

# LAB 3 INSTRUCTIONS

CKME 134 – BIG DATA ANALYTICS TOOLS

RYERSON UNIVERSITY

SPRING 2015

Instructor: Shaohua Zhang

# Session 3 - Lab & Assignments

## □ Lab Computer

- ▣ Username: datastudent
- ▣ Password: datastudent

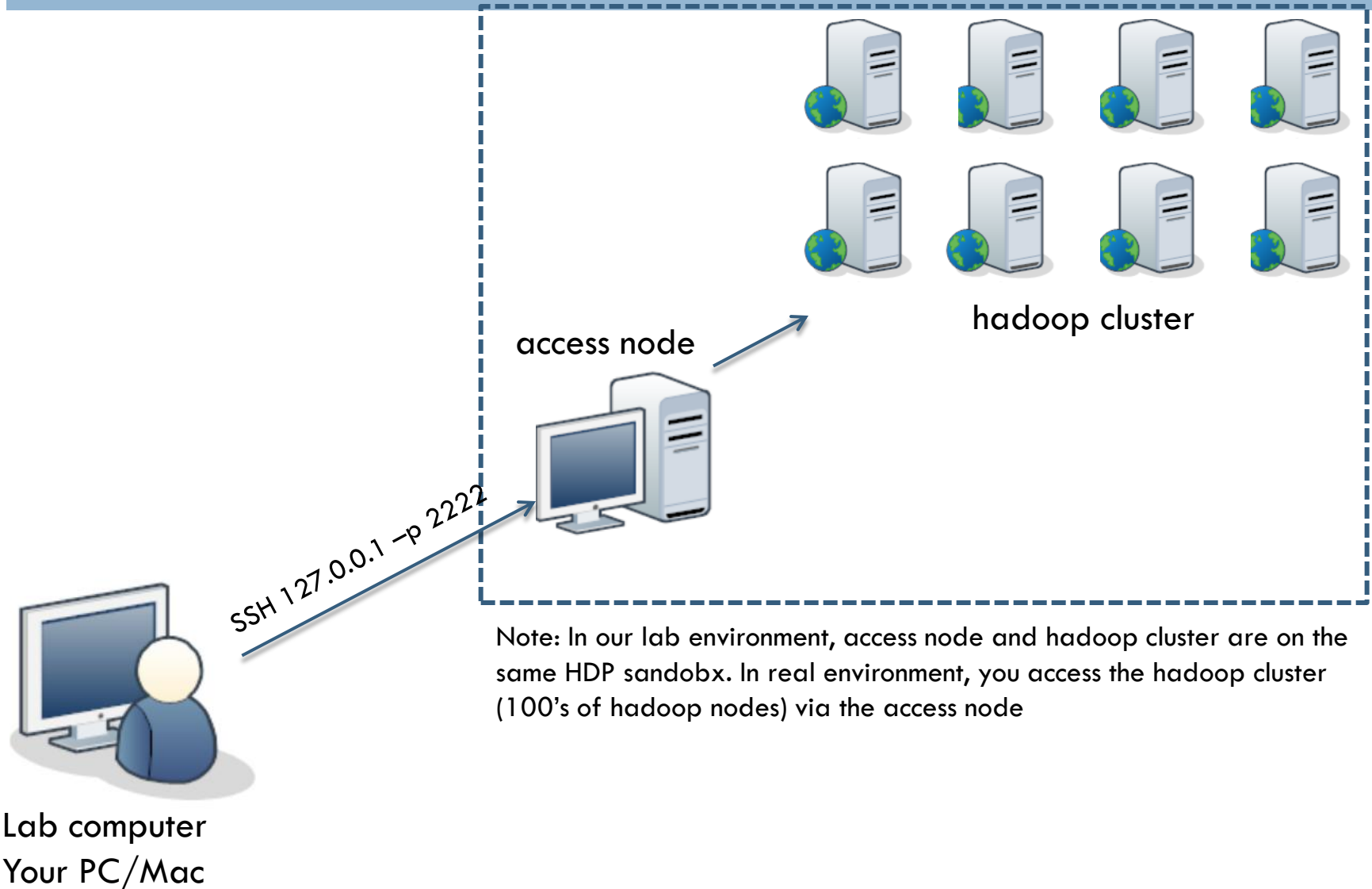
## □ HDP Sandbox

- ▣ Username: root
- ▣ Password: hadoop

## □ Lab 2

- ▣ Hadoop filesystem shell commands
- ▣ Hive Wordcount
- ▣ Pig Wordcount
- ▣ Wordcount MapReduce

