Assignment 3

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,* and *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*

   Example 7-2 An InputFormat for reading a whole file as a record

   Example 7-3. The RecordReader used by WholeFileInputFormat for reading a whole file as a record

   Example 7-4. A MapReduce program for packaging a collection of small files as a single SequenceFile

   Compile Java files:

   ```
   javac -classpath /home/student7/hadoop-common-2.6.1.jar:/home/student7/hadoop-mapreduce-client-core-2.6.1.jar:/home/student7/commons-cli-2.0.jar -d . WholeFileInputFormat.java WholeFileRecordReader.java SmallFilesToSequenceFileConverter.java JobBuilder.java
   ```

   Create a Jar file:

   ```
   jar -cvf SmallFileSequence.jar WholeFileInputFormat.class WholeFileRecordReader.class SmallFilesToSequenceFileConverter*.class JobBuilder.class
   ```

   Make a directory in HDFS:

   ```
   hdfs dfs -mkdir /home/student7/smallfiles/
   ```

   Copy a file from local drive to HDFS:

   ```
   hdfs dfs -copyFromLocal a /home/student7/smallfiles/a
   ```

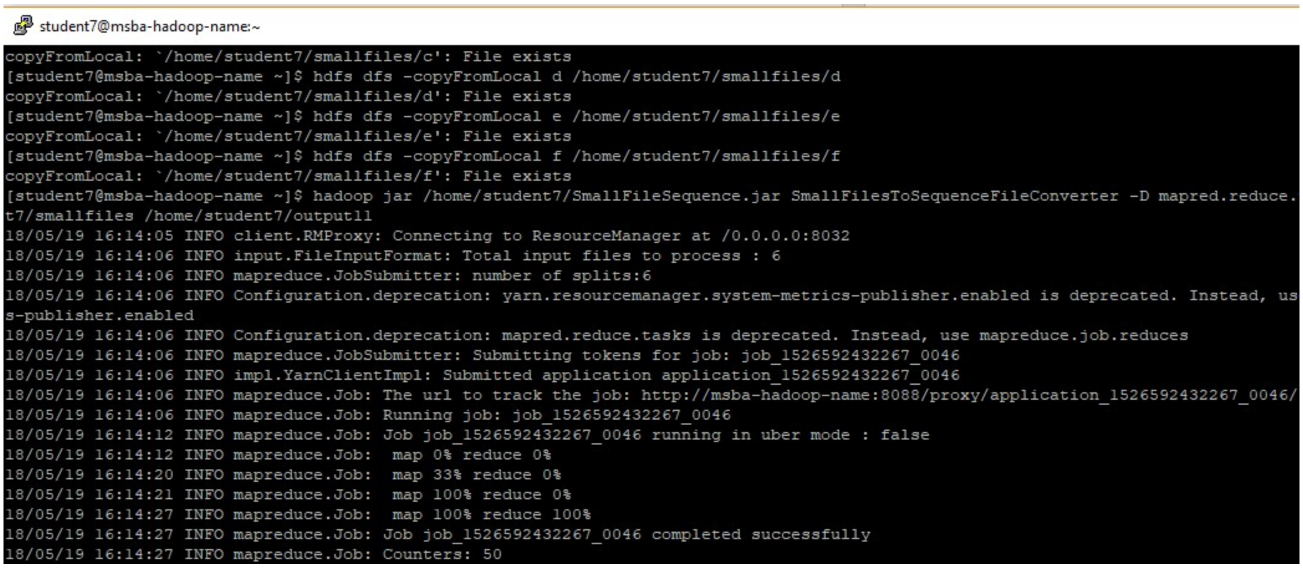
hdfs dfs -copyFromLocal b /home/student7/smallfiles/b
hdfs dfs -copyFromLocal c /home/student7/smallfiles/c
hdfs dfs -copyFromLocal d /home/student7/smallfiles/d
hdfs dfs -copyFromLocal e /home/student7/smallfiles/e
hdfs dfs -copyFromLocal f /home/student7/smallfiles/f

Run a Jar file on Hadoop:

