# The Course Project

The course project includes 3 parts. The first part is to develop a Mapper and Reducer application to retrieve Year and Temperature from original NCDC records (National climatic Data Center) and then write the Year and Temperature data into **a text file**. The second part is to load the text file **into Pig** and get the **highest and lowest** temperatures for each year. The third part is **to load the text file into Hive** and get **the average temperature** for each year.

You need to turn in 1) **the three java files** (mapper, reducer and main), 2) the commands from converting them into a Jar file to **running the Jar file** in Hadoop, 3) the text file including Year and Temperature data created by you, 4) the screenshot of **the text file** being created, 5) the screenshot of **the final Pig output** showing the year and **the highest and lowest** temperatures, and 6) the screenshot of **the final Hive output showing the year and average** temperature.

**First Part:**

**1. Compiling Java file:**

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 . MaxTemperature.java MaxTemperatureMapper.java MaxTemperatureReducer.java

**2. Create a Jar file:**

jar -cvf max-temperature.jar ./MaxTemperature*.class

**3. Copy the folder from Local Disk to HDFS**

hdfs dfs -copyFromLocal ITM6273-CourseProjectData /home/student7/

**4. Run the Jar file on Hadoop:**

hadoop jar max-temperature.jar MaxTemperature /home/student7/ITM6273-CourseProjectData /home/student7/output888/