# Lab Environment



# Lab 3 – Hadoop and WordCount

- Today's Lab
  - ▣ Hadoop Shell
  - ▣ WordCount – Hive
- Supplementary
  - ▣ WordCount – Java M/R
  - ▣ WordCount – Python Streaming

# Lab 3 – Before We Start...

- ❑ Open Virtualbox
- ❑ Start HDP Sandbox
  - ▣ If you don't see HDP Sandbox, reload it
    - Please mark the computer ID so that I can tell the lab director!
- ❑ Start Putty and connect to Sandbox access node (client)

# Lab 3 - Before We Start...

```
[root@sandbox ~]# ll
total 32
```

```
-rw----- 1 root root 2143 Dec 16 18:13 anaconda-ks.cfg
-rw-r--r-- 1 root root 9436 Dec 16 18:13 install.log
-rw-r--r-- 1 root root 3314 Dec 16 18:12 install.log.syslog
drwxr-xr-x 8 root root 4096 Dec 16 19:33 ranger_tutorial
lrwxrwxrwx 1 root root 48 Dec 16 19:15 start_ambari.sh -> /usr/lib/hue/tools/start_scripts/start_ambari.sh
lrwxrwxrwx 1 root root 47 Dec 16 19:17 start_hbase.sh -> /usr/lib/hue/tools/start_scripts/start_hbase.sh
-rwxrwxrwx 1 vagrant vagrant 241 Dec 16 19:15 start_solr.sh
-rwxrwxrwx 1 vagrant vagrant 63 Dec 16 19:17 stop_solr.sh
```

```
[root@sandbox ~]# cd /home/lab
```

```
[root@sandbox lab]# ll
total 115364
```

```
drwxr-xr-x 5 nagios games 4096 Oct 12 2010 GeoText.2010-10-12
-rw-r--r-- 1 root root 60973289 Jan 16 21:47 GeoText.2010-10-12.tgz
-rw-r--r-- 1 root root 57139942 Jan 16 15:34 full_text.txt
-rwxrwxr-- 1 root root 1027 Jan 23 21:38 sc_reducer.py
-rwxrwxr-- 1 root root 537 Jan 23 21:38 wc_mapper.py
```

```
[root@sandbox lab]# cd GeoText.2010-10-12
```

```
[root@sandbox GeoText.2010-10-12]# ll
total 55820
```

```
-rw-r--r-- 1 nagios games 2695 Oct 12 2010 README.txt
-rw-r--r-- 1 nagios games 57139942 Oct 12 2010 full_text.txt
drwxr-xr-x 2 nagios games 4096 Jan 16 21:52 geo_eval
drwxr-xr-x 2 nagios games 4096 Jan 16 21:52 preproc
drwxr-xr-x 2 nagios games 4096 Jan 16 21:52 processed_data
```

```
[root@sandbox GeoText.2010-10-12]# cat full_text.txt | head
```

```
USER_79321756 2010-03-03T04:15:26 ÜT: 47.528139,-122.197916 47.528139 -122.197916 RT @USER_2ff4faca: IF SHE DO IT 1 MORE TIME.....IMA KNOW
&gt;haha. #cutthatout
USER_79321756 2010-03-03T04:55:32 ÜT: 47.528139,-122.197916 47.528139 -122.197916 @USER_77a4822d @USER_2ff4faca okay:) lol. Saying ok to bo
USER_79321756 2010-03-03T05:13:34 ÜT: 47.528139,-122.197916 47.528139 -122.197916 RT @USER_5d4d777a: YOURE A FAG FOR GETTING IN THE MIDDLE
OU ? A FUCKING NOBODY !!!!!&gt;&gt;Lol! Dayum! Aye!
USER_79321756 2010-03-03T05:28:02 ÜT: 47.528139,-122.197916 47.528139 -122.197916 @USER_77a4822d yea ok..well answer that cheap as Sweden p
USER_79321756 2010-03-03T05:56:13 ÜT: 47.528139,-122.197916 47.528139 -122.197916 A sprite can disappear in her mouth - lil kim hmmm the
USER_79321756 2010-03-03T16:52:44 ÜT: 47.528139,-122.197916 47.528139 -122.197916 Lmao! I still get txt when AJ tweets before they even pos
s me dyin! @USER_a5b463b2 what's ur issue!
USER_79321756 2010-03-03T16:57:24 ÜT: 47.528139,-122.197916 47.528139 -122.197916 Alright twitters tryna take me over!
USER_79321756 2010-03-03T20:20:40 ÜT: 47.528139,-122.197916 47.528139 -122.197916 Just got to work. Got my pizza bagel and my raspberry ice
not til 2. I just wanna get it done!:D
USER_79321756 2010-03-03T23:23:33 ÜT: 47.528139,-122.197916 47.528139 -122.197916 Just got a txt from my cousin! Yes! So happy for you @USE
USER_79321756 2010-03-03T23:37:36 ÜT: 47.528139,-122.197916 47.528139 -122.197916 Why is this woman in the bathroom everytime I'm in the ba
```

```
[root@sandbox GeoText.2010-10-12]#
```

