Final Project

BAN\_632 -Big data tech. & aPPs

**Objective**:

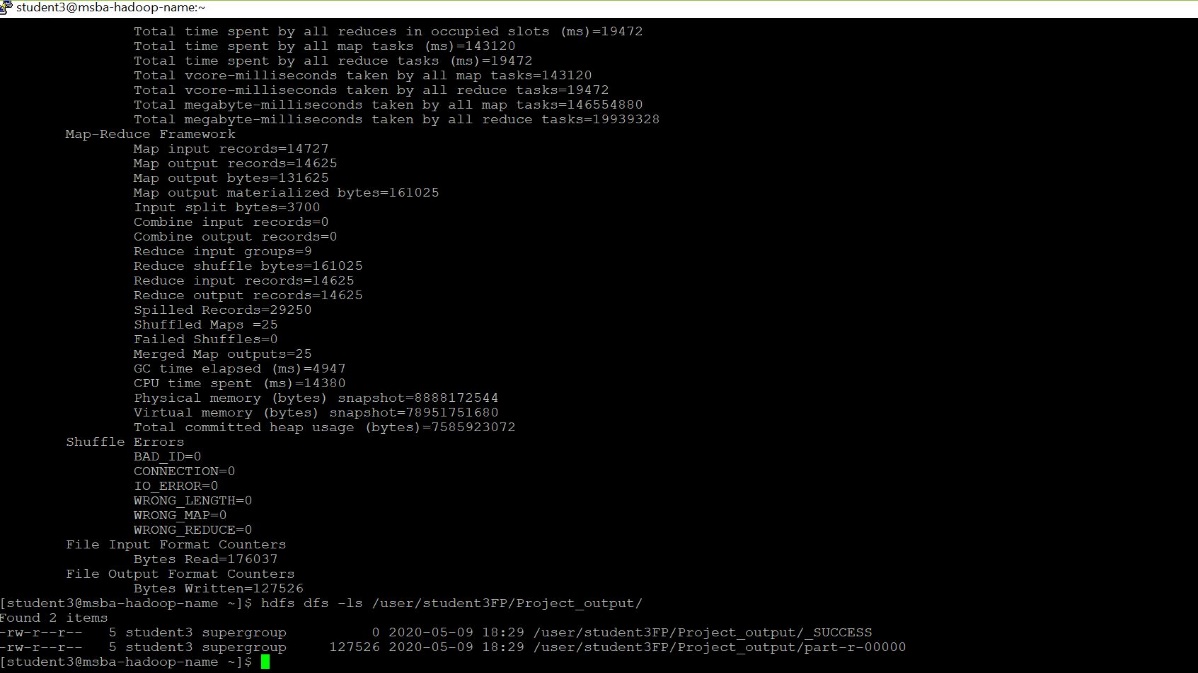
The course project includes 3 parts:

