**Name: Mandar Gadkari**

**Map Reduce Example using Python Code**

In the Cloudera Quickstart VM terminal, follow the instructions below to execute a simple word count example in Python. There will be one file to upload, and quiz questions about word count results with different numbers of reducers.

1. Review the following to create the python code

Section 1: wordcount\_mapper.py

Section 2: wordcount\_reducer.py

The reducer code has some basic parts, see the comments in the code.

Type in the following to open a text editor, and then cut and paste the above lines for wordcount\_mapper.py into the text editor, save, and exit. Repeat for wordcount\_reducer.py

> gedit wordcount\_mapper.py

> gedit wordcount\_reducer.py

Enter the following to see that the indentations line up as above

> more wordcount\_mapper.py

> more wordcount\_reducer.py

Enter the following to make it executable

> chmod +x wordcount\_mapper.py

> chmod +x wordcount\_reducer.py

Enter the following to see what directory you are in

> pwd

It should be /user/cloudera , or something like that.

2. Create some data:

> echo "A long time ago in a galaxy far far away" > /home/cloudera/testfile1

> echo "Another episode of Star Wars" > /home/cloudera/testfile2

3. Create a directory on the HDFS file system (if already exists that’s OK):

hdfs dfs -mkdir /user/cloudera/input

4. Copy the files from local filesystem to the HDFS filesystem:

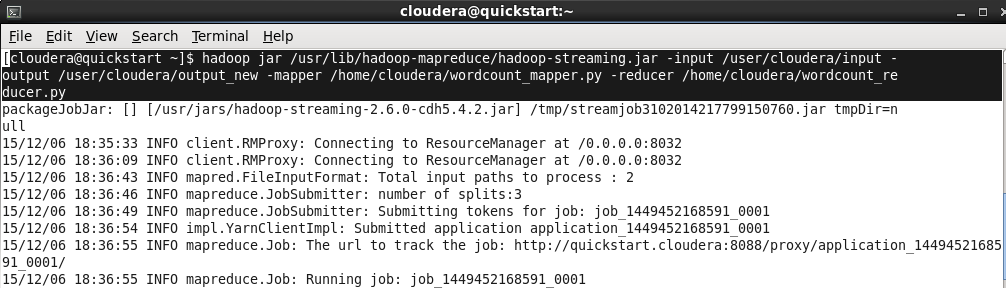
hdfs dfs -put /home/cloudera/testfile1 /user/cloudera/input

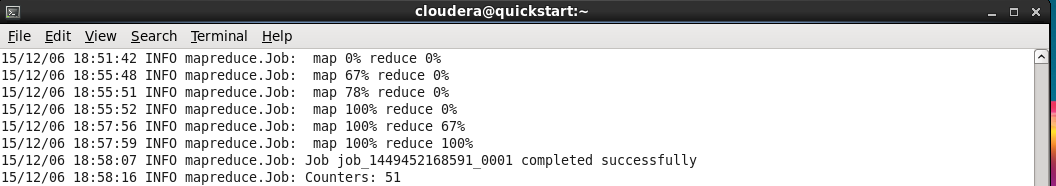
hdfs dfs -put /home/cloudera/testfile2 /user/cloudera/input

