=3.1.0; extra == "gcp"
in /opt/conda/lib/python3.7/site-packages (from apache-beam[gcp]==2.
16.0) (3.1.1)
Collecting google-cloud-bigtable<1.1.0,>=0.31.1; extra == "gcp"
  Downloading google_cloud_bigtable-1.0.0-py2.py3-none-any.whl (232
kB)
     |████████████████████████████████| 232 kB 78.9 MB/s eta 0:00:01
Collecting google-apitools<0.5.29,>=0.5.28; extra == "gcp"
  Downloading google-apitools-0.5.28.tar.gz (172 kB)
     |████████████████████████████████| 172 kB 74.3 MB/s eta 0:00:01
Requirement already satisfied: google-cloud-core<2,>=0.28.1; extra =
= "gcp" in /opt/conda/lib/python3.7/site-packages (from apache-beam
[gcp]==2.16.0) (1.3.0)
Requirement already satisfied: six>=1.9 in /opt/conda/lib/python3.7/
site-packages (from mock<3.0.0,>=1.0.1->apache-beam[gcp]==2.16.0)
(1.14.0)
Collecting pbr>=0.11
  Downloading pbr-5.4.5-py2.py3-none-any.whl (110 kB)
     |████████████████████████████████| 110 kB 53.9 MB/s eta 0:00:01
Requirement already satisfied: setuptools in /opt/conda/lib/python3.
7/site-packages (from protobuf<4,>=3.5.0.post1->apache-beam[gcp]==2.
16.0) (46.1.3)
Requirement already satisfied: pyparsing>=2.1.4 in /opt/conda/lib/py
thon3.7/site-packages (from pydot<2,>=1.2.0->apache-beam[gcp]==2.16.
0) (2.4.6)
Requirement already satisfied: pyasn1>=0.1.7 in /opt/conda/lib/pytho
n3.7/site-packages (from oauth2client<4,>=2.0.1->apache-beam[gcp]==
2.16.0) (0.4.8)
Requirement already satisfied: pyasn1-modules>=0.0.5 in /opt/conda/l
ib/python3.7/site-packages (from oauth2client<4,>=2.0.1->apache-beam
[gcp]==2.16.0) (0.2.7)
Requirement already satisfied: rsa>=3.1.4 in /opt/conda/lib/python3.
7/site-packages (from oauth2client<4,>=2.0.1->apache-beam[gcp]==2.1
6.0) (4.0)
Collecting docopt
  Downloading docopt-0.6.2.tar.gz (25 kB)
Requirement already satisfied: requests>=2.7.0 in /opt/conda/lib/pyt
hon3.7/site-packages (from hdfs<3.0.0,>=2.1.0->apache-beam[gcp]==2.1
6.0) (2.23.0)
Requirement already satisfied: numpy>=1.14 in /opt/conda/lib/python
3.7/site-packages (from pyarrow<0.15.0,>=0.11.1; python_version >=
"3.0" or platform_system != "Windows"->apache-beam[gcp]==2.16.0) (1.
18.2)
Collecting google-resumable-media<0.5.0dev,>=0.3.1
  Downloading google_resumable_media-0.4.1-py2.py3-none-any.whl (38
kB)
Requirement already satisfied: google-api-core[grpc]<2.0.0dev,>=1.6.
0 in /opt/conda/lib/python3.7/site-packages (from google-cloud-datas
tore<1.8.0,>=1.7.1; extra == "gcp"->apache-beam[gcp]==2.16.0) (1.16.
0)
Requirement already satisfied: grpc-google-iam-v1<0.13dev,>=0.12.3 i
n /opt/conda/lib/python3.7/site-packages (from google-cloud-pubsub<
1.1.0,>=0.39.0; extra == "gcp"->apache-beam[gcp]==2.16.0) (0.12.3)
Collecting fasteners>=0.14
  Downloading fasteners-0.15-py2.py3-none-any.whl (23 kB)
```

```
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/
python3.7/site-packages (from requests>=2.7.0->hdfs<3.0.0,>=2.1.0->a
pache-beam[gcp]==2.16.0) (2019.11.28)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.2
1.1 in /opt/conda/lib/python3.7/site-packages (from requests>=2.7.0-
>hdfs<3.0.0,>=2.1.0->apache-beam[gcp]==2.16.0) (1.25.7)
Requirement already satisfied: chardet<4,>=3.0.2 in /opt/conda/lib/p
ython3.7/site-packages (from requests>=2.7.0->hdfs<3.0.0,>=2.1.0->ap
ache-beam[gcp]==2.16.0) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python
3.7/site-packages (from requests>=2.7.0->hdfs<3.0.0,>=2.1.0->apache-
beam[gcp]==2.16.0) (2.9)
Requirement already satisfied: google-auth<2.0dev,>=0.4.0 in /opt/co
nda/lib/python3.7/site-packages (from google-api-core[grpc]<2.0.0de
v,>=1.6.0->google-cloud-datastore<1.8.0,>=1.7.1; extra == "gcp"->apa
che-beam[gcp]==2.16.0) (1.11.2)
Requirement already satisfied: googleapis-common-protos<2.0dev,>=1.
6.0 in /opt/conda/lib/python3.7/site-packages (from google-api-core
[grpc]<2.0.0dev,>=1.6.0->google-cloud-datastore<1.8.0,>=1.7.1; extra
== "gcp"->apache-beam[gcp]==2.16.0) (1.51.0)
Collecting monotonic>=0.1
  Downloading monotonic-1.5-py2.py3-none-any.whl (5.3 kB)
Building wheels for collected packages: avro-python3, crcmod, dill,
pyyaml, oauth2client, hdfs, httplib2, google-apitools, docopt
  Building wheel for avro-python3 (setup.py) ... done
  Created wheel for avro-python3: filename=avro_python3-1.9.2.1-py3-
none-any.whl size=43513 sha256=8db4664609126de1055ac2f8e0c1c22892c67
003044f2127ff67756a059f567f
  Stored in directory: /home/jupyter/.cache/pip/wheels/bc/49/5f/fdb5
b9d85055c478213e0158ac122b596816149a02d82e0ab1
  Building wheel for crcmod (setup.py) ... done
  Created wheel for crcmod: filename=crcmod-1.7-cp37-cp37m-linux_x86
_64.whl size=36247 sha256=69b0f7df91782f347990c2ba6f3d540b6e71c35c9f
76299767553bbc54b1ffa9
  Stored in directory: /home/jupyter/.cache/pip/wheels/dc/9a/e9/49e6
27353476cec8484343c4ab656f1e0d783ee77b9dde2d1f
  Building wheel for dill (setup.py) ... done
  Created wheel for dill: filename=dill-0.3.0-py3-none-any.whl size=
77511 sha256=a6baf1db17c42e25aedee712ca464c83af30166351c8d274a11c4d6
8eed51e45
  Stored in directory: /home/jupyter/.cache/pip/wheels/6a/3c/26/1fcc
712c80b81fe1859f2dda4415f180fe9ef3ebe9f5e202e4
  Building wheel for pyyaml (setup.py) ... done
  Created wheel for pyyaml: filename=PyYAML-3.13-cp37-cp37m-linux_x8
6_64.whl size=43086 sha256=9afc52918304a7df82fb245cbbd613ed0ff321d0e
0fe9fe25e6a4c01f46d1ddd
  Stored in directory: /home/jupyter/.cache/pip/wheels/95/cd/14/899e
daa9cdb9a65aa7224539f6e0ad488e9a7b202bb48f6ae6
  Building wheel for oauth2client (setup.py) ... done
  Created wheel for oauth2client: filename=oauth2client-3.0.0-py3-no
ne-any.whl size=106383 sha256=263f93108f1d3d113eae937e7d66d2660a4fc7
3c40021c4c8ba7163edb7b66b4
  Stored in directory: /home/jupyter/.cache/pip/wheels/86/73/7a/3b3f
76a2142176605ff38fbca574327962c71e25a43197a4c1
  Building wheel for hdfs (setup.py) ... done
  Created wheel for hdfs: filename=hdfs-2.5.8-py3-none-any.whl size=
33213 sha256=7f1849e064a3ae6b3bcf08d450b5e3284749b340514ae58959cd8da
4feae9c6f
  Stored in directory: /home/jupyter/.cache/pip/wheels/0a/7d/38/ea4e
af831518e6cd867b515b88919a9785eb66f11def5ab859
  Building wheel for httplib2 (setup.py) ... done
  Created wheel for httplib2: filename=httplib2-0.12.0-py3-none-any.
whl size=93464 sha256=8f55fb3fa1d6dff707626b68d0879a56b90d6c547b19ff
8472c47f12fea7c1fc
  Stored in directory: /home/jupyter/.cache/pip/wheels/0d/e7/b6/0dd3
0343ceca921cfbd91f355041bd9c69e0f40b49f25b7b8a
  Building wheel for google-apitools (setup.py) ... done
  Created wheel for google-apitools: filename=google_apitools-0.5.28
-py3-none-any.whl size=130110 sha256=23a14001da07ad0b8095abe2eae49ad
2314775ec3a3b0ad97966bc9ffdc5383d
  Stored in directory: /home/jupyter/.cache/pip/wheels/34/3b/69/ecd8
e6ae89d9d71102a58962c29faa7a9467ba45f99f205920
  Building wheel for docopt (setup.py) ... done
  Created wheel for docopt: filename=docopt-0.6.2-py2.py3-none-any.w
hl size=13704 sha256=2e742af7b56311a694cef01984b3b12fbdb310a88cca076
581cc3ffc4c752e81
  Stored in directory: /home/jupyter/.cache/pip/wheels/72/b0/3f/1d95
f96ff986c7dfffe46ce2be4062f38ebd04b506c77c81b9
Successfully built avro-python3 crcmod dill pyyaml oauth2client hdfs
httplib2 google-apitools docopt
ERROR: google-cloud-storage 1.26.0 has requirement google-resumable-
media<0.6dev,>=0.5.0, but you'll have google-resumable-media 0.4.1 w
hich is incompatible.
Installing collected packages: pbr, mock, avro-python3, pymongo, crc
mod, dill, pyyaml, httplib2, oauth2client, docopt, hdfs, fastavro, p
yarrow, google-resumable-media, google-cloud-bigquery, google-cloud-
datastore, google-cloud-pubsub, google-cloud-bigtable, monotonic, fa
steners, google-apitools, apache-beam
  WARNING: The script pbr is installed in '/home/jupyter/.local/bin'
which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppre
ss this warning, use --no-warn-script-location.
  WARNING: The scripts hdfscli and hdfscli-avro are installed in '/h
ome/jupyter/.local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppre
ss this warning, use --no-warn-script-location.
  WARNING: The script fastavro is installed in '/home/jupyter/.loca
l/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppre
ss this warning, use --no-warn-script-location.
  WARNING: The script plasma_store is installed in '/home/jupyter/.l
ocal/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppre
ss this warning, use --no-warn-script-location.
  WARNING: The script gen_client is installed in '/home/jupyter/.loc
al/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppre
ss this warning, use --no-warn-script-location.
Successfully installed apache-beam-2.16.0 avro-python3-1.9.2.1 crcmo
d-1.7 dill-0.3.0 docopt-0.6.2 fastavro-0.21.24 fasteners-0.15 google
-apitools-0.5.28 google-cloud-bigquery-1.17.1 google-cloud-bigtable-
1.0.0 google-cloud-datastore-1.7.4 google-cloud-pubsub-1.0.2 google-
resumable-media-0.4.1 hdfs-2.5.8 httplib2-0.12.0 mock-2.0.0 monotoni
c-1.5 oauth2client-3.0.0 pbr-5.4.5 pyarrow-0.14.1 pymongo-3.10.1 pyy
aml-3.13
Collecting tensorflow-transform==0.15.0
  Downloading tensorflow-transform-0.15.0.tar.gz (222 kB)
     |████████████████████████████████| 222 kB 4.2 MB/s eta 0:00:01
Collecting absl-py<0.9,>=0.7
  Downloading absl-py-0.8.1.tar.gz (103 kB)
     |████████████████████████████████| 103 kB 9.3 MB/s eta 0:00:01
Requirement already satisfied: apache-beam[gcp]<3,>=2.16 in /home/ju
```

