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 (i.e., the dataset we are using for this class) 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.

The original dataset for this project is available on Blackboard.

**PART 1**

Preliminary commands and steps:

*Important jars - hadoop-common-2.6.1.jar, hadoop-mapreduce-client-core-2.6.1.jar, commons-cli-2.0.jar- were placed into the local drive.*

*The 50 weather datasets were extracted into a folder directory named “WeatherData”, and then placed into the local drive.*

*User created java files –AllTemperature.java, AllTemperatureMapper.java, AllTemperatureReducer – were placed into the local drive.*

export HADOOP\_CLASSPATH=/home/student9/

Command for compiling the java files:

javac -cp src/:hadoop-common-2.6.1.jar:hadoop-mapreduce-client-core-2.6.1.jar:commons-cli-2.0.jar -d . AllTemperature.java AllTemperatureReducer.java AllTemperatureMapper.java

Command for compiling the jar:

jar -cvf all-temperature.jar AllTemperature\*.class

Command for exporting WeatherData folder into hdfs:

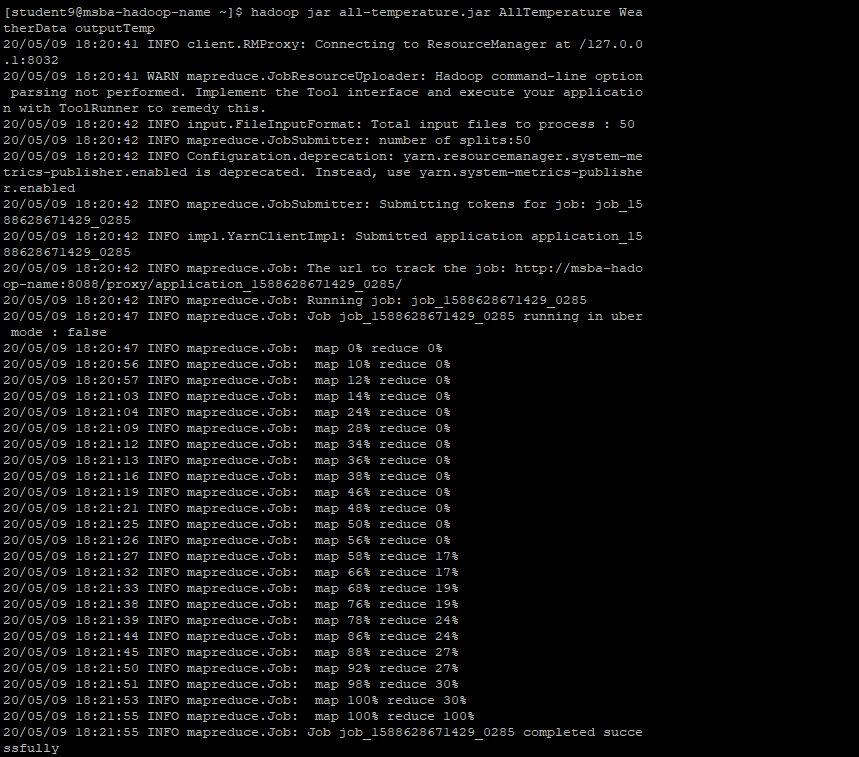
hdfs dfs -copyFromLocal WeatherData

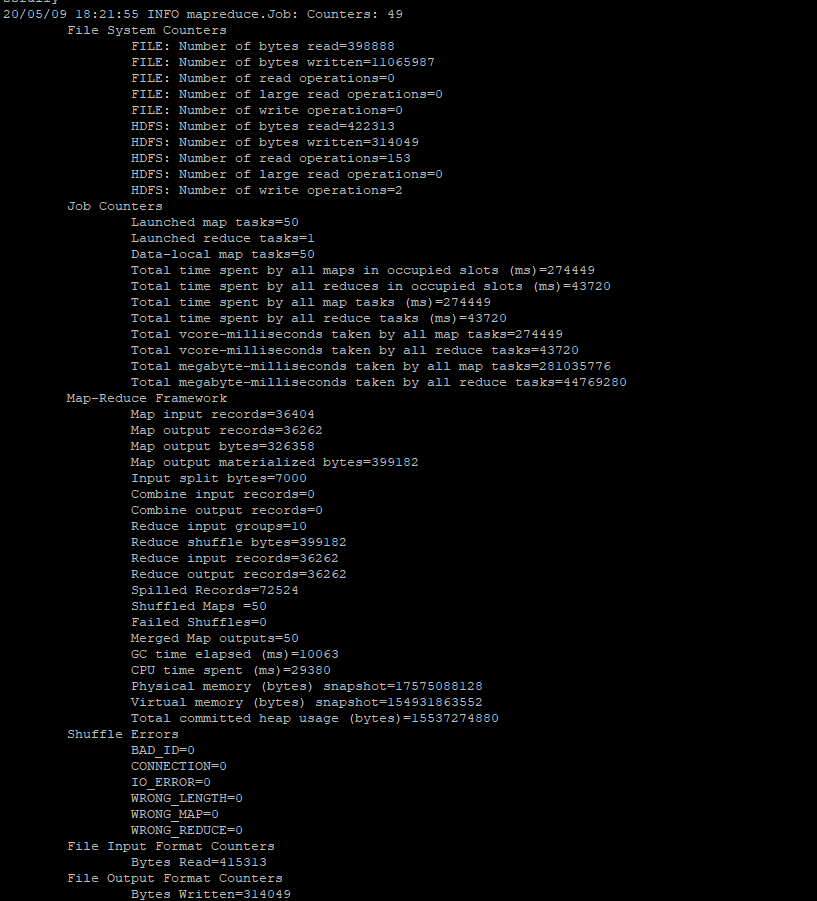
*(At this point, I deleted the three class files)*

