HW3 Hadoop Mapreduce

Student ID: r09921a10
Name: 鄭翔予

Hadoop

1. Use the images from dockerhub, which is provided by professor
2. Follow the steps on github, and the environment will be setup successfully

MapReduce

1. Check Input format
   Example: 64.242.88.10 - - [07/Mar/2004:16:05:49 -0800] "GET /twiki/bin/edit/Main/Double_bounce_sender?topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12846
   The information we cared about is 07/Mar/2004:16:05:49
2. Mapper.py
   Use the method of wordcount. I split the input line by line and the keys will end up look like 2004-03-07 T 16:00:00.000, and the value will all be 1.
3. Reducer.py
   It's the same as the reducer in wordcount since we want to know how many log at each time, which is exactly our key and the value 1 help us do the math
4. Script file
   Since Hadoop is running by JAVA and we are using python as out programming language, we need to use hadoop-streaming-3.2.1.jar to help us run the code successfully

Discussion

1. If the containers are stopped, we need to restart the containers by order mysql-> hadoop-master -> hadoop-worker -> hadoop-dev, and after entering hadoop-dev the hdfs dfsadmin –report command won't work immediately, it takes some time to wait for the connection.