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 hour, which is exactly our key and the value '1' help us do the math.
4. Script file (run command, included on github)
   Since Hadoop is running by JAVA and we are using python as our programming language, we need to use hadoop-streaming-3.2.1.jar to help us run the code successfully.

JAVA

1. The code was modified from the reference(https://www.informit.com/articles/article.aspx?p=2017061). Since it has almost the same input and output. Thus, in this part I'll explain the code literally.
2. First of all, I need to define my own class of WritableComparable to set up my key for Hadoop. Use SimpleDateFormat to transfer the original date format into "yyyy-MM-dd' T 'HH:mm:ss.SSS", which is the format of our output.
3. Inside this Class, I need to at least implement readFields() to read the input, write() to write to the output and compareTo() to sort the key.
4. After defining the class that help us reformat the input, I need to start define the class mapper since the original input format is 64.242.88.10 - - [07/Mar/2004:16:05:49 -0800]. Use indexOf to find out index of '[', '/', ':', ']' and

split the input to date, month, year, hour and use the object Calendar to help us deal with the problem of month(because its original input is 'Mar' and cannot be recognized in Java). Then set the output to (date, one), where one means the number 1.

5. The Reducer is exactly the same as wordcount because the the input key, value pair is same.
6. Last, need to configure the environment before running the code. Create a job from default configuration, and define input/output path. Then configure the job's name, mapper, reducer ,combiner(in this case, the combiner and reducer is same). After all, configure the output and the job can be run.

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.

Github

https://github.com/stanley101music/Hadoop-MapReduce

Input

http://hpc.ee.ntu.edu.tw/html/IntelligentClouds/webAccessLog/access_log

Reference

https://www.informit.com/articles/article.aspx?p=2017061