Cloud deployment 1 - Xi Zhang

Starbucks nutrition analysis

Repository : A screencast of the deployed running cluster.

Introduction

This report outlines the process followed for setting up a Google DataProc environment for a data analysis project concerning Starbucks nutrition. The aim is to process and analyse raw data related to Starbucks nutrition, leveraging Google DataProc's capabilities.

Steps

1. Setup

The first step involved enabling essential libraries within the Google Cloud environment. These included Cloud DataProc, Compute Engine, and Cloud Storage. This setup ensures that all required services are available for the data processing tasks.

2. Storage Bucket and Cluster Creation

I created a Google Cloud Storage bucket named “xi-assignment1”, the region is “us-central1”. This bucket serves as the primary storage location for both the raw data and processed outputs relevant to the project.

I also created a cluster named “cluster-assignment1” that was for managing and processing the data. The cluster was in the same region as the bucket. I chose series “N1” for both the management node and the work node. I set the Machine type for management node and the work node were “n1-standard-2” (2 vCPU,1 core, 7.5GB memory)

3. Data Upload and Path Configuration

I uploaded the raw data and queries to the “xi-assignment” bucket.

Modified the paths within the queries to ensure correct file loading and output saving in the designated “Outputs” folder.

4. Data Processing and Cleaning

My first step was to run a Pig Query (procressing&cleaning.pig) to process and clean the raw data. Upon submission and completion of the job, a file containing cleaned data was generated. This file includes data that has been pre-processed and is ready for further analysis.

5. Table Creation, Hive Queries and Pig Queries

Next, I ran the query create_tables.sql to prepare for the Hive query that follows.
Then I ran Hive simple queries and Hive complex queries.. The results of these queries are stored in the "hive_outputs" subfolder in the "output" folder. We can download them to see the results it generates.
I used the same method to run the pig queries and then got a Pig_outputs file containing all of the outputs generated.