Lab: Writing a MapReduce Streaming Program

In this lab you will write a program to calculate average word lengths for letters. You will write this as a streaming program using a scripting language of your choice rather than using Java.

Your virtual machine has Perl, Python, PHP, and Ruby installed, so you can choose any of these—or even shell scripting—to develop a Streaming solution.

For your Hadoop Streaming program you will not use Eclipse. Launch a text editor to write your Mapper script and your Reducer script. Here are some notes about solving the problem in Hadoop Streaming:

1. Input Data

   Use the same input data as in the tutorial from CS246. You can download the data by running the following command:

   ```
   curl http://www.gutenberg.org/cache/epub/100/pg100.txt | \
    perl -pe 's/^\xEF\xBB\xBF//' > pg100.txt
   ```

2. The Mapper Script

   The Mapper will receive lines of text on `stdin`. Find the words in the lines to produce the intermediate output, and emit intermediate (key, value) pairs by writing strings of the form:

   ```
   key <tab> value <newline>
   ```

   These strings should be written to `stdout`.

3. The Reducer Script

   For the reducer, multiple values with the same key are sent to your script on `stdin` as successive lines of input. Each line contains a key, a tab, a value, and a newline. All lines with the same key are sent one after another, possibly followed by lines with a different key, until the reducing input is complete. For example, the reduce script may receive the following:

   ```
   t	3
   t	4
   w	4
   w	6
   ```

   For this input, emit the following to `stdout`:

```
t    3.5
w    5.0
```

Observe that the reducer receives a key with each input line, and must "notice" when the key changes on a subsequent line (or when the input is finished) to know when the values for a given key have been exhausted. This is different than the Java version you worked on in the previous lab.

3. Run the streaming program:

```
$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
 -files pathToMapScript,pathToReduceScript \
 -input inputDir -output outputDir \
 -mapper mapBasename -reducer reduceBasename
```

(Remember, you may need to delete any previous output before running your program by issuing: `hadoop fs -rm -r dataToDelete`.)

4. Review the output in the HDFS directory you specified (`outputDir`).