**5. List the output result:**

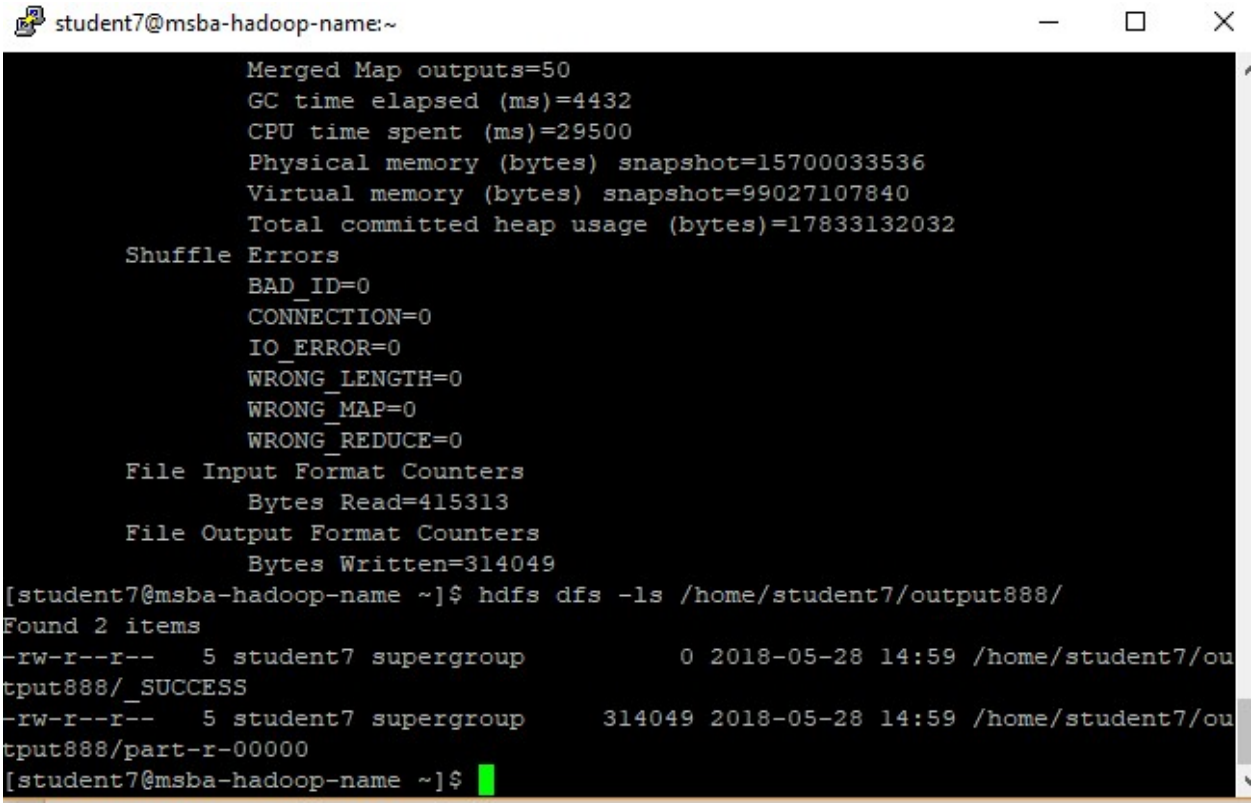
hdfs dfs -ls /home/student7/output888/

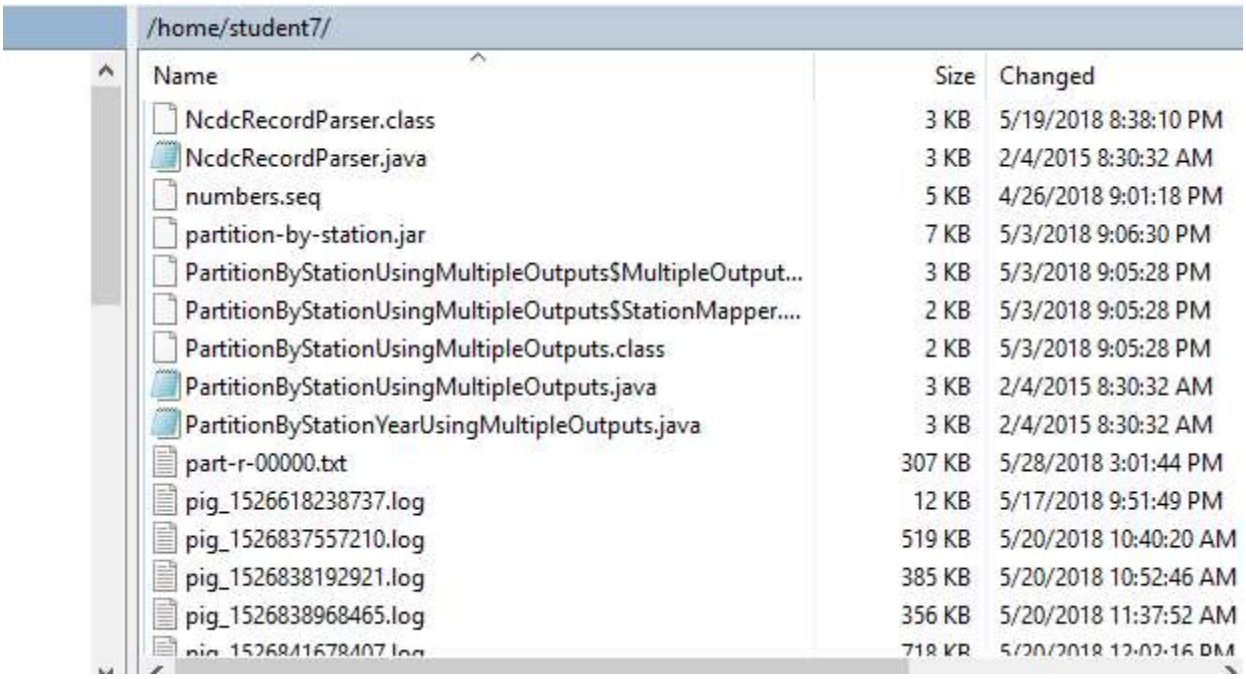
hdfs dfs -cat /home/student7/output888/part-r-00000

