**BAN 632**

**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.**

**Part 1:**

javac -classpath /home/student38/hadoop-common-2.6.1.jar:/home/student38/hadoop-mapreduce-client-core-2.6.1.jar:/home/student38/commons-cli-2.0.jar -d . YearTemperatureVrushali.java YearTemperatureVrushaliMapper.java YearTemperatureVrushaliReducer.java

jar -cvf YearTemperatureVrushali.jar ./YearTemperatureVrushali\*.class

hadoop fs -mkdir /user/c38student38

hdfs dfs -copyFromLocal CourseProjectData /user/c38student38/CourseProjectData

hadoop jar /home/student38/YearTemperatureVrushali.jar YearTemperatureVrushali /user/c38student38/CourseProjectData /user/c38student38/Project\_Output/

hdfs dfs -ls /user/c38student38/Project\_Output/

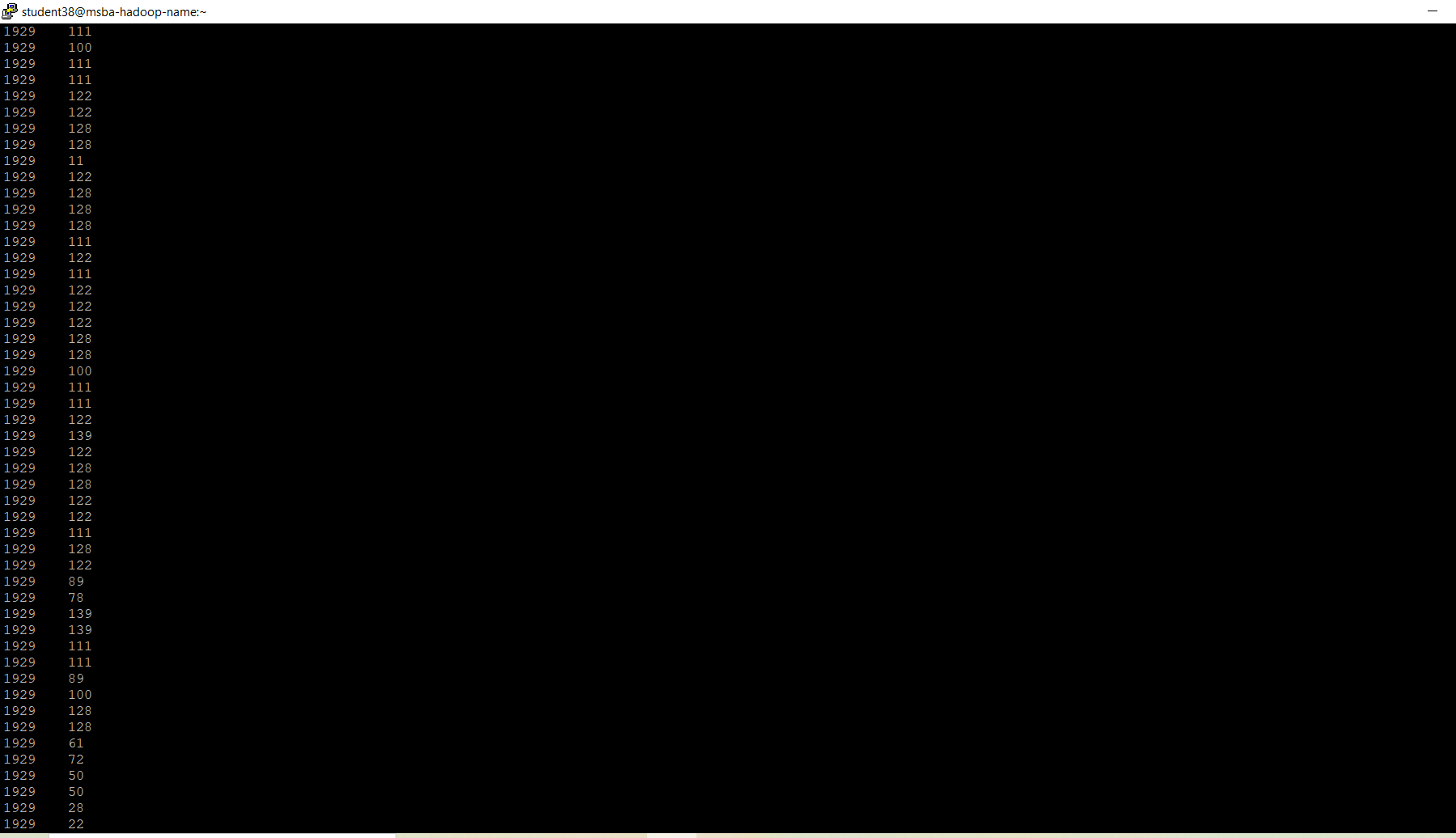
hdfs dfs -cat /user/c38student38/Project\_Output/part-r-00000

