1. Create a directory - mapred 180506

[cloudera@quickstart ~]$ pwd

/home/cloudera

[cloudera@quickstart ~]$ mkdir mapred180506

2. Change directory - mapred180506

[cloudera@quickstart ~]$ cd mapred180506

[cloudera@quickstart mapred180506]$

3. Copy all python programs to from windows machine to linux machine folder - mapred180506 using filezilla (or download from google cloud in linux directly)

4. Give executable authority to python program

[cloudera@quickstart mapred180506]$ chmod a+x \*.py

\*\*\*\*\*\*\*\*\*\*\*\* 1st program \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

5. Execution of 1st map-reduce program for word count

Count Number of Words in a file

Mapper: wordcount\_mapper.py (Converts to Key-Value pair)

Reducer: wordcount\_reducer.py (Workon/Reduce key value pair to count words)

A. Mapper

[cloudera@quickstart mapred180506]$ echo "I felt happy because i saw the others were happy and because I knew I should feel happy, but I wasn't really happy"|./wordcount\_mapper.py

I 1

felt 1

happy 1

because 1

i 1

saw 1

the 1

others 1

were 1

happy 1

and 1

because 1

I 1

knew 1

I 1

should 1

feel 1

happy, 1

but 1

I 1

wasn't 1

really 1

happy 1

B. Mapper + Sort

[cloudera@quickstart mapred180506]$ echo "I felt happy because i saw the others were happy and because I knew I should feel happy, but I wasn't really happy"|./wordcount\_mapper.py|sort

and 1

because 1

because 1

but 1

feel 1

felt 1

happy, 1

happy 1

happy 1

happy 1

i 1

I 1

I 1

I 1

I 1

knew 1

others 1

really 1

saw 1

should 1

the 1

wasn't 1

were 1

C. Mapper + Sort + Reducer

[cloudera@quickstart mapred180506]$ echo "I felt happy because i saw the others were happy and because I knew I should feel happy, but I wasn't really happy"|./wordcount\_mapper.py|sort|./wordcount\_reducer.py

and 1

because 2

but 1

feel 1

felt 1

happy, 1

happy 3

i 1

I 4

knew 1

others 1

really 1

saw 1

should 1

the 1

wasn't 1

were 1

6. Run the same Map-Reduce program to perform wordcount in Hadoop

\* Make input file available in the path /user/harish/wordcount\_input in HDFS

[cloudera@quickstart mapred180506]$ /usr/bin/hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar -files wordcount\_mapper.py,wordcount\_reducer.py -mapper wordcount\_mapper.py -reducer wordcount\_reducer.py -input /user/harish/wordcount\_input -output /user/harish/wordcount\_output

\* Output file will be available in the path /user/harish/wordcount\_output

\*\*\*\*\*\*\*\*\*\* 1st program end \*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\* 2nd program \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

NOTE: In class we did this program is MRJob (Developed by Yelp). Unfortunately cloudera VM and HortonworksVM are still on Python 2.6 and MRJob has removed support for python 2.6 in recent release.

Hence we have converted program to make it separate mapper and reducer.

7. Most used word

A. Mapper

[cloudera@quickstart mapred180506]$ echo "I felt happy because i saw the others were happy and because I knew I should feel happy but I wasn't really happy"| ./mostcommonword\_mapper.py

i,1

felt,1

happy,1

because,1

i,1

saw,1

the,1

others,1

were,1

happy,1

and,1

because,1

i,1

knew,1

i,1

should,1

feel,1

happy,1

but,1

i,1

wasn't,1

really,1

happy,1

B. Mapper + Sort

[cloudera@quickstart mapred180506]$ echo "I felt happy because i saw the others were happy and because I knew I should feel happy but I wasn't really happy"| ./mostcommonword\_mapper.py|sort

and,1

because,1

because,1

but,1

feel,1

felt,1

happy,1

happy,1

happy,1

happy,1

i,1

i,1

i,1

i,1

i,1

knew,1

others,1

really,1

saw,1

should,1

the,1

wasn't,1

were,1

C. Mapper + Sort +

[cloudera@quickstart mapred180506]$ echo "I felt happy because i saw the others were happy and because I knew I should feel happy but I wasn't really happy"| ./mostcommonword\_mapper.py|sort|./mostcommonword\_reducer.py

('i', 5)

8. You can also run in hadoop in same way as in line 6

###### USing MRJob

[cloudera@quickstart mapred180506]$ python most\_common\_word.py wordcount\_input.xls

5 i

######

\*\*\*\*\*\*\*\*\*\*\* 2nd program end \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\* 3rd program \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

NOTE:In class we did this program is MRJob (Developed by Yelp). Unfortunately cloudera VM and HortonworksVM are still on Python 2.6 and MRJob has removed support for python 2.6 in recent release.

Hence we are unable to run this program.

As a assignment, you can convert the below program into separate mappers and reducers as done for 2nd program

9. email analysis to analyze most emails sent to 'H'

[cloudera@quickstart mapred180506]$ python email\_mapreduce.py email\_input.xls

No configs found; falling back on auto-configuration

Creating temp directory /tmp/email\_mapreduce.cloudera.20180505.163659.863303

Running step 1 of 3...

Running step 2 of 3...

Running step 3 of 3...

Streaming final output from /tmp/email\_mapreduce.cloudera.20180505.163659.863303/output...

"H" "\"Coleman,\"Huma,\"Koh"

\*\*\*\*\*\*\*\*\*\*\* 3rd program end \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*