**6. Export the output to local disk and transfer into text file:**

hdfs dfs -copyToLocal /home/student7/output888/part-r-00000 /home/student7/part-r-00000.txt


student7@msba-hadoop-name:~
1930 28
1930 28
1930 28
1930 22
1930 39
1930 22
1930 11
1930 11
1930 50
1930 50
1930 11
1930 28
1930 22
1930 28
1930 0
1930 -11
1930 11
1930 0
1930 22
1930 11
1930 -22
1930 -22
1930 -22
[student7@msba-hadoop-name ~]$


student7@msba-hadoop-name:~
Merged Map outputs=50
GC time elapsed (ms)=4432
CPU time spent (ms)=29500
Physical memory (bytes) snapshot=15700033536
Virtual memory (bytes) snapshot=99027107840
Total committed heap usage (bytes)=17833132032
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=415313
File Output Format Counters
Bytes Written=314049
[student7@msba-hadoop-name ~]$ hdfs dfs -ls /home/student7/output888/
Found 2 items
-rw-r--r-- 5 student7 supergroup 0 2018-05-28 14:59 /home/student7/output888/_SUCCESS
-rw-r--r-- 5 student7 supergroup 314049 2018-05-28 14:59 /home/student7/output888/part-r-00000
[student7@msba-hadoop-name ~]$




**Second Part:**

**1. Login to pig:**

pig –x local

**2. Load in the file('part-r-00000.txt')**

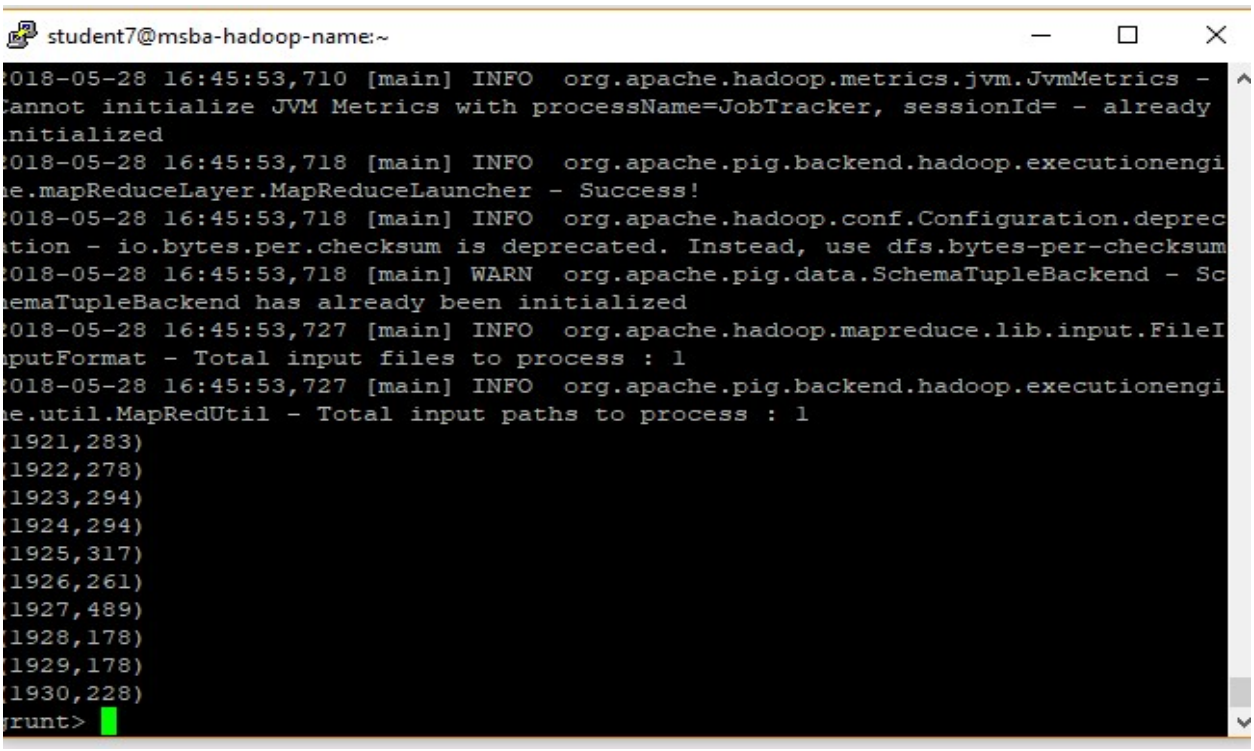
```
records = LOAD 'part-r-00000.txt'

AS (year:chararray, temperature:int);

DUMP records;
```