```
pyter/.local/lib/python3.7/site-packages (from tensorflow-transform=
=0.15.0) (2.16.0)
Requirement already satisfied: numpy<2,>=1.16 in /opt/conda/lib/pyth
on3.7/site-packages (from tensorflow-transform==0.15.0) (1.18.2)
Requirement already satisfied: protobuf<4,>=3.7 in /opt/conda/lib/py
thon3.7/site-packages (from tensorflow-transform==0.15.0) (3.11.4)
Requirement already satisfied: pydot<2,>=1.2 in /opt/conda/lib/pytho
n3.7/site-packages (from tensorflow-transform==0.15.0) (1.4.1)
Requirement already satisfied: six<2,>=1.10 in /opt/conda/lib/python
3.7/site-packages (from tensorflow-transform==0.15.0) (1.14.0)
Collecting tensorflow-metadata<0.16,>=0.15
  Downloading tensorflow_metadata-0.15.2-py2.py3-none-any.whl (29 k
B)
Requirement already satisfied: tensorflow<2.2,>=1.15 in /opt/conda/l
ib/python3.7/site-packages (from tensorflow-transform==0.15.0) (1.1
5.2)
Collecting tfx-bsl<0.16,>=0.15
  Downloading tfx_bsl-0.15.3-cp37-cp37m-manylinux2010_x86_64.whl (1.
9 MB)
     |████████████████████████████████| 1.9 MB 8.8 MB/s eta 0:00:01
Requirement already satisfied: pyarrow<0.15.0,>=0.11.1; python_versi
on >= "3.0" or platform_system != "Windows" in /home/jupyter/.local/
lib/python3.7/site-packages (from apache-beam[gcp]<3,>=2.16->tensorf
low-transform==0.15.0) (0.14.1)
Requirement already satisfied: pytz>=2018.3 in /opt/conda/lib/python
3.7/site-packages (from apache-beam[gcp]<3,>=2.16->tensorflow-transf
orm==0.15.0) (2019.3)
Requirement already satisfied: future<1.0.0,>=0.16.0 in /opt/conda/l
ib/python3.7/site-packages (from apache-beam[gcp]<3,>=2.16->tensorfl
ow-transform==0.15.0) (0.18.2)
Requirement already satisfied: oauth2client<4,>=2.0.1 in /home/jupyt
er/.local/lib/python3.7/site-packages (from apache-beam[gcp]<3,>=2.1
6->tensorflow-transform==0.15.0) (3.0.0)
Requirement already satisfied: python-dateutil<3,>=2.8.0 in /opt/con
da/lib/python3.7/site-packages (from apache-beam[gcp]<3,>=2.16->tens
orflow-transform==0.15.0) (2.8.1)
Requirement already satisfied: mock<3.0.0,>=1.0.1 in /home/jupyter/.
local/lib/python3.7/site-packages (from apache-beam[gcp]<3,>=2.16->t
ensorflow-transform==0.15.0) (2.0.0)
Requirement already satisfied: httplib2<=0.12.0,>=0.8 in /home/jupyt
er/.local/lib/python3.7/site-packages (from apache-beam[gcp]<3,>=2.1
6->tensorflow-transform==0.15.0) (0.12.0)
Requirement already satisfied: dill<0.3.1,>=0.3.0 in /home/jupyter/.
local/lib/python3.7/site-packages (from apache-beam[gcp]<3,>=2.16->t
ensorflow-transform==0.15.0) (0.3.0)
Requirement already satisfied: crcmod<2.0,>=1.7 in /home/jupyter/.lo
cal/lib/python3.7/site-packages (from apache-beam[gcp]<3,>=2.16->ten
sorflow-transform==0.15.0) (1.7)
Requirement already satisfied: grpcio<2,>=1.12.1 in /opt/conda/lib/p
ython3.7/site-packages (from apache-beam[gcp]<3,>=2.16->tensorflow-t
ransform==0.15.0) (1.27.2)
Requirement already satisfied: fastavro<0.22,>=0.21.4 in /home/jupyt
er/.local/lib/python3.7/site-packages (from apache-beam[gcp]<3,>=2.1
6->tensorflow-transform==0.15.0) (0.21.24)
Requirement already satisfied: avro-python3<2.0.0,>=1.8.1; python_ve
rsion >= "3.0" in /home/jupyter/.local/lib/python3.7/site-packages
 (from apache-beam[gcp]<3,>=2.16->tensorflow-transform==0.15.0) (1.
9.2.1)
Requirement already satisfied: hdfs<3.0.0,>=2.1.0 in /home/jupyter/.
local/lib/python3.7/site-packages (from apache-beam[gcp]<3,>=2.16->t
ensorflow-transform==0.15.0) (2.5.8)
```