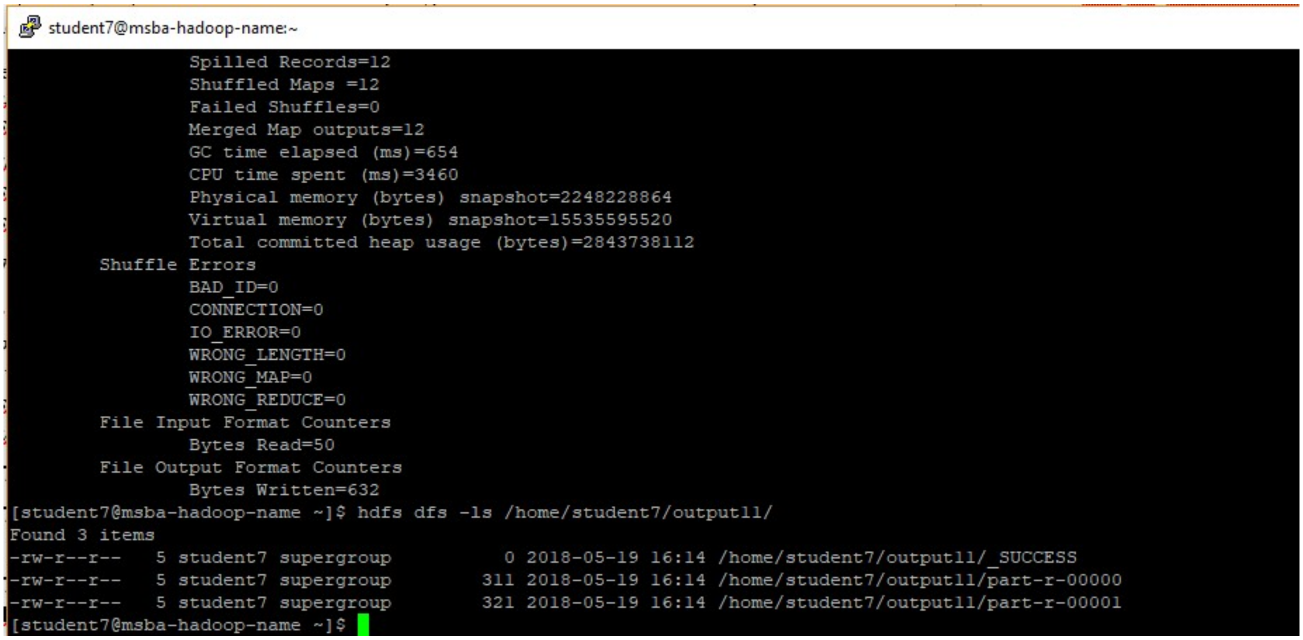
hadoop jar /home/student7/SmallFileSequence.jar SmallFilesToSequenceFileConverter -D mapred.reduce.tasks=2 /home/student7/smallfiles /home/student7/output11

List files in HDFS:
hdfs dfs -ls /home/student7/output11/

Display the output on screen:
hadoop fs -text /home/student7/output11/part-r-00000
hadoop fs -text /home/student7/output11/part-r-00001

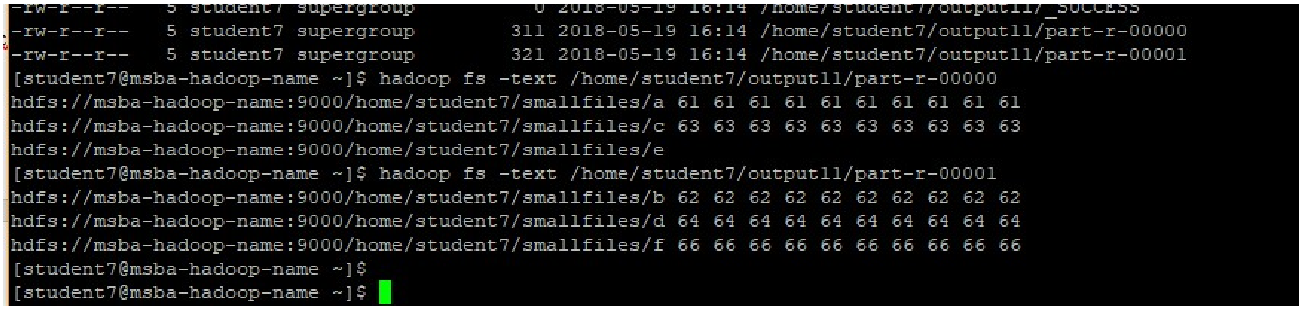


```
copyFromLocal: `/home/student7/smallfiles/c': File exists
[student7@msba-hadoop-name ~]$ hdfs dfs -copyFromLocal d /home/student7/smallfiles/d
copyFromLocal: `/home/student7/smallfiles/d': File exists
[student7@msba-hadoop-name ~]$ hdfs dfs -copyFromLocal e /home/student7/smallfiles/e
copyFromLocal: `/home/student7/smallfiles/e': File exists
[student7@msba-hadoop-name ~]$ hdfs dfs -copyFromLocal f /home/student7/smallfiles/f
copyFromLocal: `/home/student7/smallfiles/f': File exists
[student7@msba-hadoop-name ~]$ hadoop jar /home/student7/SmallFileSequence.jar SmallFilesToSequenceFileConverter -D mapred.reduce.
t7/smallfiles /home/student7/output11
18/05/19 16:14:05 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/05/19 16:14:06 INFO input.FileInputFormat: Total input files to process : 6
18/05/19 16:14:06 INFO mapreduce.JobSubmitter: number of splits:6
18/05/19 16:14:06 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, us
s-publisher.enabled
18/05/19 16:14:06 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
18/05/19 16:14:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1526592432267_0046
18/05/19 16:14:06 INFO impl.YarnClientImpl: Submitted application application_1526592432267_0046
18/05/19 16:14:06 INFO mapreduce.Job: The url to track the job: http://msba-hadoop-name:8088/proxy/application_1526592432267_0046/
18/05/19 16:14:06 INFO mapreduce.Job: Running job: job_1526592432267_0046
18/05/19 16:14:12 INFO mapreduce.Job: Job job_1526592432267_0046 running in uber mode : false
18/05/19 16:14:12 INFO mapreduce.Job:  map 0% reduce 0%
18/05/19 16:14:20 INFO mapreduce.Job:  map 33% reduce 0%
18/05/19 16:14:21 INFO mapreduce.Job:  map 100% reduce 0%
18/05/19 16:14:27 INFO mapreduce.Job:  map 100% reduce 100%
18/05/19 16:14:27 INFO mapreduce.Job: Job job_1526592432267_0046 completed successfully
18/05/19 16:14:27 INFO mapreduce.Job: Counters: 50
```