Twitter data directory

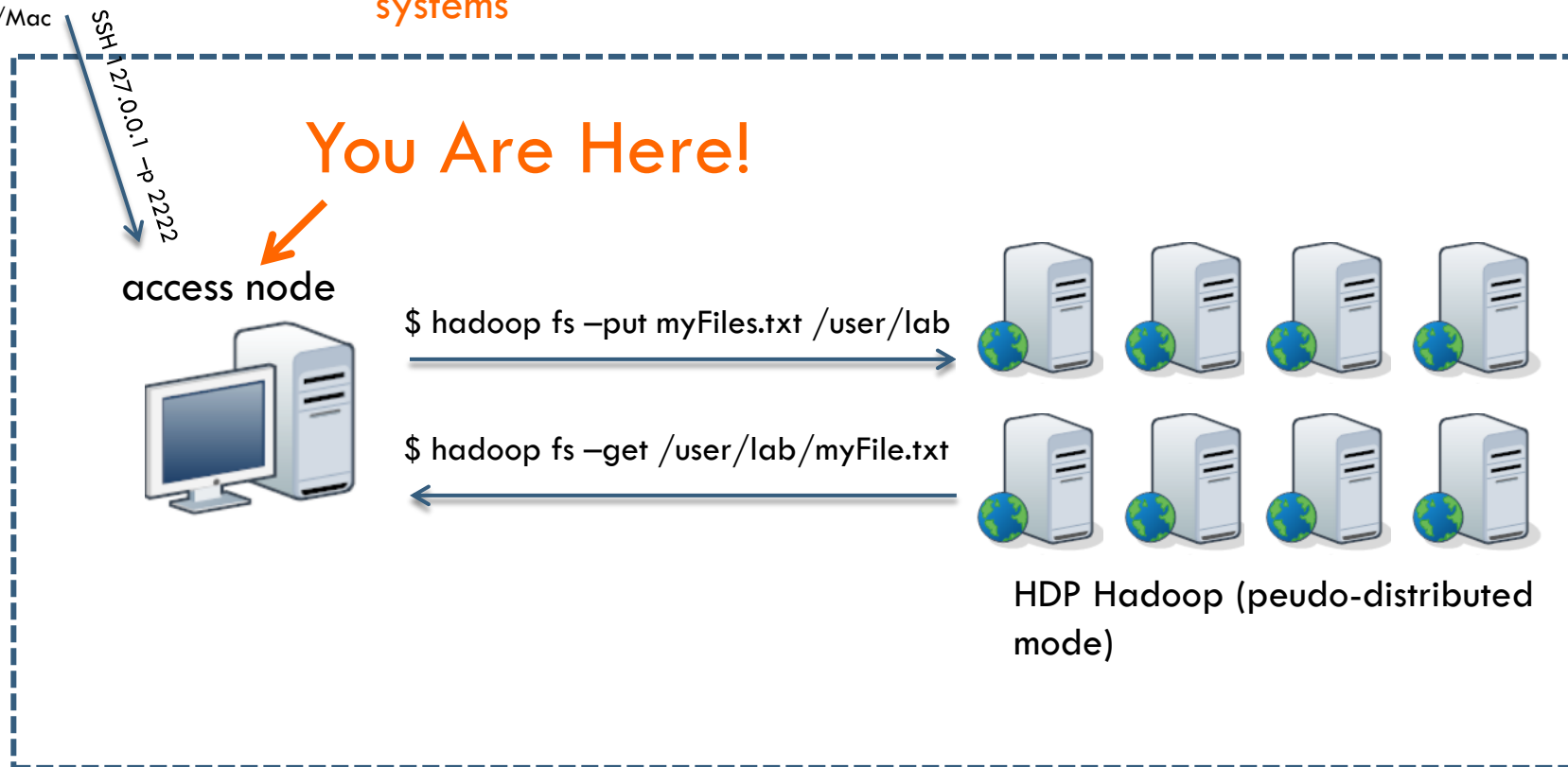
Make sure you can use cat to view the content of the data

