Homework4

Hadoop Programming
Due Date: 23:59, December 12, Thursday, 2013

In this homework, we show the example running in Ubuntu LTS 13.10 64bit on personal computer and installing Hadoop through `http://trac.nchc.org.tw/cloud/wiki/Hadoop_Lab1` using Pipes Hadoop framework.

1 Part1:Character count

In part1, you will be given a text file(.txt) and you will have to count the number of each character appears. You will need to implement Character Count based on the WordCount example. Please insert your code in the map() function as shows below.

```
void map( HadoopPipes::MapContext& context ) {
        // insert your code here
}
```

2 Input/Output

2.1 Sample Input

```
I love PPC.
I love NCTU.
```

2.2 Sample Output

```
.   2
C   2
I   2
N   1
P   2
T   1
U   1
e   2
l   2
o   2
v   2
```

3 Part2:K-means Clustering

k-means clustering is a method of vector quantization originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest distance, serving as a prototype of the cluster.

4 Description

In this homework, you will implement a simpler version of K-means Clustering algorithm. The way of implementation is similar to the WordCount. You will need to modify to corresponding map() and reduce() function. Initially, representatives points will be given. Then you will have to compare other points' distance to the given representatives. If the distance is closer to one of the representative (say A), the point belongs to GroupA. Finally, you will output each data point and its group. Notes: There are only two representatives in this homework, A and B. So there will be only GroupA and GroupB.

5 Input/Output

5.1 Input Format

1. There are two lines separated by a comma.
2. Each data is separated by a space.
3. Each column represents the distance to point A, B, C...respectively.

5.2 Sample Input

```
A 0 4 100 50 200,
B 10 0 60 70 90
```

In row 1, meaning that distance A's distance to A is 0, to B 4, to C is 100, to D is 50, to E is 200. Same as in row 2.

5.3 Output Format

1. Each row starts with data point name followed by its group name.

5.4 Sample Output

```
A   GroupA
B   GroupB
C   GroupB
D   GroupA
E   GroupB
```

6 XML config file

Create a .xml config file for each part of this homework. Place it in the home directory of Hadoop. You will need it to run your hadoop program. For example as wordcount.xml

```
<?xml version="1.0"?>
<configuration>
  <property>
    // Set the binary path on DFS
    <name>hadoop.pipes.executable</name>
    <value>$PATH_TO_YOUR_EXECUTABLE_FILE_AT_HDFS</value>
  </property>
  <property>
    <name>hadoop.pipes.java.recordreader</name>
    <value>true</value>
  </property>
  <property>
    <name>hadoop.pipes.java.recordwriter</name>
    <value>true</value>
  </property>
</configuration>
```

7 Creating makefile to compile the program

You will create a makefile to compile your program. And put the executable wordcount file into your HDFS after the compilation(here we take wordcount as an example).

```
CC = g++
# We install hadoop under /opt for example
HADOOP_INSTALL = /opt/hadoop
PLATFORM = Linux-amd64-64
CPPFLAGS = -I $(HADOOP_INSTALL)/c++/$(PLATFORM)/include

wordcount: wordcount.cpp
        $(CC) $(CPPFLAGS) $< -L $(HADOOP_INSTALL)/c++/$(PLATFORM)
        /lib -lnsl -lhadooppipes -lhadooputils -lpthread -g -O2 -o $@
```

8 Run the program

At your hadoop home directory. Follow the instruction below.

```
$ bin/hadoop pipes -conf wordcount.xml -input YOUR_INPUT_PATH
        -output YOUR_OUTPUT_PATH
```

9 Environment

We provide three ways to let you work on your homework.

1. Install Hadoop in your computer. `http://trac.nchc.org.tw/cloud/wiki/Hadoop_Lab1`
2. Using NCHC `http://hadoop.nchc.org.tw/`
3. Using NCTU Openstack. (140.113.98.51. account:studentID, pass:studentID) For 2, 3 we don't guarantee the system or connection will be stable. We suggest you to install Hadoop in your computer.

Also you can use either way of the following Hadoop framework

1. Pipes
2. Streaming

10 Submission

Please submit the following files to e3 system and package them into a directory named studentID-hw4.zip.

1. studentID-charcount.cpp
2. studentID-cluster.cpp
3. makefile
4. description.txt

In description.txt, please provide us your running platform and the Hadoop framework you use.

Your running platform

The Hadoop framework you use

11 Grading Policy

1. If your output is not correct or your program cannot compile or running, you will get 0 in that part of homework.
2. You will get full score in that part of homework if your answer is correct.
3. 40 points for part1.
4. 60 points for part2.

12 Reference

[1]Install Hadoop: http://trac.nchc.org.tw/cloud/wiki/Hadoop_Lab1
[2]Hadoop Pipes: http://cs.smith.edu/dftwiki/index.php/Hadoop_Tutorial_2.2_--_Running_C%2B%2B_Programs_on_Hadoop#Makefile
[3]Hadoop Streaming: http://hadoop.apache.org/docs/r1.1.2/streaming.html
[4]NCHC Hadoop: http://hadoop.nchc.org.tw/