student7@msba-hadoop-name:~

```
                Spilled Records=12
                Shuffled Maps =12
                Failed Shuffles=0
                Merged Map outputs=12
                GC time elapsed (ms)=654
                CPU time spent (ms)=3460
                Physical memory (bytes) snapshot=2248228864
                Virtual memory (bytes) snapshot=15535595520
                Total committed heap usage (bytes)=2843738112
        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=632
[student7@msba-hadoop-name ~]$ hdfs dfs -ls /home/student7/output11/
Found 3 items
-rw-r--r--   5 student7 supergroup          0 2018-05-19 16:14 /home/student7/output11/_SUCCESS
-rw-r--r--   5 student7 supergroup        311 2018-05-19 16:14 /home/student7/output11/part-r-00000
-rw-r--r--   5 student7 supergroup        321 2018-05-19 16:14 /home/student7/output11/part-r-00001
[student7@msba-hadoop-name ~]$
```



```
-rw-r--r--   5 student7 supergroup          0 2018-05-19 16:14 /home/student7/output11/_SUCCESS
-rw-r--r--   5 student7 supergroup        311 2018-05-19 16:14 /home/student7/output11/part-r-00000
-rw-r--r--   5 student7 supergroup        321 2018-05-19 16:14 /home/student7/output11/part-r-00001
[student7@msba-hadoop-name ~]$ hadoop fs -text /home/student7/output11/part-r-00000
hdfs://msba-hadoop-name:9000/home/student7/smallfiles/a 61 61 61 61 61 61 61 61 61 61
hdfs://msba-hadoop-name:9000/home/student7/smallfiles/c 63 63 63 63 63 63 63 63 63 63
hdfs://msba-hadoop-name:9000/home/student7/smallfiles/e
[student7@msba-hadoop-name ~]$ hadoop fs -text /home/student7/output11/part-r-00001
hdfs://msba-hadoop-name:9000/home/student7/smallfiles/b 62 62 62 62 62 62 62 62 62 62
hdfs://msba-hadoop-name:9000/home/student7/smallfiles/d 64 64 64 64 64 64 64 64 64 64
hdfs://msba-hadoop-name:9000/home/student7/smallfiles/f 66 66 66 66 66 66 66 66 66 66
[student7@msba-hadoop-name ~]$
[student7@msba-hadoop-name ~]$
```

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, and 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.