```
Requirement already satisfied: pymongo<4.0.0,>=3.8.0 in /home/jupyte
r/.local/lib/python3.7/site-packages (from apache-beam[gcp]<3,>=2.16
->tensorflow-transform==0.15.0) (3.10.1)
Requirement already satisfied: pyyaml<4.0.0,>=3.12 in /home/jupyte
r/.local/lib/python3.7/site-packages (from apache-beam[gcp]<3,>=2.16
->tensorflow-transform==0.15.0) (3.13)
Requirement already satisfied: google-cloud-pubsub<1.1.0,>=0.39.0; e
xtra == "gcp" in /home/jupyter/.local/lib/python3.7/site-packages (f
rom apache-beam[gcp]<3,>=2.16->tensorflow-transform==0.15.0) (1.0.2)
Requirement already satisfied: google-cloud-datastore<1.8.0,>=1.7.1;
extra == "gcp" in /home/jupyter/.local/lib/python3.7/site-packages
 (from apache-beam[gcp]<3,>=2.16->tensorflow-transform==0.15.0) (1.
7.4)
Requirement already satisfied: cachetools<4,>=3.1.0; extra == "gcp"
 in /opt/conda/lib/python3.7/site-packages (from apache-beam[gcp]<3,
>=2.16->tensorflow-transform==0.15.0) (3.1.1)
Requirement already satisfied: google-cloud-core<2,>=0.28.1; extra =
= "gcp" in /opt/conda/lib/python3.7/site-packages (from apache-beam
[gcp]<3,>=2.16->tensorflow-transform==0.15.0) (1.3.0)
Requirement already satisfied: google-apitools<0.5.29,>=0.5.28; extr
a == "gcp" in /home/jupyter/.local/lib/python3.7/site-packages (from
apache-beam[gcp]<3,>=2.16->tensorflow-transform==0.15.0) (0.5.28)
Requirement already satisfied: google-cloud-bigtable<1.1.0,>=0.31.1;
extra == "gcp" in /home/jupyter/.local/lib/python3.7/site-packages
 (from apache-beam[gcp]<3,>=2.16->tensorflow-transform==0.15.0) (1.
0.0)
Requirement already satisfied: google-cloud-bigquery<1.18.0,>=1.6.0;
extra == "gcp" in /home/jupyter/.local/lib/python3.7/site-packages
 (from apache-beam[gcp]<3,>=2.16->tensorflow-transform==0.15.0) (1.1
7.1)
Requirement already satisfied: setuptools in /opt/conda/lib/python3.
7/site-packages (from protobuf<4,>=3.7->tensorflow-transform==0.15.
0) (46.1.3)
Requirement already satisfied: pyparsing>=2.1.4 in /opt/conda/lib/py
thon3.7/site-packages (from pydot<2,>=1.2->tensorflow-transform==0.1
5.0) (2.4.6)
Requirement already satisfied: googleapis-common-protos in /opt/cond
a/lib/python3.7/site-packages (from tensorflow-metadata<0.16,>=0.15-
>tensorflow-transform==0.15.0) (1.51.0)
Requirement already satisfied: tensorboard<1.16.0,>=1.15.0 in /opt/c
onda/lib/python3.7/site-packages (from tensorflow<2.2,>=1.15->tensor
flow-transform==0.15.0) (1.15.0)
Requirement already satisfied: gast==0.2.2 in /opt/conda/lib/python
3.7/site-packages (from tensorflow<2.2,>=1.15->tensorflow-transform=
=0.15.0) (0.2.2)
Requirement already satisfied: astor>=0.6.0 in /opt/conda/lib/python
3.7/site-packages (from tensorflow<2.2,>=1.15->tensorflow-transform=
=0.15.0) (0.8.1)
Requirement already satisfied: keras-preprocessing>=1.0.5 in /opt/co
nda/lib/python3.7/site-packages (from tensorflow<2.2,>=1.15->tensorf
low-transform==0.15.0) (1.1.0)
Requirement already satisfied: tensorflow-estimator==1.15.1 in /opt/
conda/lib/python3.7/site-packages (from tensorflow<2.2,>=1.15->tenso
rflow-transform==0.15.0) (1.15.1)
Requirement already satisfied: google-pasta>=0.1.6 in /opt/conda/li
b/python3.7/site-packages (from tensorflow<2.2,>=1.15->tensorflow-tr
ansform==0.15.0) (0.2.0)
Requirement already satisfied: termcolor>=1.1.0 in /opt/conda/lib/py
thon3.7/site-packages (from tensorflow<2.2,>=1.15->tensorflow-transf
orm==0.15.0) (1.1.0)
Requirement already satisfied: wrapt>=1.11.1 in /opt/conda/lib/pytho
```

```
n3.7/site-packages (from tensorflow<2.2,>=1.15->tensorflow-transform
==0.15.0) (1.12.1)
Requirement already satisfied: wheel>=0.26; python_version >= "3" in
/opt/conda/lib/python3.7/site-packages (from tensorflow<2.2,>=1.15->
tensorflow-transform==0.15.0) (0.34.2)
Requirement already satisfied: keras-applications>=1.0.8 in /opt/con
da/lib/python3.7/site-packages (from tensorflow<2.2,>=1.15->tensorfl
ow-transform==0.15.0) (1.0.8)
Requirement already satisfied: opt-einsum>=2.3.2 in /opt/conda/lib/p
ython3.7/site-packages (from tensorflow<2.2,>=1.15->tensorflow-trans
form==0.15.0) (3.2.0)
Requirement already satisfied: tensorflow-serving-api<3,>=1.15 in /o
pt/conda/lib/python3.7/site-packages (from tfx-bsl<0.16,>=0.15->tens
orflow-transform==0.15.0) (1.15.0)
Requirement already satisfied: psutil<6,>=5.6 in /opt/conda/lib/pyth
on3.7/site-packages (from tfx-bsl<0.16,>=0.15->tensorflow-transform=
=0.15.0) (5.7.0)
Requirement already satisfied: rsa>=3.1.4 in /opt/conda/lib/python3.
7/site-packages (from oauth2client<4,>=2.0.1->apache-beam[gcp]<3,>=
2.16->tensorflow-transform==0.15.0) (4.0)
Requirement already satisfied: pyasn1-modules>=0.0.5 in /opt/conda/l
ib/python3.7/site-packages (from oauth2client<4,>=2.0.1->apache-beam
[gcp]<3,>=2.16->tensorflow-transform==0.15.0) (0.2.7)
Requirement already satisfied: pyasn1>=0.1.7 in /opt/conda/lib/pytho
n3.7/site-packages (from oauth2client<4,>=2.0.1->apache-beam[gcp]<3,
>=2.16->tensorflow-transform==0.15.0) (0.4.8)
Requirement already satisfied: pbr>=0.11 in /home/jupyter/.local/li
b/python3.7/site-packages (from mock<3.0.0,>=1.0.1->apache-beam[gcp]
<3,>=2.16->tensorflow-transform==0.15.0) (5.4.5)
Requirement already satisfied: requests>=2.7.0 in /opt/conda/lib/pyt
hon3.7/site-packages (from hdfs<3.0.0,>=2.1.0->apache-beam[gcp]<3,>=
2.16->tensorflow-transform==0.15.0) (2.23.0)
Requirement already satisfied: docopt in /home/jupyter/.local/lib/py
thon3.7/site-packages (from hdfs<3.0.0,>=2.1.0->apache-beam[gcp]<3,>
=2.16->tensorflow-transform==0.15.0) (0.6.2)
Requirement already satisfied: grpc-google-iam-v1<0.13dev,>=0.12.3 i
n /opt/conda/lib/python3.7/site-packages (from google-cloud-pubsub<
1.1.0,>=0.39.0; extra == "gcp"->apache-beam[gcp]<3,>=2.16->tensorflo
w-transform==0.15.0) (0.12.3)
Requirement already satisfied: google-api-core[grpc]<2.0.0dev,>=1.1
4.0 in /opt/conda/lib/python3.7/site-packages (from google-cloud-pub
sub<1.1.0,>=0.39.0; extra == "gcp"->apache-beam[gcp]<3,>=2.16->tenso
rflow-transform==0.15.0) (1.16.0)
Requirement already satisfied: fasteners>=0.14 in /home/jupyter/.loc
al/lib/python3.7/site-packages (from google-apitools<0.5.29,>=0.5.2
8; extra == "gcp"->apache-beam[gcp]<3,>=2.16->tensorflow-transform==
0.15.0) (0.15)
Requirement already satisfied: google-resumable-media<0.5.0dev,>=0.
3.1 in /home/jupyter/.local/lib/python3.7/site-packages (from google
-cloud-bigquery<1.18.0,>=1.6.0; extra == "gcp"->apache-beam[gcp]<3,>
=2.16->tensorflow-transform==0.15.0) (0.4.1)
Requirement already satisfied: markdown>=2.6.8 in /opt/conda/lib/pyt
hon3.7/site-packages (from tensorboard<1.16.0,>=1.15.0->tensorflow<
2.2,>=1.15->tensorflow-transform==0.15.0) (3.2.1)
Requirement already satisfied: werkzeug>=0.11.15 in /opt/conda/lib/p
ython3.7/site-packages (from tensorboard<1.16.0,>=1.15.0->tensorflow
<2.2,>=1.15->tensorflow-transform==0.15.0) (1.0.0)
Requirement already satisfied: h5py in /opt/conda/lib/python3.7/site
-packages (from keras-applications>=1.0.8->tensorflow<2.2,>=1.15->te
nsorflow-transform==0.15.0) (2.10.0)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/
```