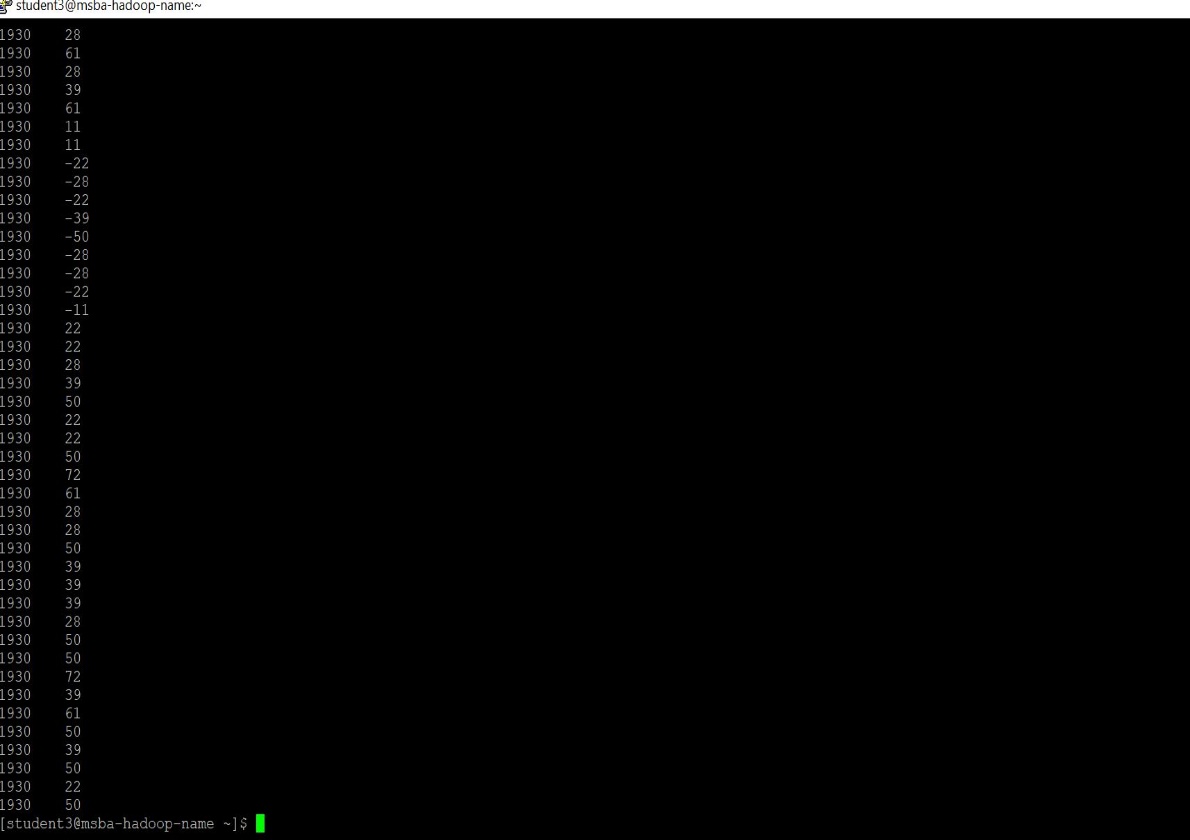
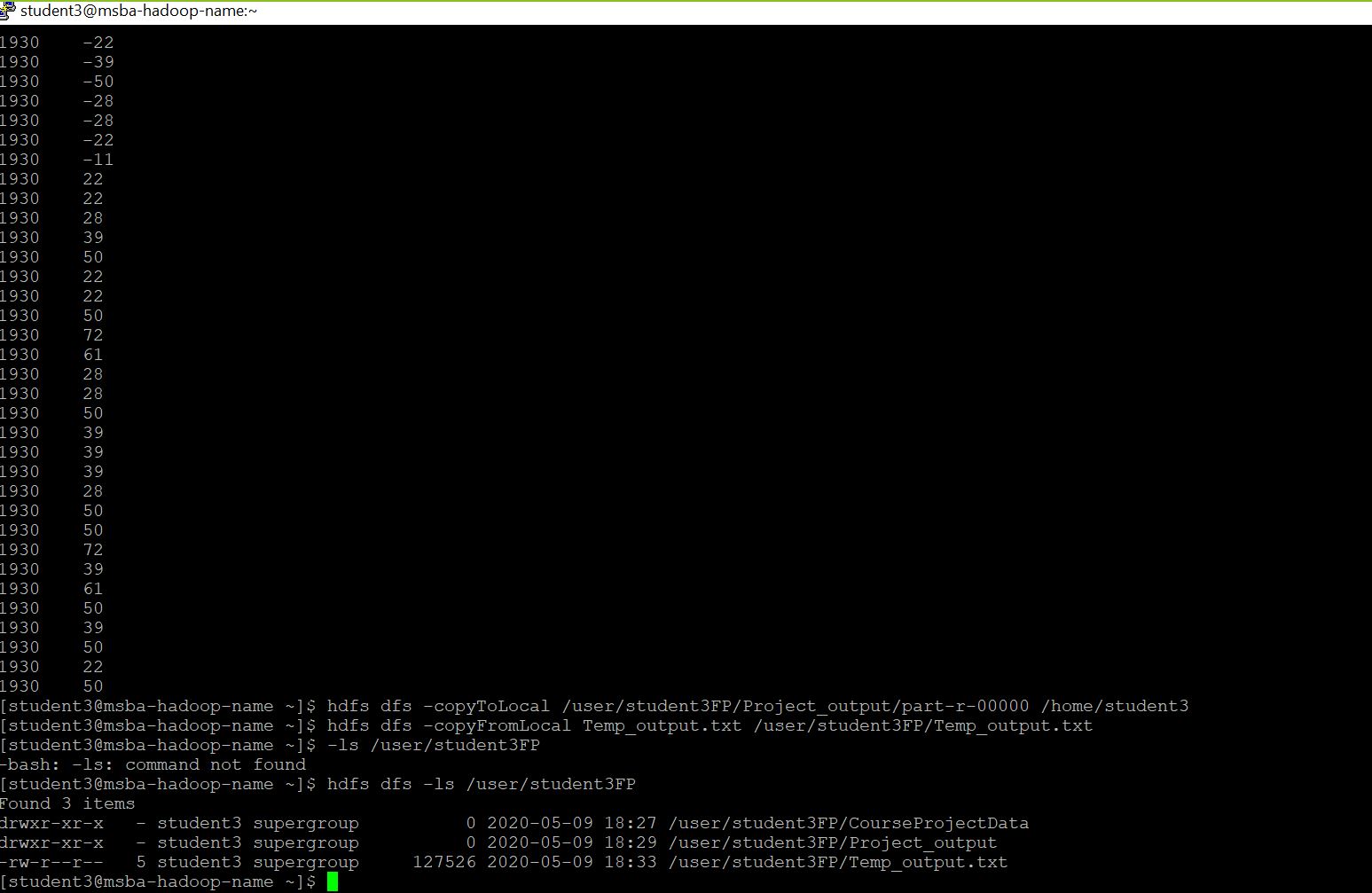
* Develop a Mapper and Reducer application to retrieve Year and Temperature from original NCDC records and then write the Year and Temperature data into a text file.
* Load the text file into Pig and get the highest and lowest temperatures for each year.
* Load the text file into Hive and get the average temperature for each year.