**3. group the records by year**

```
grouped_records = GROUP records BY year;

DUMP grouped_records;
```

**4. select the max temperature for each year and print out the result**

```
max_temp = FOREACH grouped_records GENERATE group,

MAX(records.temperature);

DUMP max_temp;
```

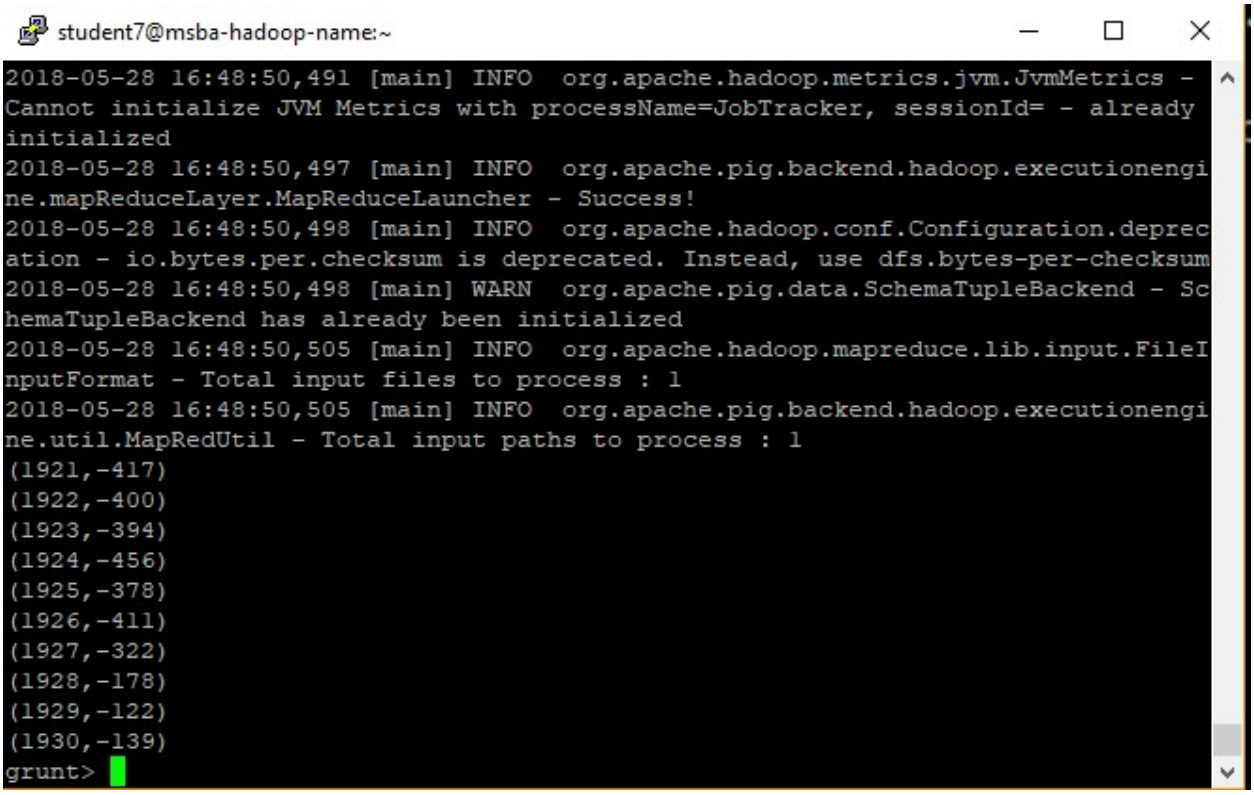



**5. select the lowest temperature for each year and print out the result**

Min_temp = FOREACH grouped_records GENERATE group,

MIN(records.temperature);