# Lab 3 – Now Let's Start...



Lab computer  
Your PC/Mac

NOTE: Your client linux file system and hadoop filesystem (HDFS) are separate environment. First, you need to learn how to move files between the two file systems



# Hadoop FileSystem Shell

- ❑ **hadoop fs -mkdir**
- ❑ **hadoop fs -ls**
- ❑ **hadoop fs -put**
- ❑ **hadoop fs -cat**
- ❑ **hadoop fs -get**
- ❑ **hadoop fs -rm**
- ❑ **hadoop fs -rmdir**
- ❑ **hadoop fs -count**
- ❑ **hadoop fs -cp**
- ❑ **hadoop fs -du**
- ❑ **hadoop fs -mv**
- ❑ **hadoop fs -tail**
- ❑ **hadoop fs -getmerge**



# Lab 3 – Getting Files into HDFS

```
[root@sandbox GeoText.2010-10-12]# hadoop fs -ls /user/
```

Found 10 items

```
drwxr-xr-x - hue hdfs 0 2015-01-15 06:14 /user/4sq
drwxrwx--- - ambari-qa hdfs 0 2014-12-16 19:04 /user/ambari-qa
drwxr-xr-x - guest guest 0 2014-12-16 19:28 /user/guest
drwxr-xr-x - hcat hdfs 0 2014-12-16 19:13 /user/hcat
drwx----- - hive hdfs 0 2014-12-16 19:08 /user/hive
drwxr-xr-x - hue hue 0 2014-12-16 19:27 /user/hue
drwxrwxr-x - oozie hdfs 0 2014-12-16 19:10 /user/oozie
drwx----- - root hdfs 0 2015-01-24 05:19 /user/root
drwxr-xr-x - solr hdfs 0 2014-12-16 19:24 /user/solr
drwxr-xr-x - hue hdfs 0 2015-01-23 21:58 /user/twitter
```

You don't see a /user/lab folder in HDFS yet

```
[root@sandbox GeoText.2010-10-12]# hadoop fs -mkdir /user/lab/
```

```
[root@sandbox GeoText.2010-10-12]# hadoop fs -ls /user
```

Create the lab folder in HDFS

Found 11 items

```
drwxr-xr-x - hue hdfs 0 2015-01-15 06:14 /user/4sq
drwxrwx--- - ambari-qa hdfs 0 2014-12-16 19:04 /user/ambari-qa
drwxr-xr-x - guest guest 0 2014-12-16 19:28 /user/guest
drwxr-xr-x - hcat hdfs 0 2014-12-16 19:13 /user/hcat
drwx----- - hive hdfs 0 2014-12-16 19:08 /user/hive
drwxr-xr-x - hue hue 0 2014-12-16 19:27 /user/hue
drwxr-xr-x - root hdfs 0 2015-01-24 05:20 /user/lab
drwxrwxr-x - oozie hdfs 0 2014-12-16 19:10 /user/oozie
drwx----- - root hdfs 0 2015-01-24 05:19 /user/root
drwxr-xr-x - solr hdfs 0 2014-12-16 19:24 /user/solr
drwxr-xr-x - hue hdfs 0 2015-01-23 21:58 /user/twitter
```

```
[root@sandbox GeoText.2010-10-12]# hadoop fs -put full_text.txt /user/lab/
```

```
[root@sandbox GeoText.2010-10-12]# hadoop fs -ls /user/lab
```