```
python3.7/site-packages (from requests>=2.7.0->hdfs<3.0.0,>=2.1.0->a
pache-beam[gcp]<3,>=2.16->tensorflow-transform==0.15.0) (2019.11.28)
Requirement already satisfied: chardet<4,>=3.0.2 in /opt/conda/lib/p
ython3.7/site-packages (from requests>=2.7.0->hdfs<3.0.0,>=2.1.0->ap
ache-beam[gcp]<3,>=2.16->tensorflow-transform==0.15.0) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python
3.7/site-packages (from requests>=2.7.0->hdfs<3.0.0,>=2.1.0->apache-
beam[gcp]<3,>=2.16->tensorflow-transform==0.15.0) (2.9)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.2
1.1 in /opt/conda/lib/python3.7/site-packages (from requests>=2.7.0-
>hdfs<3.0.0,>=2.1.0->apache-beam[gcp]<3,>=2.16->tensorflow-transform
==0.15.0) (1.25.7)
Requirement already satisfied: google-auth<2.0dev,>=0.4.0 in /opt/co
nda/lib/python3.7/site-packages (from google-api-core[grpc]<2.0.0de
v,>=1.14.0->google-cloud-pubsub<1.1.0,>=0.39.0; extra == "gcp"->apac
he-beam[gcp]<3,>=2.16->tensorflow-transform==0.15.0) (1.11.2)
Requirement already satisfied: monotonic>=0.1 in /home/jupyter/.loca
l/lib/python3.7/site-packages (from fasteners>=0.14->google-apitools
<0.5.29,>=0.5.28; extra == "gcp"->apache-beam[gcp]<3,>=2.16->tensorf
low-transform==0.15.0) (1.5)
Building wheels for collected packages: tensorflow-transform, absl-p
y
  Building wheel for tensorflow-transform (setup.py) ... done
  Created wheel for tensorflow-transform: filename=tensorflow_transf
orm-0.15.0-py3-none-any.whl size=280591 sha256=feced3fe57ea4991395fc
b980cbf59931d938b1de1e8da37c31dbb8b6fd73612
  Stored in directory: /home/jupyter/.cache/pip/wheels/25/9e/5a/3616
db66925c4a6ff4fdf1666f0b1ff869247519683aec02cd
  Building wheel for absl-py (setup.py) ... done
  Created wheel for absl-py: filename=absl_py-0.8.1-py3-none-any.whl
size=121165 sha256=05c07f583a89fca5754e82cc20641666db93177f4d10922e6
9f200077c5f8c79
  Stored in directory: /home/jupyter/.cache/pip/wheels/46/91/e3/0fce
d4f5fbc0a051a5667096826186c9ff60f2d0e9bf0f1cdc
Successfully built tensorflow-transform absl-py
Installing collected packages: absl-py, tensorflow-metadata, tfx-bs
l, tensorflow-transform
Successfully installed absl-py-0.8.1 tensorflow-metadata-0.15.2 tens
orflow-transform-0.15.0 tfx-bsl-0.15.3
```

Download .whl file for tensorflow-transform. We will pass this file to Beam Pipeline Options so it is installed on the DataFlow workers

In [2]:

```
!pip download tensorflow-transform==0.15.0 --no-deps
```

```
Collecting tensorflow-transform==0.15.0
  Using cached tensorflow-transform-0.15.0.tar.gz (222 kB)
  Saved ./tensorflow-transform-0.15.0.tar.gz
Successfully downloaded tensorflow-transform
```

Restart the kernel (click on the reload button above - beside the word "Markdown").

In [1]:

```
%%bash
pip freeze | grep -e 'flow\|beam'
```

```
apache-beam==2.16.0
tensorflow==1.15.2
tensorflow-datasets==1.2.0
tensorflow-estimator==1.15.1
tensorflow-hub==0.6.0
tensorflow-io==0.8.1
tensorflow-metadata==0.15.2
tensorflow-probability==0.8.0
tensorflow-serving-api==1.15.0
tensorflow-transform==0.15.0
```

In [2]:

```
import tensorflow as tf
import tensorflow_transform as tft
import shutil
print(tf.__version__)
```

```
1.15.2
```

In [4]:

```
# change these to those of your environment to try this notebook out

BUCKET = 'qwiklabs-gcp-03-b02dedbd6a51'
PROJECT = 'qwiklabs-gcp-03-b02dedbd6a51'
REGION = 'us-central1'
```

In [5]:

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

In [6]:

```
%%bash
gcloud config set project $PROJECT
gcloud config set compute/region $REGION
```

```
Updated property [core/project].
Updated property [compute/region].
```

In [7]:

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

Input source: BigQuery

Get data from BigQuery but defer the majority of filtering etc. to Beam. Note that the dayofweek column is now strings.

In [8]:

```
from google.cloud import bigquery


def create_query(phase, EVERY_N):
    """Creates a query with the proper splits.

    Args:
        phase: int, 1=train, 2=valid.
        EVERY_N: int, take an example EVERY_N rows.

    Returns:
        Query string with the proper splits.
    """
    base_query = """
    WITH daynames AS
    (SELECT ['Sun', 'Mon', 'Tues', 'Wed', 'Thurs', 'Fri', 'Sat'] AS daysofweek)
    SELECT
    (tolls_amount + fare_amount) AS fare_amount,
    daysofweek[ORDINAL(EXTRACT(DAYOFWEEK FROM pickup_datetime))] AS dayofweek,
    EXTRACT(HOUR FROM pickup_datetime) AS hourofday,
    pickup_longitude AS pickuplon,
    pickup_latitude AS pickuplat,
    dropoff_longitude AS dropofflon,
    dropoff_latitude AS dropofflat,
    passenger_count AS passengers,
    'notneeded' AS key
    FROM
    `nyc-tlc.yellow.trips`, daynames
    WHERE
    trip_distance > 0 AND fare_amount > 0
    """
    if EVERY_N is None:
        if phase < 2:
            # training
            query = """{0} AND ABS(MOD(FARM_FINGERPRINT(CAST
            (pickup_datetime AS STRING), 4)) < 2""".format(base_query)
        else:
            query = """{0} AND ABS(MOD(FARM_FINGERPRINT(CAST(
            pickup_datetime AS STRING), 4)) = {1}""".format(base_query, phase)
    else:
        query = """{0} AND ABS(MOD(FARM_FINGERPRINT(CAST(
        pickup_datetime AS STRING)), {1})) = {2}""".format(
            base_query, EVERY_N, phase)

    return query

query = create_query(2, 100000)
```

Let's pull this query down into a Pandas DataFrame and take a look at some of the statistics.

In [9]:

