README.md

Customer churn rate prediction using ML in serverless spark

Overview

With the advent of cloud environments, the concept of huge capital investments in infrastructure in terms of capital and maintenance is a thing of the past. Even when it comes to provisioning infrastructure on cloud services, it can get tedious and cumbersome.

In this example, you will look at executing a simple PySpark code which runs on Serverless batch (a fully managed Dataproc cluster). It is similar to executing code on a Dataproc cluster without the need to initialize, deploy or manage the underlying infrastructure.

This usecase is used to check customer churn rate prediction using ML in Serverless spark.

Services Used

- Google Cloud Storage
- Google Cloud Dataproc
- Google Cloud Bigquery
- Google Cloud Composer

3. Permissions / IAM Roles required to run the lab

Following permissions / roles are required to execute the serverless batch

- Viewer
- Dataproc Editor
- BigQuery Data Editor
- Service Account User
- Storage Admin
- Environment User and Storage Object Viewer

4. Checklist

To perform the lab, below are the list of activities to perform.

1. GCP Prerequisites
2. Spark History Server Setup
3. Uploading scripts and datasets to GCP
4. Creating a Composer Environment
5. Creating a BigQuery Dataset

Note down the values for below variables to get started with the lab:

```
PROJECT_ID=                                         #Current GCP project where we are building our use case
REGION=                                             #GCP region where all our resources will be created
SUBNET=                                             #subnet which has private google access enabled
BQ_DATASET_NAME=                                    #BigQuery dataset where all the tables will be stored
BUCKET_CODE=                                        #GCP bucket where our code, data and model files will be stored
BUCKET_PHS=                                         #bucket where our application logs created in the history server will be stored
HISTORY_SERVER_NAME=                                #name of the history server which will store our application logs
UMSA=                                               #user managed service account required for the PySpark job executions
SERVICE_ACCOUNT=$UMSA@$PROJECT_ID.iam.gserviceaccount.com
NAME=<your_name_here>                               #Your Unique Identifier
```

5. Lab Modules

The lab consists of the following modules.

1. Understand the Data
2. Solution Architecture
3. Data Preparation
4. Model Training and Evaluation
5. Examine the logs
6. Explore the output

There are 3 ways of perforing the lab.

- Using Google Cloud Shell
- Using GCP console
- Using Cloud Composer

Please chose one of the methods to execute the lab.

6. CleanUp

Delete the resources after finishing the lab.
Refer - Cleanup