Upload the full\_text file to HDFS

Found 1 items

```
-rw-r--r-- 1 root hdfs 57139942 2015-01-24 05:21 /user/lab/full_text.txt
```

```
[root@sandbox GeoText.2010-10-12]# hadoop fs -cat /user/lab/full_text.txt | head -n 5
```

Display the first few lines of HDFS file

```
USER_79321756 2010-03-03T04:15:20 UT: 47.528139,-122.197916 47.528139 -122.197916 RT @USER_2ff4faca: IF SHE DO IT 1 MORE TII
&gt;haha. #cutthatout
USER_79321756 2010-03-03T04:55:32 UT: 47.528139,-122.197916 47.528139 -122.197916 @USER_77a4822d @USER_2ff4faca okay:) lol.
USER_79321756 2010-03-03T05:13:34 UT: 47.528139,-122.197916 47.528139 -122.197916 RT @USER_5d4d777a: YOURE A FAG FOR GETTIN'
OU ? A FUCKING NOBODY !!!!&gt;&gt;Lol! Dayum! Aye!
USER_79321756 2010-03-03T05:28:02 UT: 47.528139,-122.197916 47.528139 -122.197916 @USER_77a4822d yea ok..well answer that c
USER_79321756 2010-03-03T05:56:13 UT: 47.528139,-122.197916 47.528139 -122.197916 A sprite can disappear in her mouth - lil
cat: Unable to write to output stream.
```

```
[root@sandbox GeoText.2010-10-12]# hadoop fs -get /user/lab/full_text.txt full_text_2.txt
```

```
[root@sandbox GeoText.2010-10-12]# ll
```

Download a file from HDFS to access node

total 111624

```
-rw-r--r-- 1 nagios games 2695 Oct 12 2010 README.txt
-rw-r--r-- 1 nagios games 57139942 Oct 12 2010 full_text.txt
-rw-r--r-- 1 root root 57139942 Jan 24 05:23 full_text_2.txt
drwxr-xr-x 2 nagios games 4096 Jan 16 21:52 geo_eval
drwxr-xr-x 2 nagios games 4096 Jan 16 21:52 preproc
drwxr-xr-x 2 nagios games 4096 Jan 16 21:52 processed_data
```

Make a copy of the HDFS file and rename

```
[root@sandbox GeoText.2010-10-12]# hadoop fs -cp /user/lab/full_text.txt /user/lab/full_text_2.txt
```

```
[root@sandbox GeoText.2010-10-12]# hadoop fs -ls /user/lab
```

Found 2 items

```
-rw-r--r-- 1 root hdfs 57139942 2015-01-24 05:21 /user/lab/full_text.txt
-rw-r--r-- 1 root hdfs 57139942 2015-01-24 05:24 /user/lab/full_text_2.txt
```

Delete a file in HDFS

```
[root@sandbox GeoText.2010-10-12]# hadoop fs -rm /user/lab/full_text_2.txt
```

```
15/01/24 05:25:20 INFO fs.TrashPot cvDefault: Namenode trash configuration: Deletion interval = 360 minutes, Emptier interval = 0 minutes.
Moved: 'hdfs://sandbox.hortonworks.com:8020/user/lab/full_text_2.txt' to trash at: hdfs://sandbox.hortonworks.com:8020/user/root/.Trash/Current
[root@sandbox GeoText.2010-10-12]#
```



```
# hive
conf: HiveConf{}
```



```
show tables;
```



```
create table full_text (line string);
```



```
load data inpath '/user/lab/full_text.txt' overwrite into table full_text;
```



```
select * from full_text limit 2;
```

79321756	2010-03-03T04:15:26	ÜT: 47.528139,-122.197916	47.528139
----------	---------------------	---------------------------	-----------



```
> create table wordcount as
> select word, count(1) as count from
```

```
hive> create table wordcount as
> select word, count(1) as count from
> (select explode(split(line, '[\s+ +\t+]')) as word from full_text) w
> group by word
> order by count desc;
```

← The WordCount query

Query ID = root\_20150124053838\_1e737143-4823-4df5-8483-959c0e33e8bb

Total jobs = 2

Launching Job 1 out of 2 ← Launching a MapReduce job

Number of reduce tasks not specified. Estimated from input data size: 1

In order to change the average load for a reducer (in bytes):

set hive.exec.reducers.bytes.per.reducer=<number>

In order to limit the maximum number of reducers:

set hive.exec.reducers.max=<number>

In order to set a constant number of reducers:

set mapreduce.job.reduces=<number>

Starting Job = job\_1422074871964\_0001, Tracking URL = http://sandbox.hortonworks.com:8088/proxy/application\_1422074871964\_0001/