```
df_valid = bigquery.Client().query(query).to_dataframe()
display(df_valid.head())
df_valid.describe()
```

	fare_amount	dayofweek	hourofday	pickuplon	pickuplat	dropofflon	dropofflat	passeng
0	8.5	Sat	0	-74.004418	40.742525	-73.987448	40.760442	
1	5.0	Mon	0	-74.012780	40.701832	-74.013807	40.709285	
2	29.3	Sat	0	-73.983300	40.744700	-73.960800	40.617400	
3	17.5	Thurs	0	-73.976814	40.739868	-73.957535	40.704876	
4	5.5	Sun	0	-73.948690	40.717057	-73.952610	40.726865	

Out[9]:

	fare_amount	hourofday	pickuplon	pickuplat	dropofflon	dropofflat
count	11181.000000	11181.000000	11181.000000	11181.000000	11181.000000	11181.000000
mean	11.242599	13.244075	-72.576852	39.973146	-72.748974	40.006091
std	9.447462	6.548354	10.133452	5.777329	12.981577	5.664887
min	2.500000	0.000000	-78.133333	-73.991278	-751.400000	-73.977970
25%	6.000000	9.000000	-73.991849	40.734954	-73.991236	40.734008
50%	8.500000	14.000000	-73.981824	40.752640	-73.980164	40.753427
75%	12.500000	19.000000	-73.967418	40.766700	-73.964153	40.767832
max	143.000000	23.000000	40.806487	41.366138	40.785400	41.366138

Create ML dataset using tf.transform and Dataflow

Let's use Cloud Dataflow to read in the BigQuery data and write it out as TFRecord files. Along the way, let's use tf.transform to do scaling and transforming. Using tf.transform allows us to save the metadata to ensure that the appropriate transformations get carried out during prediction as well.

NOTE: You may ignore any WARNING related to "tensorflow" in the output after executing the code cell below.

`transformed_data` is type `pcollection` .

In [10]:

```
import datetime
import tensorflow as tf
import apache_beam as beam
import tensorflow_transform as tft
import tensorflow_metadata as tfmd
from tensorflow_transform.beam import impl as beam_impl


def is_valid(inputs):
    """Check to make sure the inputs are valid.

    Args:
        inputs: dict, dictionary of TableRow data from BigQuery.

    Returns:
        True if the inputs are valid and False if they are not.
    """
    try:
        pickup_longitude = inputs['pickuplon']
        dropoff_longitude = inputs['dropofflon']
        pickup_latitude = inputs['pickuplat']
        dropoff_latitude = inputs['dropofflat']
        hourofday = inputs['hourofday']
        dayofweek = inputs['dayofweek']
        passenger_count = inputs['passengers']
        fare_amount = inputs['fare_amount']
        return fare_amount >= 2.5 and pickup_longitude > -78 \
            and pickup_longitude < -70 and dropoff_longitude > -78 \
            and dropoff_longitude < -70 and pickup_latitude > 37 \
            and pickup_latitude < 45 and dropoff_latitude > 37 \
            and dropoff_latitude < 45 and passenger_count > 0
    except:
        return False


def preprocess_tft(inputs):
    """Preproccess the features and add engineered features with tf transform.

    Args:
        dict, dictionary of TableRow data from BigQuery.

    Returns:
        Dictionary of preprocessed data after scaling and feature engineering.
    """
    import datetime
    print(inputs)
    result = {}
    result['fare_amount'] = tf.identity(inputs['fare_amount'])
    # build a vocabulary
    result['dayofweek'] = tft.string_to_int(inputs['dayofweek'])
    result['hourofday'] = tf.identity(inputs['hourofday'])  # pass through
    # scaling numeric values
    result['pickuplon'] = (tft.scale_to_0_1(inputs['pickuplon']))
    result['pickuplat'] = (tft.scale_to_0_1(inputs['pickuplat']))
    result['dropofflon'] = (tft.scale_to_0_1(inputs['dropofflon']))
    result['dropofflat'] = (tft.scale_to_0_1(inputs['dropofflat']))
    result['passengers'] = tf.cast(inputs['passengers'], tf.float32)  # a cast
    # arbitrary TF func
    result['key'] = tf.as_string(tf.ones_like(inputs['passengers']))
    # engineered features
    latdiff = inputs['pickuplat'] - inputs['dropofflat']
    londiff = inputs['pickuplon'] - inputs['dropofflon']
    result['latdiff'] = tft.scale_to_0_1(latdiff)
    result['londiff'] = tft.scale_to_0_1(londiff)
    dist = tf.sqrt(latdiff * latdiff + londiff * londiff)
    result['euclidean'] = tft.scale_to_0_1(dist)
    return result


def preprocess(in_test_mode):
    """Sets up preprocess pipeline.

    Args:
        in_test_mode: bool, False to launch DataFlow job, True to run locally.
    """
    import os
    import os.path
    import tempfile
    from apache_beam.io import tfrecordio
    from tensorflow_transform.coders import example_proto_coder
    from tensorflow_transform.tf_metadata import dataset_metadata
    from tensorflow_transform.tf_metadata import dataset_schema
    from tensorflow_transform.beam import tft_beam_io
    from tensorflow_transform.beam.tft_beam_io import transform_fn_io

    job_name = 'preprocess-taxi-features' + '-'
    job_name += datetime.datetime.now().strftime('%y%m%d-%H%M%S')
    if in_test_mode:
        import shutil
        print('Launching local job ... hang on')
        OUTPUT_DIR = './preproc_tft'
        shutil.rmtree(OUTPUT_DIR, ignore_errors=True)
        EVERY_N = 100000
    else:
        print('Launching Dataflow job {} ... hang on'.format(job_name))
        OUTPUT_DIR = 'gs://{0}/taxifare/preproc_tft/'.format(BUCKET)
        import subprocess
        subprocess.call('gsutil rm -r {}'.format(OUTPUT_DIR).split())
        EVERY_N = 10000

    options = {
        'staging_location': os.path.join(OUTPUT_DIR, 'tmp', 'staging'),
        'temp_location': os.path.join(OUTPUT_DIR, 'tmp'),
        'job_name': job_name,
        'project': PROJECT,
        'num_workers': 1,
        'max_num_workers': 1,
        'teardown_policy': 'TEARDOWN_ALWAYS',
        'no_save_main_session': True,
        'direct_num_workers': 1,
        'extra_packages': ['tensorflow-transform-0.15.0.tar.gz']
        }

    opts = beam.pipeline.PipelineOptions(flags=[], **options)
    if in_test_mode:
        RUNNER = 'DirectRunner'
    else:
        RUNNER = 'DataflowRunner'

    # Set up raw data metadata
    raw_data_schema = {
        colname: dataset_schema.ColumnSchema(
            tf.string, [], dataset_schema.FixedColumnRepresentation())
        for colname in 'dayofweek,key'.split(',')
    }

    raw_data_schema.update({
        colname: dataset_schema.ColumnSchema(
            tf.float32, [], dataset_schema.FixedColumnRepresentation())
        for colname in
        'fare_amount,pickuplon,pickuplat,dropofflon,dropofflat'.split(',')
    })

    raw_data_schema.update({
        colname: dataset_schema.ColumnSchema(
            tf.int64, [], dataset_schema.FixedColumnRepresentation())
        for colname in 'hourofday,passengers'.split(',')
    })

    raw_data_metadata = dataset_metadata.DatasetMetadata(
        dataset_schema.Schema(raw_data_schema))

    # Run Beam
    with beam.Pipeline(RUNNER, options=opts) as p:
        with beam_impl.Context(temp_dir=os.path.join(OUTPUT_DIR, 'tmp')):
            # Save the raw data metadata
            (raw_data_metadata |
                'WriteInputMetadata' >> tft_beam_io.WriteMetadata(
                    os.path.join(
                        OUTPUT_DIR, 'metadata/rawdata_metadata'), pipeline=p))

