Assignment - Big Data Architecture on the Cloud Platform

Shaun Pritchard

Rasmussen College

QMB4200

Terry Brookhouser

October 15, 2020

**Assignment - Big Data Architecture on the Cloud Platform**

**Architecture diagram:**

Diagram

Description automatically generated

**Summary:**

Essentially, we are not giving the type of data that needs to be transformed into parquet format. So, I will assume it is 200 terabytes of either CVS or SQL data just to keep it simple. Parquet is an open source file format available to any project in the Hadoop ecosystem.

So, we would need to run an ELT process implanting cloud services such as Azure or AWS,S3, and Apache Spark on EC2 virtual machines to host, table, sort, and distribute the data. In the Extract Load Transform (ELT) process, you first extract the data, and then you immediately move it into a centralized data repository.

Essentially the process would start by cleaning the data. We do this by calculating the mean values for missing data, removing duplicate data, sorting, and removing data that is error prone.

After this process we would need to load the table data into HDFS create a cloud-based data lake. HDFS is Hadoop system process to load and distribute the data across our cluster computers which is distributed on our cloud service network. After this we would need to process the data with a resource negotiator to implement map reduce algorithm and scripting using Apache Spark to transform the data loaded across the Clusters Network.

This would essentially take minimal time to transform the data from one format to another becuase it would happen in parallel. It would essentially take more time to clean and load the data than transform this data into another file type such as Parquet format once it is loaded and distributed across the cluster.

Through my research I found a script to handle doing this which would require loading the table data from an interface such as PySpark API or Java Database Connectivity (JDBC) API to access and run spark scripts across the cluster to convert the data into a parquet file

**Apache Spark Script to transform data into parquet format:**

from pyspark.sql import SparkSession

spark = SparkSession.builder.getOrCreate()

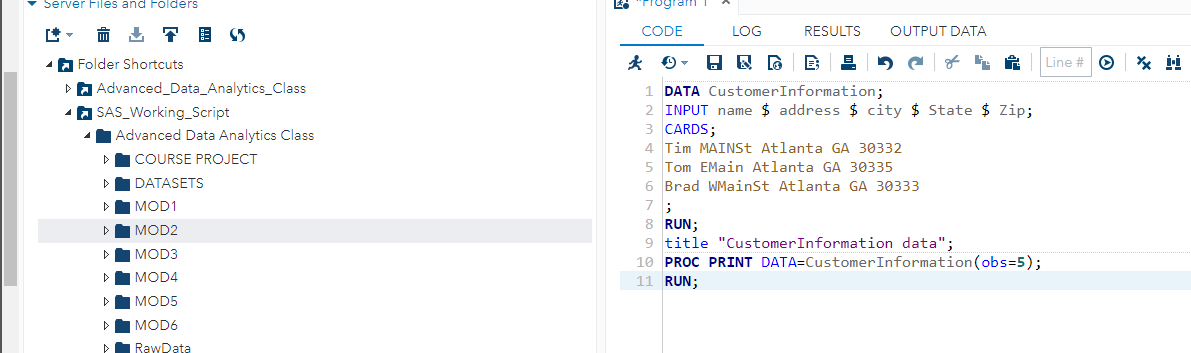
df = spark.read.jdbc("YOUR\_MYSQL\_JDBC\_CONN\_STRING", "YOUR\_TABLE",properties={"user": "YOUR\_USER", "password": "YOUR\_PASSWORD"}) df.write.parquet("YOUR\_HDFS\_FILE")

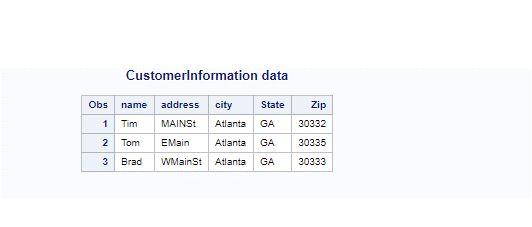
Once the data is loaded into HDFS it can be transformed into another format very quickly *(Apache Spark, 2020).*

**SAS output:**

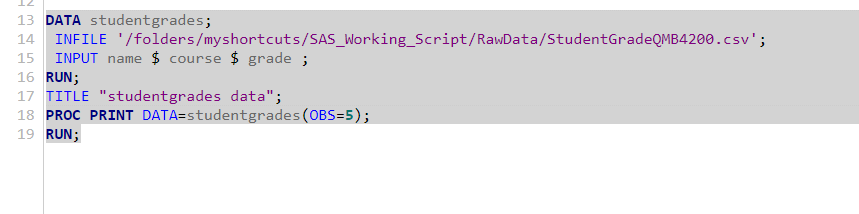
Here are the scripts is created in SAS with completed runs of data and output for both types of input scripting. The first one is the script where I labeled and created data. The second on shows importing a file and running the script on it.

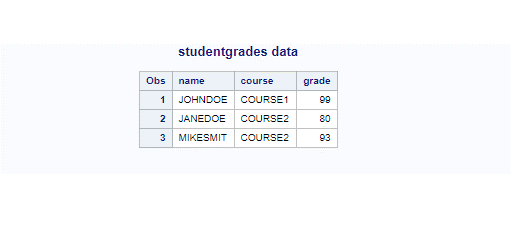
**Script1:**





**Script 2:**





# References

Apache Spark. (2020). *Apache Spark*. Retrieved from https://spark.apache.org: https://spark.apache.org/docs/latest/api/java/index.html