Compile Java files:

```
javac -classpath /home/student7/hadoop-common-2.6.1.jar:/home/student7/hadoop-mapreduce-client-core-2.6.1.jar:/home/student7/commons-cli-2.0.jar -d . MaxTemperatureReducer.java MaxTemperatureWithCounters.java NcdcRecordParser.java JobBuilder.java
```

Create a Jar file:

```
jar -cvf MaxTempWithCounter.jar MaxTemperatureWithCounters*.class NcdcRecordParser.class JobBuilder.class MaxTemperatureReducer.class
```

Copy 1930 folder from local drive to HDFS:

```
hdfs dfs -copyFromLocal 1930 /home/student7_lu/
```

List files in HDFS:

```
hdfs dfs -ls  /home/student7_lu/1930
```

Export

```
export HADOOP_CLASSPATH=/home/student7_lu
```

Run a Jar file on Hadoop:

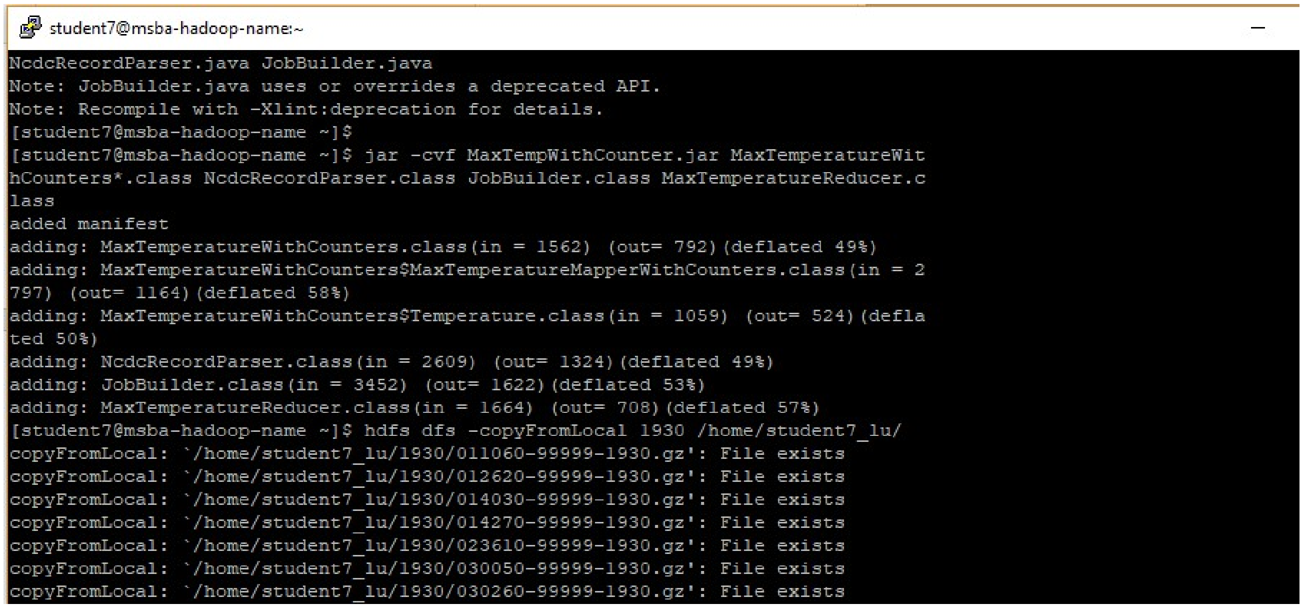
```
hadoop jar MaxTempWithCounter.jar MaxTemperatureWithCounters /home/student7_lu/1930 /home/student7_lu/output1930
```

