**Week 4 Assignment**

Monroe College

CS 675: Big Data Management and Analytics

Madeline Lin

Professor Syed Farid

October 2nd, 2022

**Part 0 MapReduce Tutorial**

MapReduce is one of the core building blocks of processing in Hadoop framework. Big data is regarded as colossal

load of data that can be hardly processed using the traditional data processing units. The trending social media like Facebook, WhatsApp and YouTube.

This tutorial discusses about MapReduce’s examples, advantages, program explained and use case (i.e. KMeans Algorithm).

MapReduce framework allows us to perform parallel computations on large data sets in a distributed environment without bothering about the issues like reliability, fault tolerance, etc.

The example that this tutorial uses is a word count example. First, they divide the input into three splits and distribute the work among all the map nodes. Then, they tokenize the words in each of the mappers and give a hardcoded value. After mapper phase, a partition process is taken place. Finally, all the output key/value pairs are collected and written in the output file. The advantages of MapReduce are parallel processing and data locality, which is very cost-effective and efficient.

Lastly, the tutorial also demonstrates a use case of KMeans Clustering using Hadoop’s MapReduce.

My feel reading this tutorial is that it gives me an overview of how the MapReduce framework works and I understand better with the real case.

**Part 1 YARN**

Based on the group discussion, we are unable to install Yarn unfortunately. We had issues in our previous assignment of installing Hadoop. Without solving the issues from the second assignment, our group suppose that we could not install Yarn if the Hadoop has not been installed successfully.

The first link (i.e. <https://hadoop.apache.org/docs/r3.3.3/hadoop-yarn/hadoop-yarn-common/yarn-default.xml>) does not make quite sense to me. It is just a huge table without showing much instructions of Yarn configuration.

The second link (i.e. <https://classic.yarnpkg.com/lang/en/docs/getting-started/>) shows us the steps of installing Yarn through the npm package manager. But before that, the Hadoop should be installed. Would you mind letting us know how to solve the Ambari issue from the past assignment?

**Part 2 Ch2 Overview of Big Data Analytics**

Inside the data analytics process, the section talks about data governance, data visualization, and so forth. Data governance ensures the proper collection and protection of the data. Another less well-known role is data steward which specializes in understanding the data to the byte, exactly where it is coming from, all transformations that occur, and what the business really needs from the column or field of data. There are seven Vs of big data, i.e. Variety, Velocity, Volume, Veracity, Variability, Visualization and Value.

Apache Hadoop framework uses HDFS (Hadoop Distributed File System) and depends on MapReduce. MapReduce computes a large amount of data in a Hadoop cluster and uses YARN to schedule the mappers and reducers as tasks, using the containers.

This lecture note also discusses about how to install Hive, Derby, and how to create a database, table in Hive. Then it talks about Select statement, Where clause, Insert statement, primitive types, complex types, built-in operators and functions, language capabilities, and how to retrieve information. Last but not least, it talks about Apache Spark, which is a unified distributed computing engine across different workloads and platforms. Hadoop and Apache Spark are both popular big data frameworks. Finally, it gives an overview of how to use Tableau to perform visualizations.