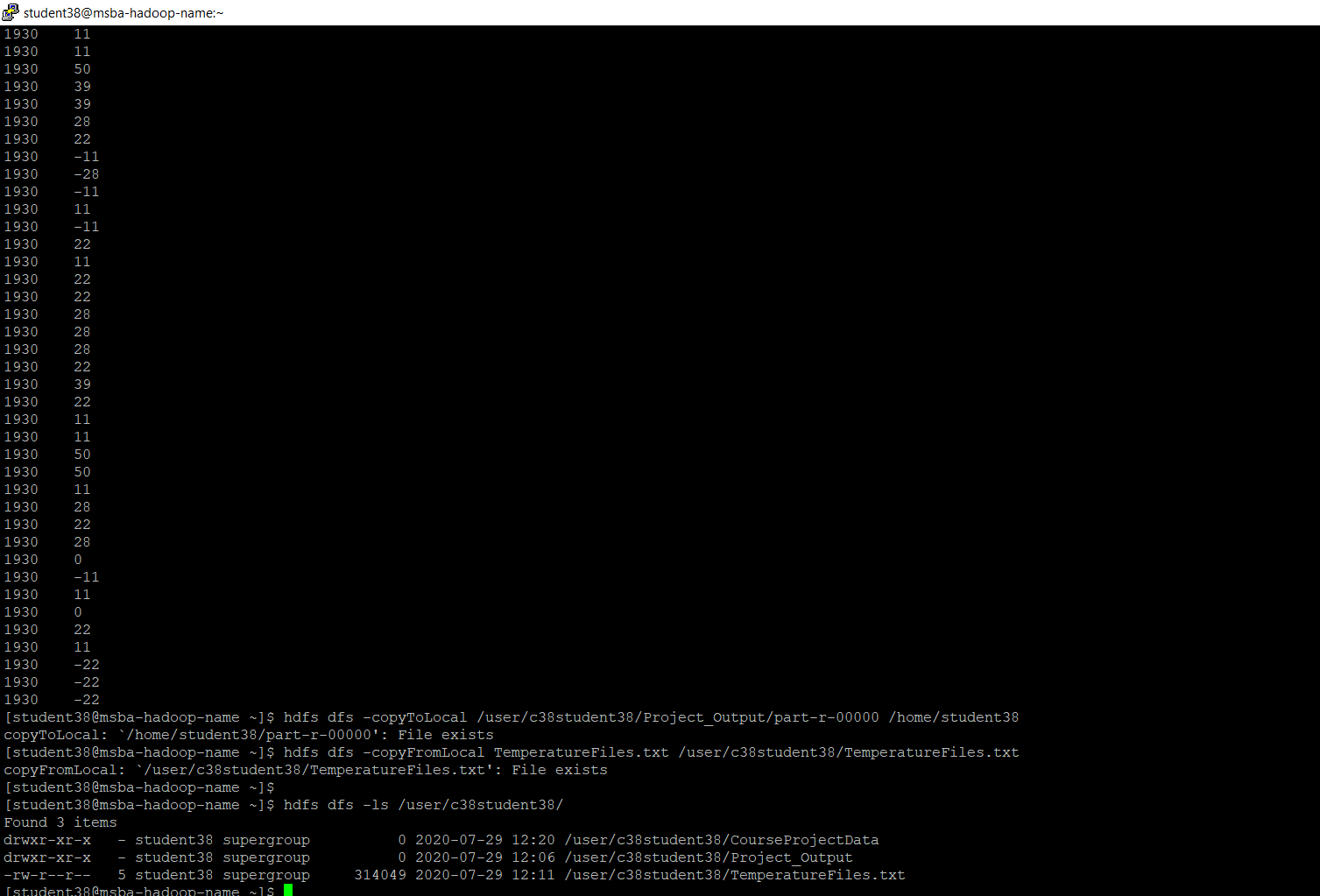
hdfs dfs -copyToLocal /user/c38student38/Project\_Output/part-r-00000 /home/student38

hdfs dfs -copyFromLocal TemperatureFiles.txt /user/c38student38/TemperatureFiles.txt

hdfs dfs -ls /user/c38student38/

**Output:**





**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**The second part is to load the text file into Pig and get the highest and lowest temperatures for each year.**