DUMP Min_temp;




**Third Part:**

**1. Enter into Hive**

hive

**2. Create the table with data**

DROP TABLE IF EXISTS Temperature;

CREATE TABLE Temperature (year STRING, temperature INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t';

**3. Populate Hive with the data**

LOAD DATA LOCAL INPATH 'part-r-00000.txt'

OVERWRITE INTO TABLE Temperature;

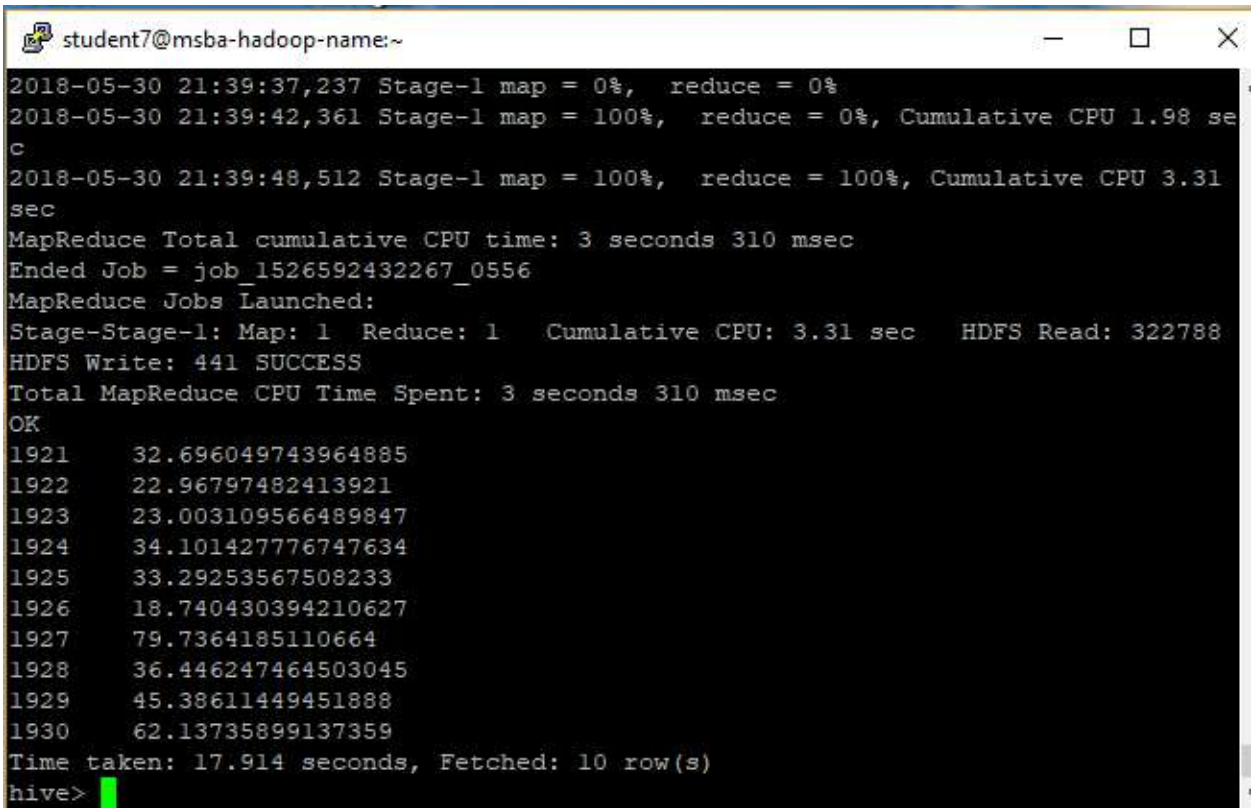
**4. Average temperature for each year**

SELECT year, AVG(temperature)

FROM Temperature

GROUP BY year

SORT BY  year;




Conclusion:

I  obtained the highest and lowest temperature and average temperature for each year by loading  data from original NCDC records into pig and hive in Hadoop.