List the output folder

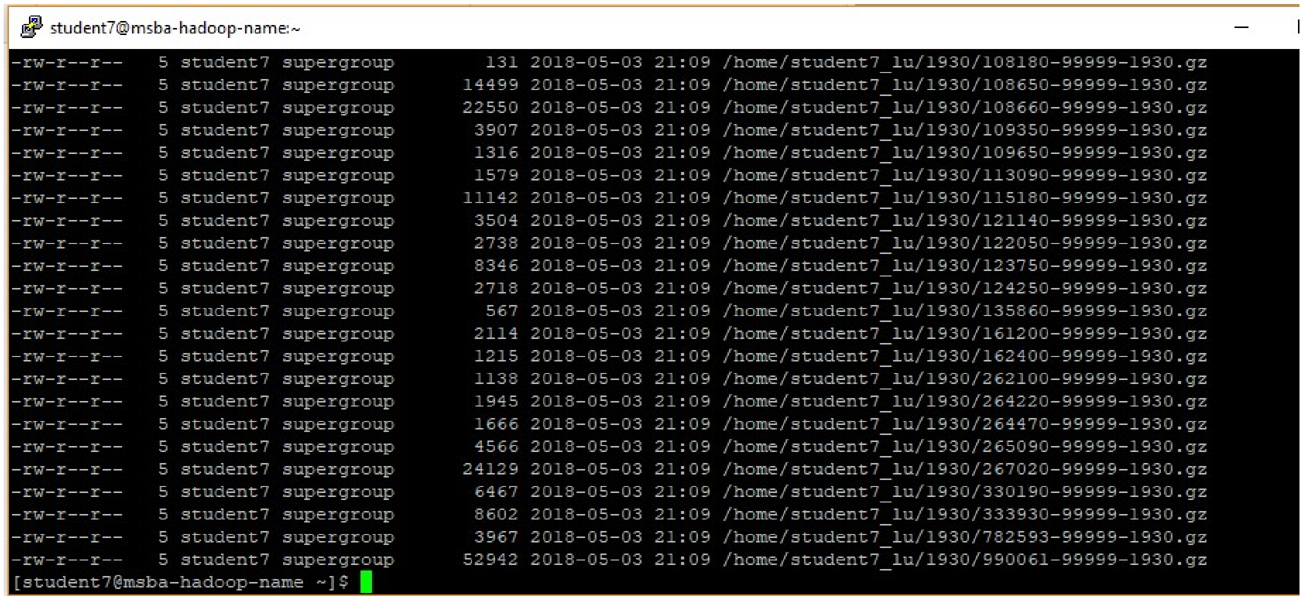
```
hdfs dfs -ls  /home/student7_lu/output1930/
```

