README.md

Retail store analytics

1. 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 deals with the analysis of retail store data.

2. Services Used

- Google Cloud Storage
- Google Cloud Dataproc
- Google Cloud BigQuery

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

4. Checklist

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

1. GCP Prerequisites
2. Spark History Server Setup
3. Creating a GCS Bucket
4. Creating a BigQuery Dataset
5. Metastore Creation

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_NAME=                                      #user managed service account required for the PySpark job executions
SERVICE_ACCOUNT=$UMSA_NAME@$PROJECT_ID.iam.gserviceaccount.com
NAME=<your_name_here>                           #Your Unique Identifier
```

Note: The region to submit serverless spark job, VPC Subnet and Staging bucket should be same.

5. Lab Modules

The lab consists of the following modules.

- Understand the Data
- Solution Architecture
- Executing ETL
- Examine the logs
- Explore the output

There is 1 way to perform the lab - Using GCP sessions through Big Query

6. CleanUp

Delete the resources after finishing the lab.

Refer - Cleanup