

ITM_6273_02
Homework Assignment #3
Quach, Hoa
NetID: gt4789

1. **Work Examples 7-2, 7-3, and 7-4 on CSUEB Hadoop. Type out all the commands in each step of the process and print out a screenshot of the final results in CSUEB Hadoop.**

Hint 1: need to create a jar file including five classes:

WholeFileInputFormat.class,
WholeFileRecordReader.class,
SmallFilesToSequenceFileConverter.class,
SmallFilesToSequenceFileConverter\$SequenceFileMapper.class,
JobBuilder.class

Note. *SmallFilesToSequenceFileConverter.class* is the main class.

SmallFilesToSequenceFileConverter\$SequenceFileMapper.class is a nested/inner class of *SmallFilesToSequenceFileConverter.class*.

JobBuilder.java can be found in Hadoop-Book-Master/common/src/main/java

Smallfiles folder can be found in Hadoop-Book-Master/input

Hint 2: *-conf conf/Hadoop-localhost.xml* is not needed in the hadoop jar command. So, your command will be like this:

*hadoop jar /home/jwu/hadoop-example.jar SmallFilesToSequenceFileConverter -D
mapred.reduce.tasks=2 /home/jwu/smallfiles /home/jwu/output11*

Commands

****Note:** All .java files have been placed in the root directory of /home/hadoop/ directory in AWS.

```
mkdir /home/hadoop/Assignment_3_1
```

```
javac -cp src/:hadoop-common-2.6.1.jar:hadoop-mapreduce-  
client-core-2.6.1.jar:commons-cli-2.0.jar -d  
Assignment_3_1/ *.java
```

```
jar -cvf  
Assignment_3_1/SmallFilesToSequenceFileConverter.jar -C  
Assignment_3_1/ .
```

```
hdfs dfs -mkdir -p /user/hadoop/input
```

****After manually uploading smallfiles directory from local desktop to AWS Hadoop server via WinSCP:**

```
hdfs dfs -copyFromLocal /home/hadoop/smallfiles
/user/hadoop/input
```

```
hdfs dfs -ls /user/hadoop/input/smallfiles
```

```
hadoop jar
Assignment_3_1/SmallFilesToSequenceFileConverter.jar
SmallFilesToSequenceFileConverter -D mapred.reduce.tasks=2
/user/hadoop/input/smallfiles/ /user/hadoop/output
```

```
hdfs dfs -ls /user/hadoop/output
```

```
hadoop fs -text /user/hadoop/output/part-r-00000
```

```
hadoop fs -text /user/hadoop/output/part-r-00001
```

```
hadoop@ip-172-31-52-87:~$
Reduce input records=6
Reduce output records=6
Spilled Records=12
Shuffled Maps =12
Failed Shuffles=0
Merged Map outputs=12
GC time elapsed (ms)=1044
CPU time spent (ms)=7330
Physical memory (bytes) snapshot=3007627264
Virtual memory (bytes) snapshot=18858094592
Total committed heap usage (bytes)=3589799936

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=50
File Output Format Counters
  Bytes Written=728
[hadoop@ip-172-31-52-87 ~]$ hdfs dfs -ls /user/hadoop/output
Found 3 items
-rw-r--r-- 1 hadoop hadoop          0 2016-05-12 18:08 /user/hadoop/output/_SU
CESS
-rw-r--r-- 1 hadoop hadoop        359 2016-05-12 18:08 /user/hadoop/output/part-
r-00000
-rw-r--r-- 1 hadoop hadoop        369 2016-05-12 18:08 /user/hadoop/output/part-
r-00001
[hadoop@ip-172-31-52-87 ~]$ hadoop fs -text /user/hadoop/output/part-r-00000
hdfs://ip-172-31-52-87.ec2.internal:8020/user/hadoop/input/smallfiles/a 61 61 61
61 61 61 61 61 61 61
hdfs://ip-172-31-52-87.ec2.internal:8020/user/hadoop/input/smallfiles/c 63 63 63
63 63 63 63 63 63 63
hdfs://ip-172-31-52-87.ec2.internal:8020/user/hadoop/input/smallfiles/e
[hadoop@ip-172-31-52-87 ~]$ hadoop fs -text /user/hadoop/output/part-r-00001
hdfs://ip-172-31-52-87.ec2.internal:8020/user/hadoop/input/smallfiles/b 62 62 62
62 62 62 62 62 62 62
hdfs://ip-172-31-52-87.ec2.internal:8020/user/hadoop/input/smallfiles/d 64 64 64
64 64 64 64 64 64 64
hdfs://ip-172-31-52-87.ec2.internal:8020/user/hadoop/input/smallfiles/f 66 66 66
66 66 66 66 66 66 66
[hadoop@ip-172-31-52-87 ~]$
```

2. Work Example 8-1 on CSUEB Hadoop. Type out all the commands in each step of the process and print out a screenshot of the final results (the counters) in CSUEB Hadoop.