            # Read training data from bigquery and filter rows
            raw_data = (p | 'train_read' >> beam.io.Read(
                    beam.io.BigQuerySource(
                        query=create_query(1, EVERY_N),
                        use_standard_sql=True)) |
                        'train_filter' >> beam.Filter(is_valid))

            raw_dataset = (raw_data, raw_data_metadata)

            # Analyze and transform training data
            transformed_dataset, transform_fn = (
                raw_dataset | beam_impl.AnalyzeAndTransformDataset(
                    preprocess_tft))
            transformed_data, transformed_metadata = transformed_dataset

            # Save transformed train data to disk in efficient tfrecord format
            transformed_data | 'WriteTrainData' >> tfrecordio.WriteToTFRecord(
                os.path.join(OUTPUT_DIR, 'train'), file_name_suffix='.gz',
                coder=example_proto_coder.ExampleProtoCoder(
                    transformed_metadata.schema))

            # Read eval data from bigquery and filter rows
            raw_test_data = (p | 'eval_read' >> beam.io.Read(
                beam.io.BigQuerySource(
                    query=create_query(2, EVERY_N),
                    use_standard_sql=True)) | 'eval_filter' >> beam.Filter(
                        is_valid))

            raw_test_dataset = (raw_test_data, raw_data_metadata)
            # Transform eval data
            transformed_test_dataset = (
                (raw_test_dataset, transform_fn) | beam_impl.TransformDataset()
                )
            transformed_test_data, _ = transformed_test_dataset

            # Save transformed train data to disk in efficient tfrecord format
            (transformed_test_data |
                'WriteTestData' >> tfrecordio.WriteToTFRecord(
                    os.path.join(OUTPUT_DIR, 'eval'), file_name_suffix='.gz',
                    coder=example_proto_coder.ExampleProtoCoder(
                        transformed_metadata.schema)))

            # Save transformation function to disk for use at serving time
            (transform_fn |
                'WriteTransformFn' >> transform_fn_io.WriteTransformFn(
                    os.path.join(OUTPUT_DIR, 'metadata')))

# Change to True to run locally
preprocess(in_test_mode=False)
```

```
Launching Dataflow job preprocess-taxi-features-200413-132314 ... ha
ng on
WARNING:tensorflow:From <ipython-input-10-609e78ab05aa>:124: ColumnS
chema (from tensorflow_transform.tf_metadata.dataset_schema) is depr
ecated and will be removed in a future version.
Instructions for updating:
ColumnSchema is a deprecated, use from_feature_spec to create a `Sch
ema`
WARNING:tensorflow:From <ipython-input-10-609e78ab05aa>:141: Schema
(from tensorflow_transform.tf_metadata.dataset_schema) is deprecated
and will be removed in a future version.
Instructions for updating:
Schema is a deprecated, use schema_utils.schema_from_feature_spec to
create a `Schema`
{'dayofweek': <tf.Tensor 'inputs/inputs/dayofweek_copy:0' shape=(?,)
dtype=string>, 'dropofflat': <tf.Tensor 'inputs/inputs/dropofflat_co
py:0' shape=(?,) dtype=float32>, 'dropofflon': <tf.Tensor 'inputs/in
puts/dropofflon_copy:0' shape=(?,) dtype=float32>, 'fare_amount': <t
f.Tensor 'inputs/inputs/F_fare_amount_copy:0' shape=(?,) dtype=float
32>, 'hourofday': <tf.Tensor 'inputs/inputs/hourofday_copy:0' shape=
(?,) dtype=int64>, 'key': <tf.Tensor 'inputs/inputs/key_copy:0' shap
e=(?,) dtype=string>, 'passengers': <tf.Tensor 'inputs/inputs/passen
gers_copy:0' shape=(?,) dtype=int64>, 'pickuplat': <tf.Tensor 'input
s/inputs/pickuplat_copy:0' shape=(?,) dtype=float32>, 'pickuplon': <
tf.Tensor 'inputs/inputs/pickuplon_copy:0' shape=(?,) dtype=float32
>}
WARNING:tensorflow:From <ipython-input-10-609e78ab05aa>:50: string_t
o_int (from tensorflow_transform.mappers) is deprecated and will be
removed in a future version.
Instructions for updating:
Use `tft.compute_and_apply_vocabulary()` instead.

WARNING:tensorflow:From <ipython-input-10-609e78ab05aa>:50: string_t
o_int (from tensorflow_transform.mappers) is deprecated and will be
removed in a future version.
Instructions for updating:
Use `tft.compute_and_apply_vocabulary()` instead.

WARNING:tensorflow:From /home/jupyter/.local/lib/python3.7/site-pack
ages/tensorflow_transform/tf_utils.py:678: where (from tensorflow.py
thon.ops.array_ops) is deprecated and will be removed in a future ve
rsion.
Instructions for updating:
Use tf.where in 2.0, which has the same broadcast rule as np.where

WARNING:tensorflow:From /home/jupyter/.local/lib/python3.7/site-pack
ages/tensorflow_transform/tf_utils.py:678: where (from tensorflow.py
thon.ops.array_ops) is deprecated and will be removed in a future ve
rsion.
Instructions for updating:
Use tf.where in 2.0, which has the same broadcast rule as np.where

WARNING:tensorflow:From /opt/conda/lib/python3.7/site-packages/tenso
rflow_core/python/saved_model/signature_def_utils_impl.py:201: build
_tensor_info (from tensorflow.python.saved_model.utils_impl) is depr
ecated and will be removed in a future version.
Instructions for updating:
This function will only be available through the v1 compatibility li
brary as tf.compat.v1.saved_model.utils.build_tensor_info or tf.comp
at.v1.saved_model.build_tensor_info.
```

```
WARNING:tensorflow:From /opt/conda/lib/python3.7/site-packages/tenso
rflow_core/python/saved_model/signature_def_utils_impl.py:201: build
_tensor_info (from tensorflow.python.saved_model.utils_impl) is depr
ecated and will be removed in a future version.
Instructions for updating:
This function will only be available through the v1 compatibility li
brary as tf.compat.v1.saved_model.utils.build_tensor_info or tf.comp
at.v1.saved_model.build_tensor_info.

INFO:tensorflow:Assets added to graph.

INFO:tensorflow:Assets added to graph.

INFO:tensorflow:No assets to write.

INFO:tensorflow:No assets to write.

INFO:tensorflow:SavedModel written to: gs://qwiklabs-gcp-03-b02dedbd
6a51/taxifare/preproc_tft/tmp/tftransform_tmp/72115c031a9646d8a6d62a
10b2d147a4/saved_model.pb

INFO:tensorflow:SavedModel written to: gs://qwiklabs-gcp-03-b02dedbd
6a51/taxifare/preproc_tft/tmp/tftransform_tmp/72115c031a9646d8a6d62a
10b2d147a4/saved_model.pb

INFO:tensorflow:Assets added to graph.

INFO:tensorflow:Assets added to graph.

INFO:tensorflow:No assets to write.

INFO:tensorflow:No assets to write.