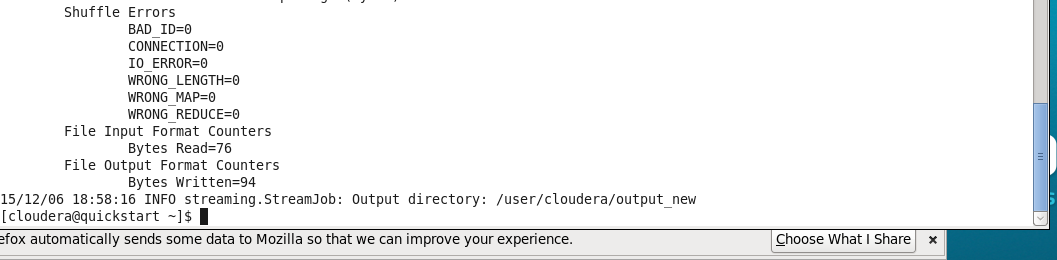
5. You can see your files on HDFS

hdfs dfs -ls /user/cloudera/input

6. Run the Hadoop WordCount example with the input and output specified.







7. Check the output file to see the results:

hdfs dfs -cat /user/cloudera/output\_new/part-r-00000



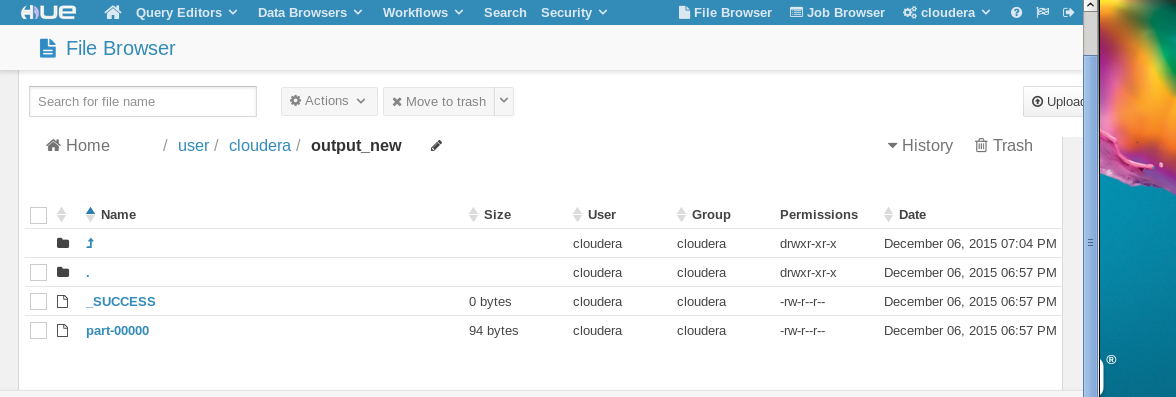
8. View the output directory:

hdfs dfs -ls /user/cloudera/output\_new

Look at the files there and check out the contents, e.g.:

hdfs dfs -cat /user/cloudera/output\_new/part-r-00000

**Checking results in Hue( HDFS) interface:**

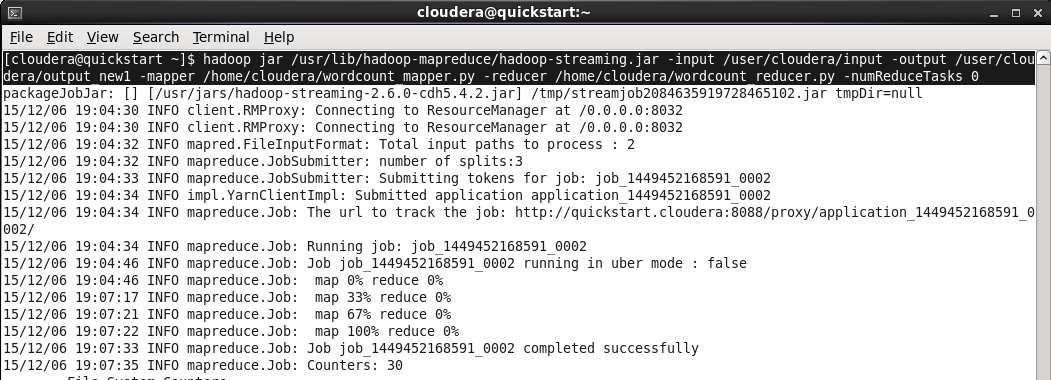


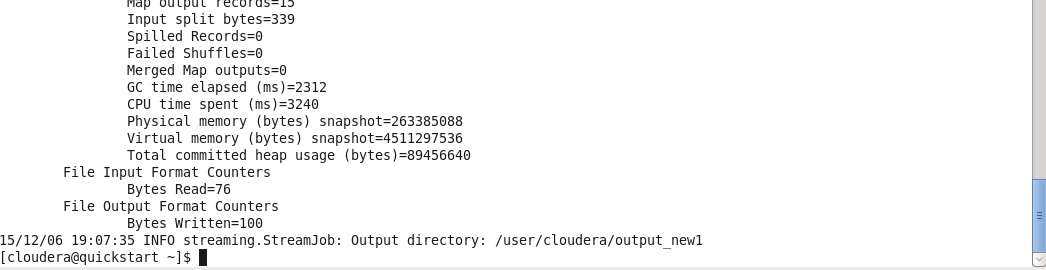
9. Streaming options:

Try: hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar --help

or see hadoop.apache.org/docs/r1.2.1/

Let’s change the number of reduce tasks to see its effects. Setting it to 0 will execute no reducer and only produce the map output. (Note the output directory is changed in the snippet below because Hadoop doesn’t like to overwrite output)

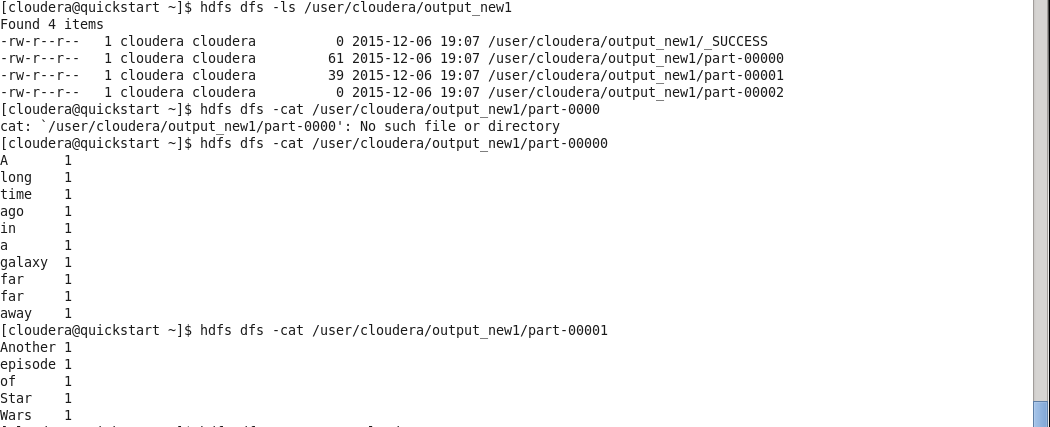




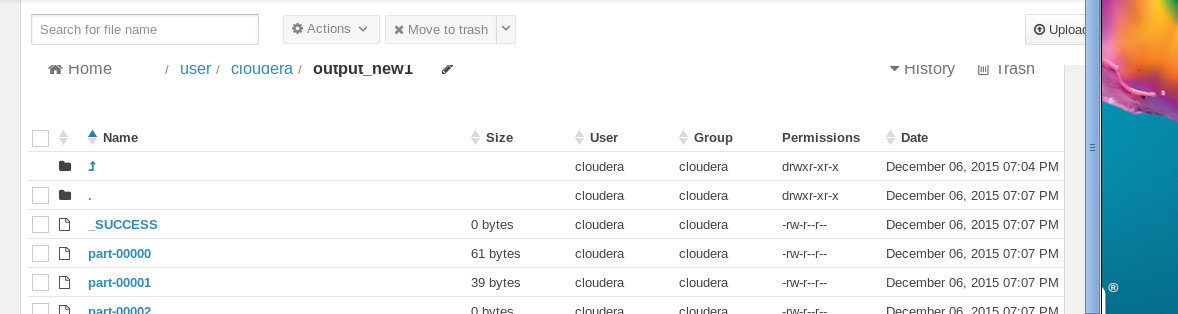
**Check the results:**

hdfs dfs -cat /user/cloudera/output\_new1/part-r-00000

hdfs dfs -cat /user/cloudera/output\_new1/part-r-00001



**Checking results in Hue( HDFS) interface:**

****