Display the output

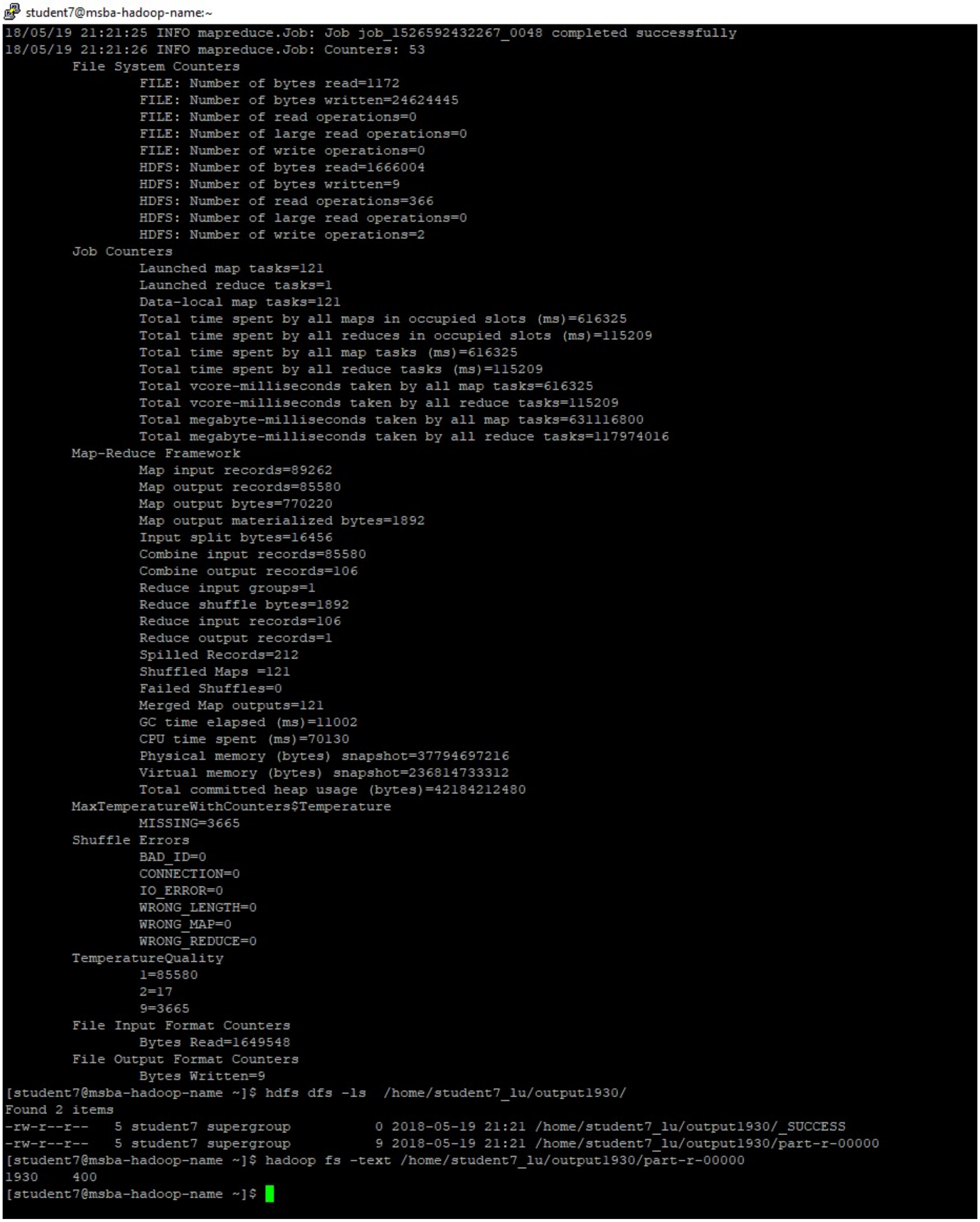
```
hadoop fs -text /home/student7_lu/output1930/part-r-00000
```



```
NcdcRecordParser.java JobBuilder.java
Note: JobBuilder.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
[student7@msba-hadoop-name ~]$
[student7@msba-hadoop-name ~]$ jar -cvf MaxTempWithCounter.jar MaxTemperatureWit
hCounters*.class NcdcRecordParser.class JobBuilder.class MaxTemperatureReducer.c
lass
added manifest
adding: MaxTemperatureWithCounters.class(in = 1562) (out= 792)(deflated 49%)
adding: MaxTemperatureWithCounters$MaxTemperatureMapperWithCounters.class(in = 2
797) (out= 1164)(deflated 58%)
adding: MaxTemperatureWithCounters$Temperature.class(in = 1059) (out= 524)(defla
ted 50%)
adding: NcdcRecordParser.class(in = 2609) (out= 1324)(deflated 49%)
adding: JobBuilder.class(in = 3452) (out= 1622)(deflated 53%)
adding: MaxTemperatureReducer.class(in = 1664) (out= 708)(deflated 57%)
[student7@msba-hadoop-name ~]$ hdfs dfs -copyFromLocal 1930 /home/student7_lu/
copyFromLocal: `/home/student7_lu/1930/011060-99999-1930.gz': File exists
copyFromLocal: `/home/student7_lu/1930/012620-99999-1930.gz': File exists
copyFromLocal: `/home/student7_lu/1930/014030-99999-1930.gz': File exists
copyFromLocal: `/home/student7_lu/1930/014270-99999-1930.gz': File exists
copyFromLocal: `/home/student7_lu/1930/023610-99999-1930.gz': File exists
copyFromLocal: `/home/student7_lu/1930/030050-99999-1930.gz': File exists
copyFromLocal: `/home/student7_lu/1930/030260-99999-1930.gz': File exists
```



```
-rw-r--r--   5 student7 supergroup        131 2018-05-03 21:09 /home/student7_lu/1930/108180-99999-1930.gz
-rw-r--r--   5 student7 supergroup      14499 2018-05-03 21:09 /home/student7_lu/1930/108650-99999-1930.gz
-rw-r--r--   5 student7 supergroup      22550 2018-05-03 21:09 /home/student7_lu/1930/108660-99999-1930.gz
-rw-r--r--   5 student7 supergroup       3907 2018-05-03 21:09 /home/student7_lu/1930/109350-99999-1930.gz
-rw-r--r--   5 student7 supergroup       1316 2018-05-03 21:09 /home/student7_lu/1930/109650-99999-1930.gz
-rw-r--r--   5 student7 supergroup       1579 2018-05-03 21:09 /home/student7_lu/1930/113090-99999-1930.gz
-rw-r--r--   5 student7 supergroup      11142 2018-05-03 21:09 /home/student7_lu/1930/115180-99999-1930.gz
-rw-r--r--   5 student7 supergroup       3504 2018-05-03 21:09 /home/student7_lu/1930/121140-99999-1930.gz
-rw-r--r--   5 student7 supergroup       2738 2018-05-03 21:09 /home/student7_lu/1930/122050-99999-1930.gz
-rw-r--r--   5 student7 supergroup       8346 2018-05-03 21:09 /home/student7_lu/1930/123750-99999-1930.gz
-rw-r--r--   5 student7 supergroup       2718 2018-05-03 21:09 /home/student7_lu/1930/124250-99999-1930.gz
-rw-r--r--   5 student7 supergroup        567 2018-05-03 21:09 /home/student7_lu/1930/135860-99999-1930.gz
-rw-r--r--   5 student7 supergroup       2114 2018-05-03 21:09 /home/student7_lu/1930/161200-99999-1930.gz
-rw-r--r--   5 student7 supergroup       1215 2018-05-03 21:09 /home/student7_lu/1930/162400-99999-1930.gz
-rw-r--r--   5 student7 supergroup       1138 2018-05-03 21:09 /home/student7_lu/1930/262100-99999-1930.gz
-rw-r--r--   5 student7 supergroup       1945 2018-05-03 21:09 /home/student7_lu/1930/264220-99999-1930.gz
-rw-r--r--   5 student7 supergroup       1666 2018-05-03 21:09 /home/student7_lu/1930/264470-99999-1930.gz
-rw-r--r--   5 student7 supergroup       4566 2018-05-03 21:09 /home/student7_lu/1930/265090-99999-1930.gz