Hint: need to create a jar file including five classes:

MaxTemperatureWithCounters.class,
MaxTemperatureMapperWithCounters.class,
NcdcRecordParser.class,
JobBuilder.class,
MaxTemperatureReducer.class.

Note. 1) *NcdcRecordParser.java* can be found in Hadoop-Book-Master/common/src/main/java, 2) you must use the data of year 1930 as input data to run the program (download data at: <ftp://ftp.ncdc.noaa.gov/pub/data/noaa/>), and 3) your results shall be different from those run over the complete dataset of 100 years, which are shown on page 265 in the textbook.

Note:

******To download all 1930 files from <ftp://ftp.ncdc.noaa.gov/pub/data/noaa/1930> to /user/hadoop/ (AWS)

```
mkdir /home/hadoop/1930

cd 1930

wget -r --no-directories
ftp://ftp.ncdc.noaa.gov/pub/data/noaa/1930/

ls -l

cd ../
```

Commands using AWS

****Note:** All .java files have been placed in the root directory of /home/hadoop/ directory in AWS and problem#1 java codes have been removed (except JobBuilder.java).

```
mkdir /home/hadoop/MaxTemperature

javac -cp src/:hadoop-common-2.6.1.jar:hadoop-mapreduce-
client-core-2.6.1.jar:commons-cli-2.0.jar -d
MaxTemperature/ *.java

jar -cvf MaxTemperature/MaxTemperatureWithCounters.jar -C
MaxTemperature/ .
```

```
hdfs dfs -mkdir -p /user/hadoop/input2
```

```
hdfs dfs -copyFromLocal /home/hadoop/1930/  
/user/hadoop/input2
```

```
hdfs dfs -ls /user/hadoop/input2/1930/
```

```
hadoop jar MaxTemperature/MaxTemperatureWithCounters.jar  
MaxTemperatureWithCounters /user/hadoop/input2/1930/  
/user/hadoop/output1930
```

```
hdfs dfs -ls /user/hadoop/output1930
```

```
hadoop@ip-172-31-52-87:~$  
Total time spent by all reduce tasks (ms)=196125  
Total vcore-milliseconds taken by all map tasks=1269125  
Total vcore-milliseconds taken by all reduce tasks=196125  
Total megabyte-milliseconds taken by all map tasks=1827540000  
Total megabyte-milliseconds taken by all reduce tasks=564840000  
Map-Reduce Framework  
  Map input records=89262  
  Map output records=85580  
  Map output bytes=770220  
  Map output materialized bytes=6974  
  Input split bytes=18150  
  Combine input records=85580  
  Combine output records=106  
  Reduce input groups=1  
  Reduce shuffle bytes=6974  
  Reduce input records=106  
  Reduce output records=1  
  Spilled Records=212  
  Shuffled Maps =363  
  Failed Shuffles=0  
  Merged Map outputs=363  
  GC time elapsed (ms)=17350  
  CPU time spent (ms)=144040  
  Physical memory (bytes) snapshot=53904248832  
  Virtual memory (bytes) snapshot=257645109248  
  Total committed heap usage (bytes)=62495129600  
MaxTemperatureWithCounters$Temperature  
  MISSING=3665  
Shuffle Errors  
  BAD_ID=0  
  CONNECTION=0  
  IO_ERROR=0  
  WRONG_LENGTH=0  
  WRONG_MAP=0  
  WRONG_REDUCE=0  
TemperatureQuality  
  1=85580  
  2=17  
  9=3665  
File Input Format Counters  
  Bytes Read=1649548  
File Output Format Counters  
  Bytes Written=9  
[hadoop@ip-172-31-52-87 ~]$ hdfs dfs -ls /user/hadoop/output1930
```