INFO:tensorflow:SavedModel written to: gs://qwiklabs-gcp-03-b02dedbd
6a51/taxifare/preproc_tft/tmp/tftransform_tmp/c748e641aca74b97b5c02f
35efc2ad29/saved_model.pb

INFO:tensorflow:SavedModel written to: gs://qwiklabs-gcp-03-b02dedbd
6a51/taxifare/preproc_tft/tmp/tftransform_tmp/c748e641aca74b97b5c02f
35efc2ad29/saved_model.pb
```

This will take **10-15 minutes**. You cannot go on in this lab until your DataFlow job has succesfully completed.

You may monitor the progress of the Dataflow job in the GCP console on the Dataflow page.

When you see the Jupyter notebook status has returned to "Idle" you may proceed to the next step.

In [11]:

```
%%bash
# ls preproc_tft
gsutil ls gs://${BUCKET}/taxifare/preproc_tft/
```

```
gs://qwiklabs-gcp-03-b02dedbd6a51/taxifare/preproc_tft/
gs://qwiklabs-gcp-03-b02dedbd6a51/taxifare/preproc_tft/eval-00000-of
-00001.gz
gs://qwiklabs-gcp-03-b02dedbd6a51/taxifare/preproc_tft/train-00000-o
f-00003.gz
gs://qwiklabs-gcp-03-b02dedbd6a51/taxifare/preproc_tft/train-00001-o
f-00003.gz
gs://qwiklabs-gcp-03-b02dedbd6a51/taxifare/preproc_tft/train-00002-o
f-00003.gz
gs://qwiklabs-gcp-03-b02dedbd6a51/taxifare/preproc_tft/metadata/
gs://qwiklabs-gcp-03-b02dedbd6a51/taxifare/preproc_tft/tmp/
```

Train off preprocessed data

Now that we have our data ready and verified it is in the correct location we can train our taxifare model locally.

NOTE: You may ignore any WARNING related to "tensorflow" in any of the outputs that follow from this point.

In [12]:

```
%%bash
rm -r ./taxi_trained
export PYTHONPATH=${PYTHONPATH}:$PWD
python3 -m tft_trainer.task \
    --train_data_path="gs://${BUCKET}/taxifare/preproc_tft/train*" \
    --eval_data_path="gs://${BUCKET}/taxifare/preproc_tft/eval*"  \
    --output_dir=./taxi_trained \
rm: cannot remove './taxi_trained': No such file or directory
INFO:tensorflow:Using default config.
INFO:tensorflow:Using config: {'_model_dir': './taxi_trained', '_tf_
random_seed': None, '_save_summary_steps': 100, '_save_checkpoints_s
teps': None, '_save_checkpoints_secs': 600, '_session_config': allow
_soft_placement: true
graph_options {
  rewrite_options {
    meta_optimizer_iterations: ONE
  }
}
, '_keep_checkpoint_max': 5, '_keep_checkpoint_every_n_hours': 1000
0, '_log_step_count_steps': 100, '_train_distribute': None, '_device
_fn': None, '_protocol': None, '_eval_distribute': None, '_experimen
tal_distribute': None, '_experimental_max_worker_delay_secs': None,
'_session_creation_timeout_secs': 7200, '_service': None, '_cluster_
spec': <tensorflow.python.training.server_lib.ClusterSpec object at
0x7f40d7717fd0>, '_task_type': 'worker', '_task_id': 0, '_global_id_
in_cluster': 0, '_master': '', '_evaluation_master': '', '_is_chie
f': True, '_num_ps_replicas': 0, '_num_worker_replicas': 1}
INFO:tensorflow:Not using Distribute Coordinator.
INFO:tensorflow:Running training and evaluation locally (non-distrib
uted).
INFO:tensorflow:Start train and evaluate loop. The evaluate will hap
pen after every checkpoint. Checkpoint frequency is determined based
on RunConfig arguments: save_checkpoints_steps None or save_checkpoi
nts_secs 600.
WARNING:tensorflow:From /opt/conda/lib/python3.7/site-packages/tenso
rflow_core/python/training/training_util.py:236: Variable.initialize
d_value (from tensorflow.python.ops.variables) is deprecated and wil
l be removed in a future version.
Instructions for updating:
Use Variable.read_value. Variables in 2.X are initialized automatica
lly both in eager and graph (inside tf.defun) contexts.
INFO:tensorflow:Calling model_fn.
WARNING:tensorflow:From /opt/conda/lib/python3.7/site-packages/tenso
rflow_core/python/ops/resource_variable_ops.py:1630: calling BaseRes
ourceVariable.__init__ (from tensorflow.python.ops.resource_variable
_ops) with constraint is deprecated and will be removed in a future
version.
Instructions for updating:
If using Keras pass *_constraint arguments to layers.
WARNING:tensorflow:From /opt/conda/lib/python3.7/site-packages/tenso
rflow_estimator/python/estimator/canned/head.py:437: to_float (from
tensorflow.python.ops.math_ops) is deprecated and will be removed in
a future version.
Instructions for updating:
Use `tf.cast` instead.
WARNING:tensorflow:From /opt/conda/lib/python3.7/site-packages/tenso
rflow_core/python/training/adagrad.py:76: calling Constant.__init__
(from tensorflow.python.ops.init_ops) with dtype is deprecated and w
ill be removed in a future version.
Instructions for updating:
Call initializer instance with the dtype argument instead of passing
it to the constructor
INFO:tensorflow:Done calling model_fn.
INFO:tensorflow:Create CheckpointSaverHook.
WARNING:tensorflow:From /opt/conda/lib/python3.7/site-packages/tenso
rflow_core/python/ops/array_ops.py:1475: where (from tensorflow.pyth
on.ops.array_ops) is deprecated and will be removed in a future vers
ion.
```

```
Instructions for updating:
Use tf.where in 2.0, which has the same broadcast rule as np.where
INFO:tensorflow:Graph was finalized.
2020-04-13 13:35:59.962920: I tensorflow/core/platform/profile_util
s/cpu_utils.cc:94] CPU Frequency: 2200000000 Hz
2020-04-13 13:35:59.963305: I tensorflow/compiler/xla/service/servic
e.cc:168] XLA service 0x5612b792e580 initialized for platform Host
(this does not guarantee that XLA will be used). Devices:
2020-04-13 13:35:59.963344: I tensorflow/compiler/xla/service/servic
e.cc:176]   StreamExecutor device (0): Host, Default Version
2020-04-13 13:35:59.963592: I tensorflow/core/common_runtime/process
_util.cc:136] Creating new thread pool with default inter op settin
g: 2. Tune using inter_op_parallelism_threads for best performance.
INFO:tensorflow:Running local_init_op.
INFO:tensorflow:Done running local_init_op.
INFO:tensorflow:Saving checkpoints for 0 into ./taxi_trained/model.c
kpt.
INFO:tensorflow:loss = 1861.8728, step = 1
INFO:tensorflow:global_step/sec: 129.392
INFO:tensorflow:loss = 568.9155, step = 101 (0.773 sec)
INFO:tensorflow:global_step/sec: 241.602
INFO:tensorflow:loss = 102.84067, step = 201 (0.414 sec)
INFO:tensorflow:Saving checkpoints for 300 into ./taxi_trained/mode
l.ckpt.
INFO:tensorflow:Calling model_fn.
INFO:tensorflow:Done calling model_fn.
INFO:tensorflow:Starting evaluation at 2020-04-13T13:36:04Z
INFO:tensorflow:Graph was finalized.
INFO:tensorflow:Restoring parameters from ./taxi_trained/model.ckpt-
300
INFO:tensorflow:Running local_init_op.
INFO:tensorflow:Done running local_init_op.
INFO:tensorflow:Evaluation [5/50]
INFO:tensorflow:Evaluation [10/50]
INFO:tensorflow:Evaluation [15/50]
INFO:tensorflow:Evaluation [20/50]
INFO:tensorflow:Evaluation [25/50]
INFO:tensorflow:Evaluation [30/50]
INFO:tensorflow:Evaluation [35/50]
INFO:tensorflow:Evaluation [40/50]
INFO:tensorflow:Evaluation [45/50]
INFO:tensorflow:Evaluation [50/50]
INFO:tensorflow:Finished evaluation at 2020-04-13-13:36:05
INFO:tensorflow:Saving dict for global step 300: average_loss = 19.0
52462, global_step = 300, label/mean = 5.318125, loss = 304.8394, pr
ediction/mean = 1.0794666
INFO:tensorflow:Saving 'checkpoint_path' summary for global step 30
0: ./taxi_trained/model.ckpt-300
INFO:tensorflow:Calling model_fn.
INFO:tensorflow:Done calling model_fn.
WARNING:tensorflow:From /opt/conda/lib/python3.7/site-packages/tenso
rflow_core/python/saved_model/signature_def_utils_impl.py:201: build
_tensor_info (from tensorflow.python.saved_model.utils_impl) is depr
ecated and will be removed in a future version.
Instructions for updating:
This function will only be available through the v1 compatibility li
brary as tf.compat.v1.saved_model.utils.build_tensor_info or tf.comp
at.v1.saved_model.build_tensor_info.
INFO:tensorflow:Signatures INCLUDED in export for Classify: None
INFO:tensorflow:Signatures INCLUDED in export for Regress: None
INFO:tensorflow:Signatures INCLUDED in export for Predict: ['predic
```

