Assignment: Count Words with MapReduce

Motivation: Although you have already run your first Hadoop program with Hello World, we want you to be able to experience the kinds of analysis Hadoop is used for in the real world. We'll do this by having you copy in some code you may not understand just yet, but which gives you a preview of (at least the basics) of what we will learn to do in Hadoop in the next class!

**Steps**

1. Open a Terminal (Rightclick on Desktop or click Terminal icon in the top toolbar). Follow the steps below, copying and pasting or typing the code below to execute a simple word count example.

2. Change Directory to the Hadoop library directory (type): cd /usr/lib/hadoop-mapreduce/

3. Execute the Hadoop jar command to run the WordCount example (type): hadoop jar hadoop-mapreduce-examples.jar wordcount

4. The wordcount example will output the warning that it needs input and output parameters (you should see this):

Usage: wordcount <in> <out>

5. Let's create a couple of text files with a few words in it for testing, or use a log file (type both commands, one at a time):

echo "Hello world in HDFS" > /home/cloudera/testfile1

echo "Hadoop word count example in HDFS" > /home/cloudera/testfile2

6. Create directory for the input files on the HDFS file system (type):

**hdfs dfs -mkdir /user/cloudera/input**

7. Copy the files from local filesystem to the HDFS filesystem (type both commands, one at a time):

**hdfs dfs -put /home/cloudera/testfile1 /user/cloudera/input**

**hdfs dfs -put /home/cloudera/testfile2 /user/cloudera/input**

Run the Program

8. Run the Hadoop WordCount example with the input and output specified (type):

**hadoop jar hadoop-mapreduce-examples.jar wordcount /user/cloudera/input /user/cloudera/output**

9. Hadoop prints out a large amount of logging information. After completion view the output directory (type):

**hdfs dfs -ls /user/cloudera/output**

10. Check the output file to see the results (type):

**hdfs dfs -cat /user/cloudera/output/partr00000**

11. Put the output in a location, in this case your cloudera user's home directory,

hdfs dfs get /user/cloudera/output/partr00000 /home/cloudera/wordcount.txt