-rw-r--r--   5 student7 supergroup      24129 2018-05-03 21:09 /home/student7_lu/1930/267020-99999-1930.gz
-rw-r--r--   5 student7 supergroup       6467 2018-05-03 21:09 /home/student7_lu/1930/330190-99999-1930.gz
-rw-r--r--   5 student7 supergroup       8602 2018-05-03 21:09 /home/student7_lu/1930/333930-99999-1930.gz
-rw-r--r--   5 student7 supergroup       3967 2018-05-03 21:09 /home/student7_lu/1930/782593-99999-1930.gz
-rw-r--r--   5 student7 supergroup      52942 2018-05-03 21:09 /home/student7_lu/1930/990061-99999-1930.gz
[student7@msba-hadoop-name ~]$
```



```
18/05/19 21:21:25 INFO mapreduce.Job: Job job_1526592432267_0048 completed successfully
18/05/19 21:21:26 INFO mapreduce.Job: Counters: 53
        File System Counters
                FILE: Number of bytes read=1172
                FILE: Number of bytes written=24624445
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1666004
                HDFS: Number of bytes written=9
                HDFS: Number of read operations=366
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=121
                Launched reduce tasks=1
                Data-local map tasks=121
                Total time spent by all maps in occupied slots (ms)=616325
                Total time spent by all reduces in occupied slots (ms)=115209
                Total time spent by all map tasks (ms)=616325
                Total time spent by all reduce tasks (ms)=115209
                Total vcore-milliseconds taken by all map tasks=616325
                Total vcore-milliseconds taken by all reduce tasks=115209
                Total megabyte-milliseconds taken by all map tasks=631116800
                Total megabyte-milliseconds taken by all reduce tasks=117974016
        Map-Reduce Framework
                Map input records=89262
                Map output records=85580
                Map output bytes=770220
                Map output materialized bytes=1892
                Input split bytes=16456
                Combine input records=85580
                Combine output records=106
                Reduce input groups=1
                Reduce shuffle bytes=1892
                Reduce input records=106
                Reduce output records=1
                Spilled Records=212
                Shuffled Maps =121
                Failed Shuffles=0
                Merged Map outputs=121
                GC time elapsed (ms)=11002
                CPU time spent (ms)=70130
                Physical memory (bytes) snapshot=37794697216
                Virtual memory (bytes) snapshot=236814733312
                Total committed heap usage (bytes)=42184212480
        MaxTemperatureWithCounters$Temperature
                MISSING=3665
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        TemperatureQuality
                1=85580
                2=17
                9=3665
        File Input Format Counters
                Bytes Read=1649548
        File Output Format Counters
                Bytes Written=9
[student7@msba-hadoop-name ~]$ hdfs dfs -ls  /home/student7_lu/output1930/
Found 2 items
-rw-r--r--   5 student7 supergroup          0 2018-05-19 21:21 /home/student7_lu/output1930/_SUCCESS
-rw-r--r--   5 student7 supergroup          9 2018-05-19 21:21 /home/student7_lu/output1930/part-r-00000
[student7@msba-hadoop-name ~]$ hadoop fs -text /home/student7_lu/output1930/part-r-00000
1930    400
[student7@msba-hadoop-name ~]$ 
```