```
t']
INFO:tensorflow:Signatures INCLUDED in export for Train: None
INFO:tensorflow:Signatures INCLUDED in export for Eval: None
INFO:tensorflow:Signatures EXCLUDED from export because they cannot
be be served via TensorFlow Serving APIs:
INFO:tensorflow:'serving_default' : Regression input must be a singl
e string Tensor; got {'dayofweek': <tf.Tensor 'dayofweek:0' shape=
(?,) dtype=int64>, 'hourofday': <tf.Tensor 'hourofday:0' shape=(?,)
dtype=int64>, 'pickuplon': <tf.Tensor 'pickuplon:0' shape=(?,) dtype
=float32>, 'pickuplat': <tf.Tensor 'pickuplat:0' shape=(?,) dtype=fl
oat32>, 'dropofflon': <tf.Tensor 'dropofflon:0' shape=(?,) dtype=flo
at32>, 'dropofflat': <tf.Tensor 'dropofflat:0' shape=(?,) dtype=floa
t32>, 'passengers': <tf.Tensor 'passengers:0' shape=(?,) dtype=float
32>, 'londiff': <tf.Tensor 'sub:0' shape=(?,) dtype=float32>, 'latdi
ff': <tf.Tensor 'sub_1:0' shape=(?,) dtype=float32>, 'euclidean': <t
f.Tensor 'Sqrt:0' shape=(?,) dtype=float32>}
INFO:tensorflow:'regression' : Regression input must be a single str
ing Tensor; got {'dayofweek': <tf.Tensor 'dayofweek:0' shape=(?,) dt
ype=int64>, 'hourofday': <tf.Tensor 'hourofday:0' shape=(?,) dtype=i
nt64>, 'pickuplon': <tf.Tensor 'pickuplon:0' shape=(?,) dtype=float3
2>, 'pickuplat': <tf.Tensor 'pickuplat:0' shape=(?,) dtype=float32>,
'dropofflon': <tf.Tensor 'dropofflon:0' shape=(?,) dtype=float32>,
'dropofflat': <tf.Tensor 'dropofflat:0' shape=(?,) dtype=float32>,
'passengers': <tf.Tensor 'passengers:0' shape=(?,) dtype=float32>,
'londiff': <tf.Tensor 'sub:0' shape=(?,) dtype=float32>, 'latdiff':
<tf.Tensor 'sub_1:0' shape=(?,) dtype=float32>, 'euclidean': <tf.Ten
sor 'Sqrt:0' shape=(?,) dtype=float32>}
WARNING:tensorflow:Export includes no default signature!
INFO:tensorflow:Restoring parameters from ./taxi_trained/model.ckpt-
300
INFO:tensorflow:Assets added to graph.
INFO:tensorflow:No assets to write.
INFO:tensorflow:SavedModel written to: ./taxi_trained/export/exporte
r/temp-b'1586784965'/saved_model.pb
INFO:tensorflow:Loss for final step: 4.64241.
```

In [13]:

```
!ls $PWD/taxi_trained/export/exporter
```

```
1586784965
```

Now let's create fake data in JSON format and use it to serve a prediction with gcloud ai-platform local predict

In [14]:

```
%%writefile /tmp/test.json
{"dayofweek":0, "hourofday":17, "pickuplon": -73.885262, "pickuplat": 40.773008,
"dropofflon": -73.987232, "dropofflat": 40.732403, "passengers": 2.0}
```

```
Writing /tmp/test.json
```

In [15]:

```
%%bash
sudo find "/usr/lib/google-cloud-sdk/lib/googlecloudsdk/command_lib/ml_engine" -
name '*.pyc' -delete
```

In [16]:

```
%%bash
model_dir=$(ls $PWD/taxi_trained/export/exporter/)
gcloud ai-platform local predict \
    --model-dir=./taxi_trained/export/exporter/${model_dir} \
    --json-instances=/tmp/test.json
```

```
PREDICTIONS
[11.677545547485352]
```

```
If the signature defined in the model is not serving_default then yo
u must specify it via --signature-name flag, otherwise the command m
ay fail.
WARNING: WARNING:tensorflow:From /usr/lib/google-cloud-sdk/lib/third
_party/ml_sdk/cloud/ml/prediction/frameworks/tf_prediction_lib.py:4
8: The name tf.saved_model.tag_constants.SERVING is deprecated. Plea
se use tf.saved_model.SERVING instead.

WARNING:tensorflow:From /usr/lib/google-cloud-sdk/lib/third_party/ml
_sdk/cloud/ml/prediction/frameworks/tf_prediction_lib.py:50: The nam
e tf.saved_model.signature_constants.DEFAULT_SERVING_SIGNATURE_DEF_K
EY is deprecated. Please use tf.saved_model.DEFAULT_SERVING_SIGNATUR
E_DEF_KEY instead.

WARNING:tensorflow:
The TensorFlow contrib module will not be included in TensorFlow 2.
0.
For more information, please see:
  * https://github.com/tensorflow/community/blob/master/rfcs/2018090
7-contrib-sunset.md
  * https://github.com/tensorflow/addons
  * https://github.com/tensorflow/io (for I/O related ops)
If you depend on functionality not listed there, please file an issu
e.

WARNING:tensorflow:From /usr/lib/google-cloud-sdk/lib/third_party/ml
_sdk/cloud/ml/prediction/frameworks/tf_prediction_lib.py:607: The na
me tf.gfile.IsDirectory is deprecated. Please use tf.io.gfile.isdir
instead.

WARNING:tensorflow:From /usr/lib/google-cloud-sdk/lib/third_party/ml
_sdk/cloud/ml/prediction/frameworks/tf_prediction_lib.py:224: The na
me tf.saved_model.loader.maybe_saved_model_directory is deprecated.
Please use tf.compat.v1.saved_model.loader.maybe_saved_model_directo
ry instead.

WARNING:tensorflow:From /usr/lib/google-cloud-sdk/lib/third_party/ml
_sdk/cloud/ml/prediction/frameworks/tf_prediction_lib.py:231: The na
me tf.Session is deprecated. Please use tf.compat.v1.Session instea
d.

WARNING:tensorflow:From /usr/lib/google-cloud-sdk/lib/third_party/ml
_sdk/cloud/ml/prediction/frameworks/tf_prediction_lib.py:231: The na
me tf.Session is deprecated. Please use tf.compat.v1.Session instea
d.

2020-04-13 13:36:10.913127: I tensorflow/core/platform/profile_util
s/cpu_utils.cc:94] CPU Frequency: 2200000000 Hz
2020-04-13 13:36:10.913550: I tensorflow/compiler/xla/service/servic
e.cc:168] XLA service 0x55ab99d1fdb0 initialized for platform Host
(this does not guarantee that XLA will be used). Devices:
2020-04-13 13:36:10.913625: I tensorflow/compiler/xla/service/servic
e.cc:176]   StreamExecutor device (0): Host, Default Version
2020-04-13 13:36:10.913737: I tensorflow/core/common_runtime/process
_util.cc:136] Creating new thread pool with default inter op settin
g: 2. Tune using inter_op_parallelism_threads for best performance.
WARNING:tensorflow:From /usr/lib/google-cloud-sdk/lib/third_party/ml
_sdk/cloud/ml/prediction/frameworks/tf_prediction_lib.py:233: load
(from tensorflow.python.saved_model.loader_impl) is deprecated and w
ill be removed in a future version.
Instructions for updating:
```

```
This function will only be available through the v1 compatibility li
brary as tf.compat.v1.saved_model.loader.load or tf.compat.v1.saved_
model.load. There will be a new function for importing SavedModels i
n Tensorflow 2.0.
WARNING:tensorflow:From /usr/lib/google-cloud-sdk/lib/third_party/ml
_sdk/cloud/ml/prediction/frameworks/tf_prediction_lib.py:233: load
(from tensorflow.python.saved_model.loader_impl) is deprecated and w
ill be removed in a future version.
Instructions for updating:
This function will only be available through the v1 compatibility li
brary as tf.compat.v1.saved_model.loader.load or tf.compat.v1.saved_
model.load. There will be a new function for importing SavedModels i
n Tensorflow 2.0.
```