**Part 2:**

pig -x local

record = LOAD 'TemperatureFiles.txt' AS (year:chararray, temperature:int);

record\_filtered = FILTER record BY temperature != 9999;

DUMP record\_filtered;

record\_grouped = GROUP record\_filtered BY year;

DUMP record\_grouped;

maximum\_temperature = FOREACH record\_grouped GENERATE group, MAX(record\_filtered.temperature);

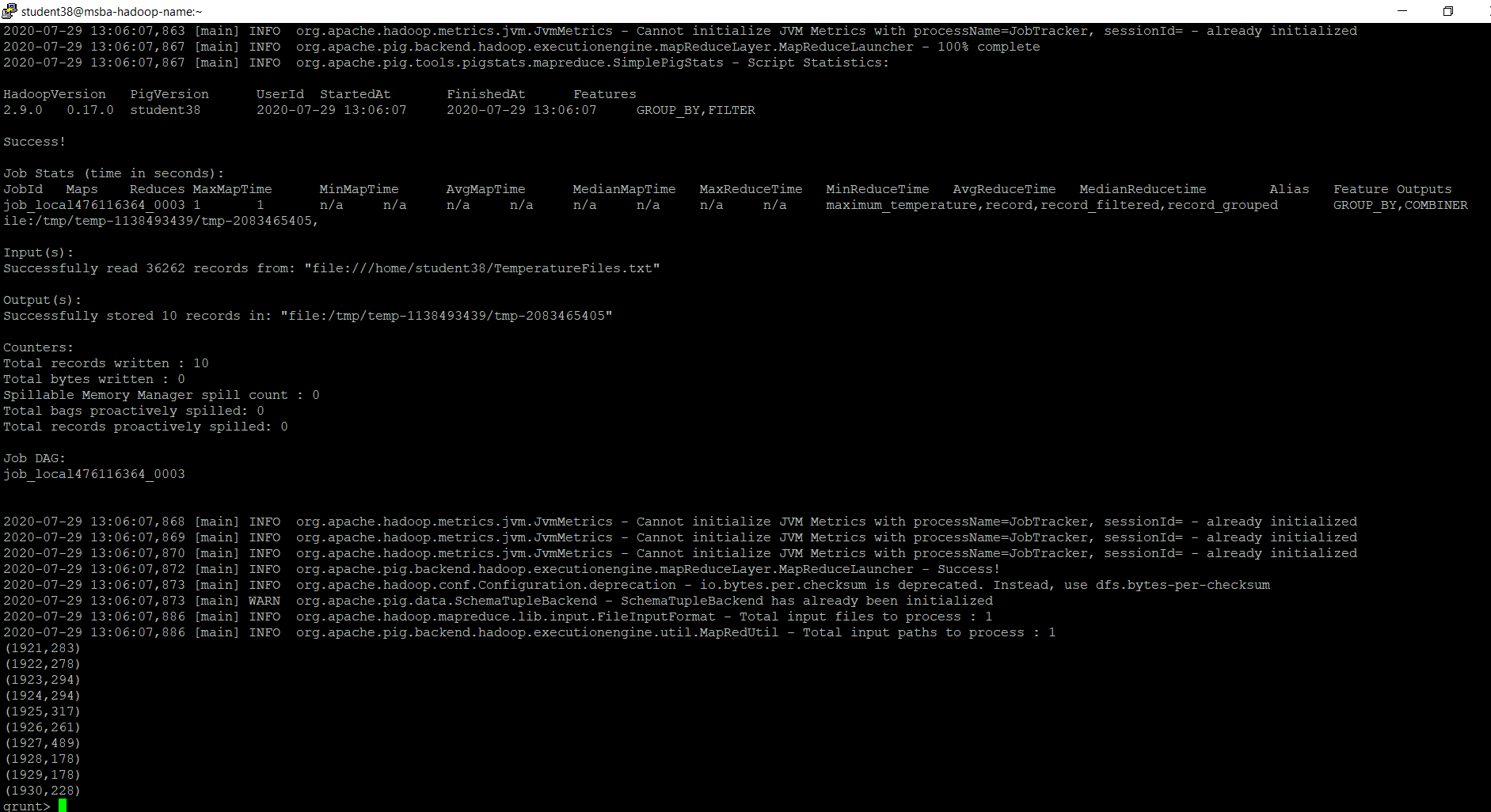
DUMP maximum\_temperature;

minimum\_temperature = FOREACH record\_grouped GENERATE group, MIN(record\_filtered.temperature);