Command for running the jar:

hadoop jar all-temperature.jar AllTemperature WeatherData outputTemp

Screenshots of the text file being created for the ncdc weather data:





Command and screenshot for creating the .txt file and then sending it to the local drive:

hdfs dfs -cat outputTemp/part-r-00000 >/home/student9/ExtractedWeatherData.txt



**PART 2**

Preliminary command for pig:

*(ExtractedWeatherData.txt is already in local drive)*

pig -x local

Commands for loading the year and weather data into a table “WeatherData”:

grunt> WeatherData = LOAD 'ExtractedWeatherData.txt'

>> AS (year:chararray, temperature:int);

Commands for isolating the maximum and the minimum temperature for each year using pig:

grunt> filtered\_WeatherData = Filter WeatherData By temperature != 9999;

grunt> grouped\_WeatherData = GROUP filtered\_WeatherData BY year;

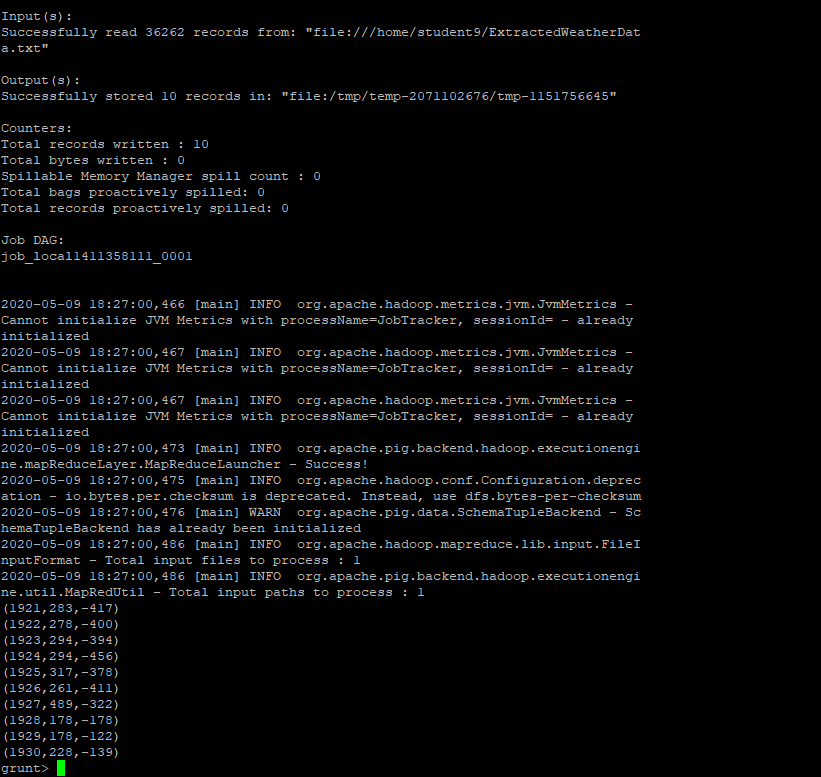
grunt> maxmin\_temp = FOREACH grouped\_WeatherData GENERATE group,

>> MAX(filtered\_WeatherData.temperature),

>> MIN(filtered\_WeatherData.temperature);

grunt> DUMP maxmin\_temp

Screenshot of the maximum and minimum temperature for each year in pig:

****

**PART 3**

Preliminary commands and steps:

*(ExtractedWeatherData.txt is already in the local drive)*

hive

Commands for creating the table “WeatherData”:

hive> CREATE TABLE WeatherData(year STRING, temperature INT)

> ROW FORMAT DELIMITED

> FIELDS TERMINATED BY '\t';

Commands for populating the WeatherData table with the data from ExtractedWeatherData.txt:

hive> LOAD DATA LOCAL INPATH 'ExtractedWeatherData.txt'

> OVERWRITE INTO TABLE WeatherData;

Commands for selecting average temperature for each year:

hive> SELECT year, AVG(temperature)

> FROM WeatherData

> WHERE temperature != 9999

> GROUP BY year;

Screenshot of the average temperature for each year in hive:

