CHAPTER 41

Google Cloud Machine Learning Engine (Cloud MLE)

The Google Cloud Machine Learning Engine, simply known as Cloud MLE, is a managed Google infrastructure for training and serving "large-scale" machine learning models. Cloud ML Engine is a part of GCP AI Platform. This managed infrastructure can train large-scale machine learning models built with TensorFlow, Keras, Scikit-learn, or XGBoost. It also provides modes of serving or consuming the trained models either as an online or batch prediction service. Using online prediction, the infrastructure scales in response to request throughout, while with the batch mode, Cloud MLE can provide inference for TBs of data.

Two important features of Cloud MLE is the ability to perform distribution training and automatic hyper-parameter tuning of your models while training. The big advantage of automatic hyper-parameter tuning is the ability to find the best set of parameters that minimize the model cost or loss function. This saves time of development hours in iterative experiments.

The Cloud MLE Train/Deploy Process

The high-level overview of the train/deploy process on Cloud MLE is depicted in Figure 41-1:

1. The data for training/inference is kept on GCS.
2. The execution script uses the application logic to train the model on Cloud MLE using the training data.

E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_41

3. The trained model is stored on GCS.
4. A prediction service is created on Cloud MLE using the trained model.
5. The external application sends data to the deployed model for inference.

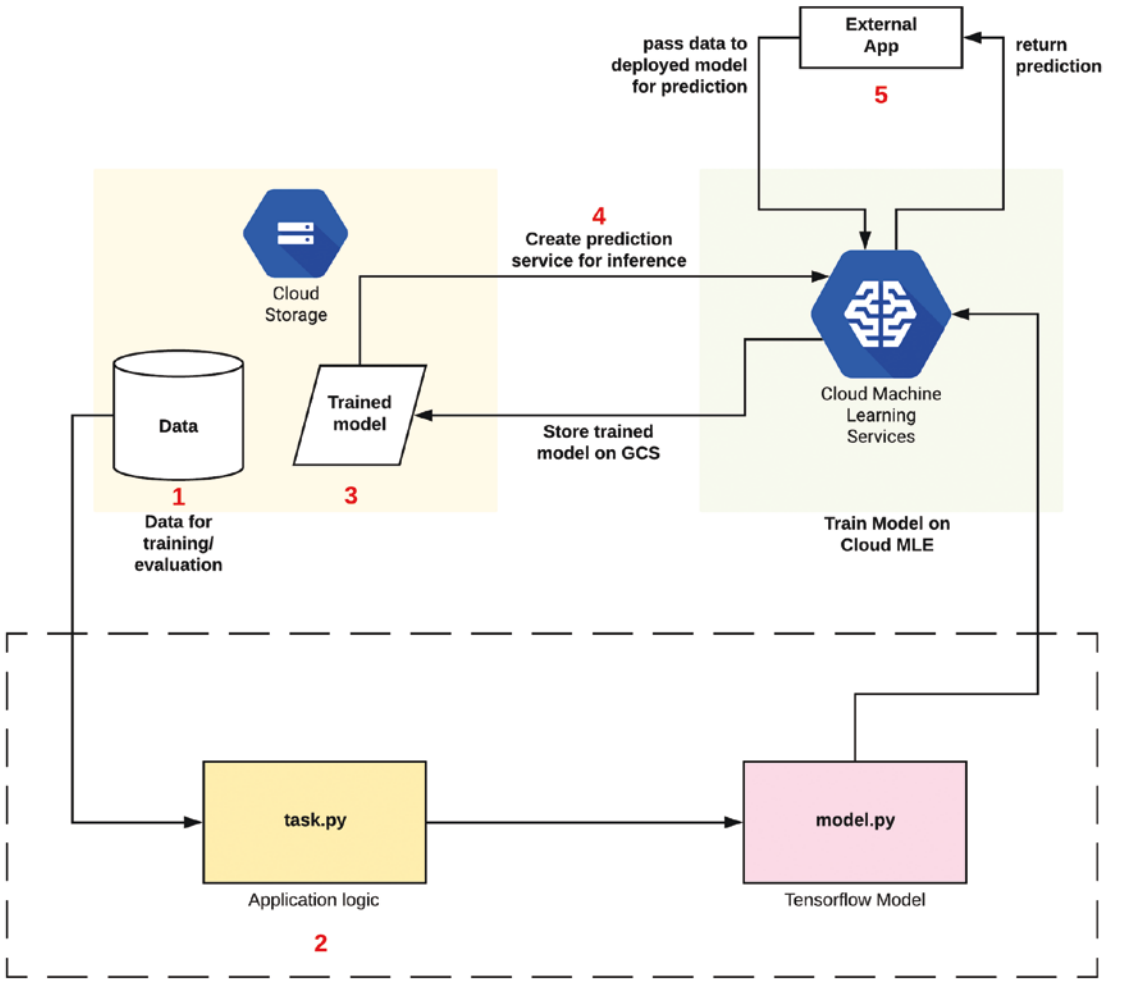

Figure 41-1. *The train/deploy process on Cloud MLE*

Preparing for Training and Serving on Cloud MLE

In this contrived example, we'll use the famous Iris dataset to train and serve a TensorFlow model using the Estimator API on Cloud MLE. To begin, let's walk through the following steps:

1. Create a bucket on GCS by running the gsutil mb command on the cloud terminal. Replace it with unique bucket name.

   ```
   export bucket_name=iris-dataset'
   gsutil mb gs://$bucket_name
   ```

2. Transfer training and test data from the code repository to the GCP bucket.
3. Move the train data.

   ```
   gsutil cp train_data.csv gs://$bucket_name
   ```

4. Move the train data.

   ```
   gsutil cp test_data.csv gs://$bucket_name
   ```

5. Move the hold-out data for batch predictions.

   ```
   gsutil cp hold_out_test.csv gs://$bucket_name
   ```

6. Enable the Cloud Machine Learning API to be able to create and use machine learning models on GCP Cloud MLE:
   a. Go to APIs & Services.
   b. Click "Enable APIs & Services".
   c. Search for "Cloud Machine Learning Engine".
   d. Click ENABLE API as shown in Figure 41-2.