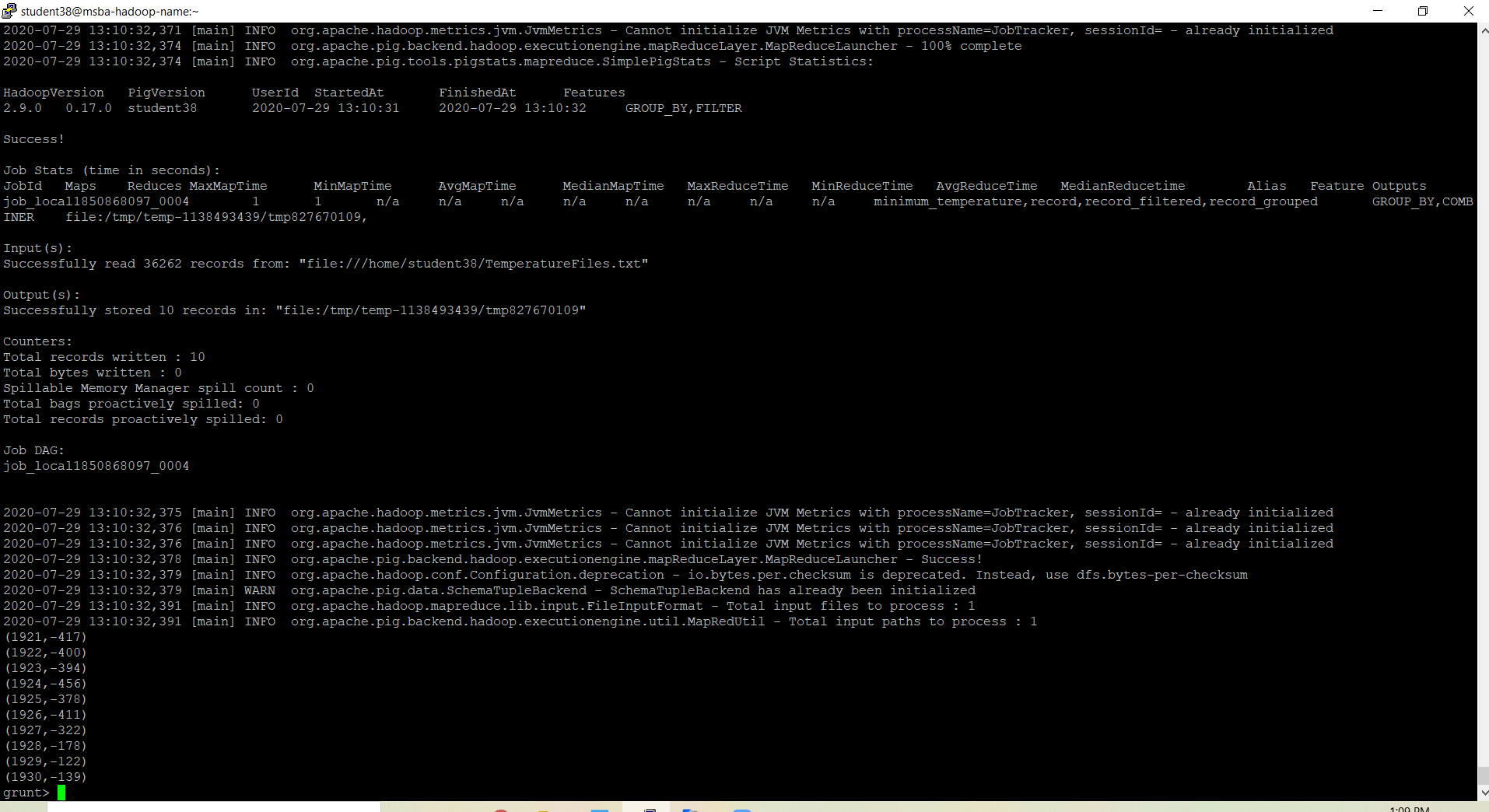
DUMP minimum\_temperature;

**Output:**

**Maximum Temperature**:



**Minimum Temperature:**



**The third part is to load the text file into Hive and get the average temperature for each year.**

**Part 3:**

ls -l | grep meta

mv metastore\_db metastore\_db.old

schematool -dbType derby -initSchema

hive

Drop table if exists project\_records\_student38;

CREATE TABLE project\_records\_student38 (year STRING, temperature INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t';

load data local inpath 'TemperatureFiles.txt' overwrite into table project\_records\_student38;

SELECT year, round(AVG(temperature),2) FROM project\_records\_student38 WHERE temperature != 9999 GROUP BY year;

**Output:**

