

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 ~]$
```

/home/student7/		
Name	Size	Changed
NcdcRecordParser.class	3 KB	5/19/2018 8:38:10 PM
NcdcRecordParser.java	3 KB	2/4/2015 8:30:32 AM
numbers.seq	5 KB	4/26/2018 9:01:18 PM
partition-by-station.jar	7 KB	5/3/2018 9:06:30 PM
PartitionByStationUsingMultipleOutputs\$MultipleOutput...	3 KB	5/3/2018 9:05:28 PM
PartitionByStationUsingMultipleOutputs\$StationMapper....	2 KB	5/3/2018 9:05:28 PM
PartitionByStationUsingMultipleOutputs.class	2 KB	5/3/2018 9:05:28 PM
PartitionByStationUsingMultipleOutputs.java	3 KB	2/4/2015 8:30:32 AM
PartitionByStationYearUsingMultipleOutputs.java	3 KB	2/4/2015 8:30:32 AM
part-r-00000.txt	307 KB	5/28/2018 3:01:44 PM
pig_1526618238737.log	12 KB	5/17/2018 9:51:49 PM
pig_1526837557210.log	519 KB	5/20/2018 10:40:20 AM
pig_1526838192921.log	385 KB	5/20/2018 10:52:46 AM
pig_1526838968465.log	356 KB	5/20/2018 11:37:52 AM
pig_1526841678407.log	718 KB	5/20/2018 12:02:16 PM

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;`

```
student7@msba-hadoop-name:~  
2018-05-28 16:45:53,710 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics -  
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already  
initialized  
2018-05-28 16:45:53,718 [main] INFO org.apache.pig.backend.hadoop.executionengi  
ne.mapReduceLayer.MapReduceLauncher - Success!  
2018-05-28 16:45:53,718 [main] INFO org.apache.hadoop.conf.Configuration.deprec  
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
2018-05-28 16:45:53,718 [main] WARN org.apache.pig.data.SchemaTupleBackend - Sc  
emaTupleBackend has already been initialized  
2018-05-28 16:45:53,727 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileI  
nputFormat - Total input files to process : 1  
2018-05-28 16:45:53,727 [main] INFO org.apache.pig.backend.hadoop.executionengi  
ne.util.MapRedUtil - Total input paths to process : 1  
(1921,283)  
(1922,278)  
(1923,294)  
(1924,294)  
(1925,317)  
(1926,261)  
(1927,489)  
(1928,178)  
(1929,178)  
(1930,228)  
grunt>
```

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;

```
student7@msba-hadoop-name:~  
2018-05-28 16:48:50,491 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics -  
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already  
initialized  
2018-05-28 16:48:50,497 [main] INFO org.apache.pig.backend.hadoop.executionengi  
ne.mapReduceLayer.MapReduceLauncher - Success!  
2018-05-28 16:48:50,498 [main] INFO org.apache.hadoop.conf.Configuration.deprec  
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
2018-05-28 16:48:50,498 [main] WARN org.apache.pig.data.SchemaTupleBackend - Sc  
hemaTupleBackend has already been initialized  
2018-05-28 16:48:50,505 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileI  
nputFormat - Total input files to process : 1  
2018-05-28 16:48:50,505 [main] INFO org.apache.pig.backend.hadoop.executionengi  
ne.util.MapRedUtil - Total input paths to process : 1  
(1921,-417)  
(1922,-400)  
(1923,-394)  
(1924,-456)  
(1925,-378)  
(1926,-411)  
(1927,-322)  
(1928,-178)  
(1929,-122)  
(1930,-139)  
grunt>
```

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;

```
student7@msba-hadoop-name:~  
2018-05-30 21:39:37,237 Stage-1 map = 0%, reduce = 0%  
2018-05-30 21:39:42,361 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.98 sec  
2018-05-30 21:39:48,512 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.31 sec  
MapReduce Total cumulative CPU time: 3 seconds 310 msec  
Ended Job = job_1526592432267_0556  
MapReduce Jobs Launched:  
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.31 sec HDFS Read: 322788  
HDFS Write: 441 SUCCESS  
Total MapReduce CPU Time Spent: 3 seconds 310 msec  
OK  
1921 32.696049743964885  
1922 22.96797482413921  
1923 23.003109566489847  
1924 34.101427776747634  
1925 33.29253567508233  
1926 18.740430394210627  
1927 79.7364185110664  
1928 36.446247464503045  
1929 45.38611449451888  
1930 62.13735899137359  
Time taken: 17.914 seconds, Fetched: 10 row(s)  
hive>
```

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.