The Programming Environment Requirement

This assignment aims to practice the MapReduce algorithm, its implementation (by Java or Python, or C++ or other programming languages) and runtime (such as Hadoop or AWS EMR, or Azure HDInsight or MongoDB). Since all the MapReduce runtime runs on a file system, please make sure you understand the file system of the runtime chosen. For Hadoop, HDFS is the file system. For AWS EMR, S3 is the file system. For MongoDB, the underlying database is used to store data. Access to data in these three modes is covered by the lecture slides and recorded videos. For Azure, HDInsight provides HDFS over Azure storage. Please refers to the Azure document https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/apache-hadoop-develop-deploy- java-mapreduce-linux and https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/apache-hadoop-develop-deploy-java-mapreduce-linux.

To get started for this assignment, please make sure you have the following aspects ready:

1. 1) Choose a MapReduce runtime
2. 2) Upload the data to the file system of the runtime
3. 3) Add the SDK library to your IDE to use the packages and libraries of

   MapReduce APIs. The SDK library should be available from the MapReduce runtime provider.

The Dataset

The data set is from a Github project, under the directory of Workload Data.

https://github.com/haniehalipour/Online-Machine-Learning-for-Cloud-Resource- Provisioning-of-Microservice-Backend-Systems

The workload data contains the workload generated from two industrial benchmarks NDBench from Netflix and Dell DVD store from Dell. Both benchmarks are deployed on a cluster of cloud VMs on AWS and Azure clouds. The workload has been split to training sets and testing sets for the machine learning purpose.

In each of the workload file, the first 4 columns contain the following attributes.

CPUUtilization_Average,NetworkIn_Average,NetworkOut_Average,MemoryUtilization _Average

In this assignment, we **only use the Dell DVD datasets**, training and testing.

Technical Requirements

Develop one MapReduce program, given the CPU usage in the step function of 10, such as (0, 10] (11, 20], ... (91,100] for both training and testing data;

1. 1) Output the number of samples in each range;
2. 2) Output Maximum, Minimum, Median and Standard Deviation for the attribute of

   MemoryUtilization_Average in each range;