3. Work “A Load UDF” example on pages 396 and 397. Type out all the commands in each step of the process and print out a screenshot of the final results in CSUEB Hadoop.

Copy the input to HDFS

```
hdfs dfs -copyFromLocal /home/student7/sample.txt /home/student7/
```

Compile the Java files

```
javac -classpath /home/student7/hadoop-common-2.6.1.jar:/home/student7/hadoop-mapreduce-client-core-2.6.1.jar:/home/student7/commons-cli-2.0.jar:/home/student7/pig-0.11.0.jar:/home/student7/commons-logging-1.2.jar -d . Range.java CutLoadFunc.java
```

Create the Jar files

```
jar -cvf pig-cut.jar com/hadoopbook/pig/CutLoadFunc.class
```

```
jar -cvf pig-range.jar com/hadoopbook/pig/Range.class
```

Enter grunt

```
pig -x local
```

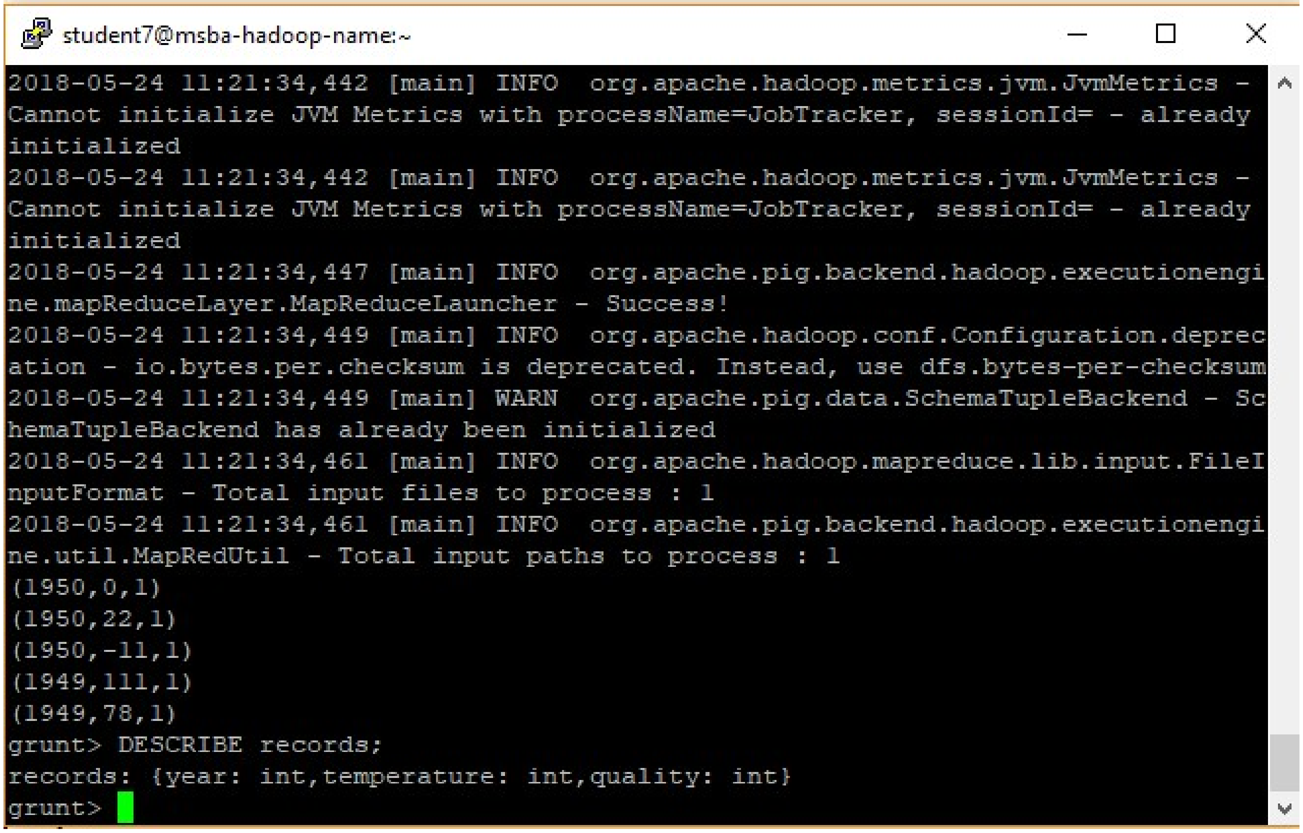
Load records in grunt

```
records = LOAD 'sample.txt' USING com.hadoopbook.pig.CutLoadFunc('16-19,88-92,93-93')
AS(year:int,temperature:int,quality:int);
```

Dump the records

```
DUMP records;
```

```
DESCRIBE records;
```



```
2018-05-24 11:21:34,442 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2018-05-24 11:21:34,442 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2018-05-24 11:21:34,447 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2018-05-24 11:21:34,449 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-05-24 11:21:34,449 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2018-05-24 11:21:34,461 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input files to process : 1
2018-05-24 11:21:34,461 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(1950,0,1)
(1950,22,1)
(1950,-11,1)
(1949,111,1)
(1949,78,1)
grunt> DESCRIBE records;
records: {year: int,temperature: int,quality: int}
grunt> 
```