Serving ML Predictions in batch and real-time

Learning Objectives

1. Copy trained model into your bucket
2. Deploy AI Platform trained model

Introduction

In this notebook, we will create a prediction service that calls your trained model deployed in Cloud to serve predictions.

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.

Copy trained model

Set necessary variables

In [1]:
```
PROJECT = "qwiklabs-gcp-01-6b5616cd9dc9"  # Replace with your PROJECT
BUCKET = PROJECT
REGION = "us-central1"            # Choose an available region for Cloud MLE
TFVERSION = "2.6"                 # TF version for CMLE to use
```

In [2]:
```
import os
os.environ["BUCKET"] = BUCKET
os.environ["PROJECT"] = PROJECT
os.environ["REGION"] = REGION
os.environ["TFVERSION"] = TFVERSION
```

Create a bucket and copy trained model in it

In [4]:
```
%%bash
if ! gsutil ls -r gs://${BUCKET} | grep -q gs://${BUCKET}/babyweight/trained_mod
    gsutil mb -l ${REGION} gs://${BUCKET}
    # copy canonical model if you didn't do previous notebook
    # TODO: Your code goes here
    gsutil -m cp -R gs://cloud-training-demos/babyweight/trained_model gs://${BU

fi
```

```
Creating gs://qwiklabs-gcp-01-6b5616cd9dc9/...
ServiceException: 409 A Cloud Storage bucket named 'qwiklabs-gcp-01-6b5616cd9dc
9' already exists. Try another name. Bucket names must be globally unique across
all Google Cloud projects, including those outside of your organization.
Copying gs://cloud-training-demos/babyweight/trained_model/export/exporter/15293
55466/variables/variables.index...
Copying gs://cloud-training-demos/babyweight/trained_model/checkpoint...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-342784.ind
```

```
ex...
Copying gs://cloud-training-demos/babyweight/trained_model/eval/events.out.tfeve
nts.1529348264.cmle-training-master-a137ac0fff-0-9q8r4...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-376661.ind
ex...
Copying gs://cloud-training-demos/babyweight/trained_model/events.out.tfevents.1
529347276.cmle-training-master-a137ac0fff-0-9q8r4...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-342784.met
a...
Copying gs://cloud-training-demos/babyweight/trained_model/export/exporter/15293
55466/saved_model.pb...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-376661.dat
a-00000-of-00003...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-376661.met
a...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-390628.dat
a-00000-of-00003...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-342784.dat
a-00001-of-00003...
Copying gs://cloud-training-demos/babyweight/trained_model/graph.pbtxt...
Copying gs://cloud-training-demos/babyweight/trained_model/export/exporter/15293
55466/variables/variables.data-00000-of-00001...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-390628.dat
a-00001-of-00003...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-342784.dat
a-00000-of-00003...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-390628.dat
a-00002-of-00003...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-342784.dat
a-00002-of-00003...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-376661.dat
a-00001-of-00003...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-376661.dat
a-00002-of-00003...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-390628.ind
ex...
Copying gs://cloud-training-demos/babyweight/trained_model/model.ckpt-390628.met
a...
- [22/22 files][  6.5 MiB/  6.5 MiB] 100% Done
Operation completed over 22 objects/6.5 MiB.
```

Deploy trained model

We'll now deploy our model. This will take a few minutes. Once the cell below completes, you should be able to see your newly deployed model in the 'Models' portion of the AI Platform section of the GCP console.

In [ ]:

```
%%bash
# Set necessary variables:
MODEL_NAME="babyweight"
MODEL_VERSION="ml_on_gcp"
MODEL_LOCATION=$(gsutil ls gs://${BUCKET}/babyweight/export/exporter/ | tail -1)

# Set the region to global by executing the following command:
gcloud config set ai_platform/region global

echo "Deploying the model '$MODEL_NAME', version '$MODEL_VERSION' from $MODEL_LO
echo "... this will take a few minutes"
```

```
# Deploy trained model:
gcloud ai-platform models create ${MODEL_NAME} --regions $REGION
# Create a new AI Platform version.
gcloud ai-platform versions create ${MODEL_VERSION} \
  --model ${MODEL_NAME} \
  --origin ${MODEL_LOCATION} \
  --runtime-version $TFVERSION
```