Kill Command = /usr/hdp/2.2.0.0-2041/hadoop/bin/hadoop job -kill job\_1422074871964\_0001

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2015-01-24 05:38:56,317 Stage-1 map = 0%, reduce = 0%

2015-01-24 05:39:13,601 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 15.37 sec

2015-01-24 05:39:14,650 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 16.65 sec

2015-01-24 05:39:26,471 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 24.05 sec

MapReduce Total cumulative CPU time: 24 seconds 50 msec

Ended Job = job\_1422074871964\_0001

Launching Job 2 out of 2

Number of reduce tasks determined at compile time: 1

In order to change the average load for a reducer (in bytes):

set hive.exec.reducers.bytes.per.reducer=<number>

In order to limit the maximum number of reducers:

set hive.exec.reducers.max=<number>

In order to set a constant number of reducers:

set mapreduce.job.reduces=<number>

Starting Job = job\_1422074871964\_0002, Tracking URL = http://sandbox.hortonworks.com:8088/proxy/application\_1422074871964\_0002/

Kill Command = /usr/hdp/2.2.0.0-2041/hadoop/bin/hadoop job -kill job\_1422074871964\_0002

Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1

2015-01-24 05:39:35,374 Stage-2 map = 0%, reduce = 0%

2015-01-24 05:39:44,985 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 5.22 sec

2015-01-24 05:39:56,195 Stage-2 map = 100%, reduce = 80%, Cumulative CPU 12.43 sec

2015-01-24 05:39:58,370 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 14.27 sec

MapReduce Total cumulative CPU time: 14 seconds 270 msec

Ended Job = job\_1422074871964\_0002

Moving data to: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/wordcount

Table default.wordcount stats: [numFiles=1, numRows=865487, totalSize=13935272, rawDataSize=13069785]

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 24.05 sec HDFS Read: 57140175 HDFS Write: 27997246 SUCCESS

Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 14.27 sec HDFS Read: 27997661 HDFS Write: 13935355 SUCCESS

Total MapReduce CPU Time Spent: 38 seconds 320 msec

OK

Time taken: 85.974 seconds

hive> select \* from wordcount limit 5;

ok

Done!

588285

ÜT: 339083

I 110427

a 81038

the 78480

Select query to display the first 5 word count

# Lab 3 – Supplementary

- To learn more about Hadoop shell commands, check out the documentations

<http://hadoop.apache.org/docs/r2.6.0/hadoop-project-dist/hadoop-common/FileSystemShell.html>

# Lab 3 – Supplementary

## - Java M/R WordCount

### 1. Download the Shakespeare file

- ▣ Shakespeare dataset has been uploaded to Blackboard “datasets” page

### 2. Put the shakespeare file into HDFS

### 3. Find your mapreduce-example jar file

- ▣ `$ find /usr -name *mapreduce-example*`

### 4. Run java M/R wordcount example


- ▣ `$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.0.jar wordcount /user/shaohua/shakespeare /user/shaohua/shakespeare-wc-out`

Change the path accordingly

### ▣ View results

# Lab 3 – Supplementary

## - Python Streaming - WordCount

1. Download the Shakespeare file
  - ▣ Shakespeare dataset has been uploaded to Blackboard “datasets” page
2. Put the shakespeare file into HDFS
3. Find your mapreduce-example jar file
  - ▣ `$ find /usr -name *hadoop-streaming*`
4. Run Hadoop streaming wordcount  Change the path accordingly
  - ▣ `$ hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/mapreduce/hadoop-streaming-2.6.0.jar -file /Users/DSinmotion/Ryerson/Demos/scripts/wc_mapper.py -mapper /Users/DSinmotion/Ryerson/Demos/scripts/wc_mapper.py -file /Users/DSinmotion/Ryerson/Demos/scripts/wc_reducer.py -reducer /Users/DSinmotion/Ryerson/Demos/scripts/wc_reducer.py -input /user/shaohua/shakespeare -output /user/shaohua/shakespeare-wc-out-1`
5. View results