**Part 1:**

**Commands Used and their results (Screenshots):**

* javac -cp /home/student3/hadoop-common-2.6.1.jar:/home/student3/hadoop-mapreduce-client-core-2.6.1.jar:/home/student3/commons-cli-2.0.jar -d . YearTemperatureSruthi.java YearTemperatureSruthiMapper.java YearTemperatureSruthiReducer.java
* jar -cvf YearTemperatureSruthi.jar ./YearTemperatureSruthi\*.class
* hadoop fs -mkdir /user/student3FP
* hdfs dfs -copyFromLocal CourseProjectData /user/student3FP/CourseProjectData
* hadoop jar /home/student3/YearTemperatureSruthi.jar YearTemperatureSruthi /user/student3FP/CourseProjectData /user/student3FP/Project \_output/
* hdfs dfs -ls /user/student3FP/Project \_output/



* hdfs dfs -cat /user/student3FP/Project \_output/part-r-00000
* hdfs dfs -copyToLocal /user/student3FP/Project \_output/part-r-00000 /home/student3
* hdfs dfs -copyFromLocal Temp\_output.txt /user/student3FP/Temp\_output.txt
* 
* 



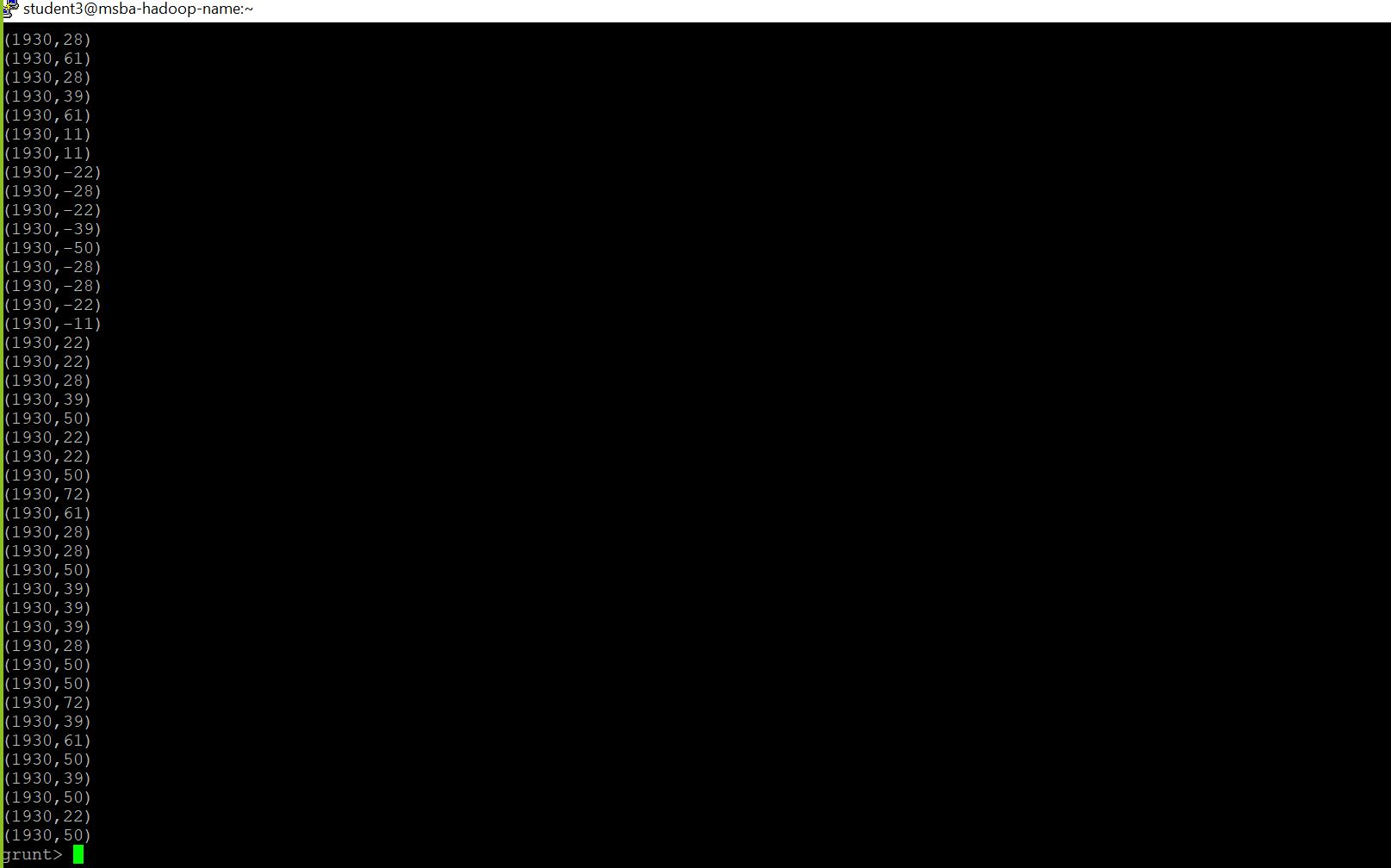
**Part:2**

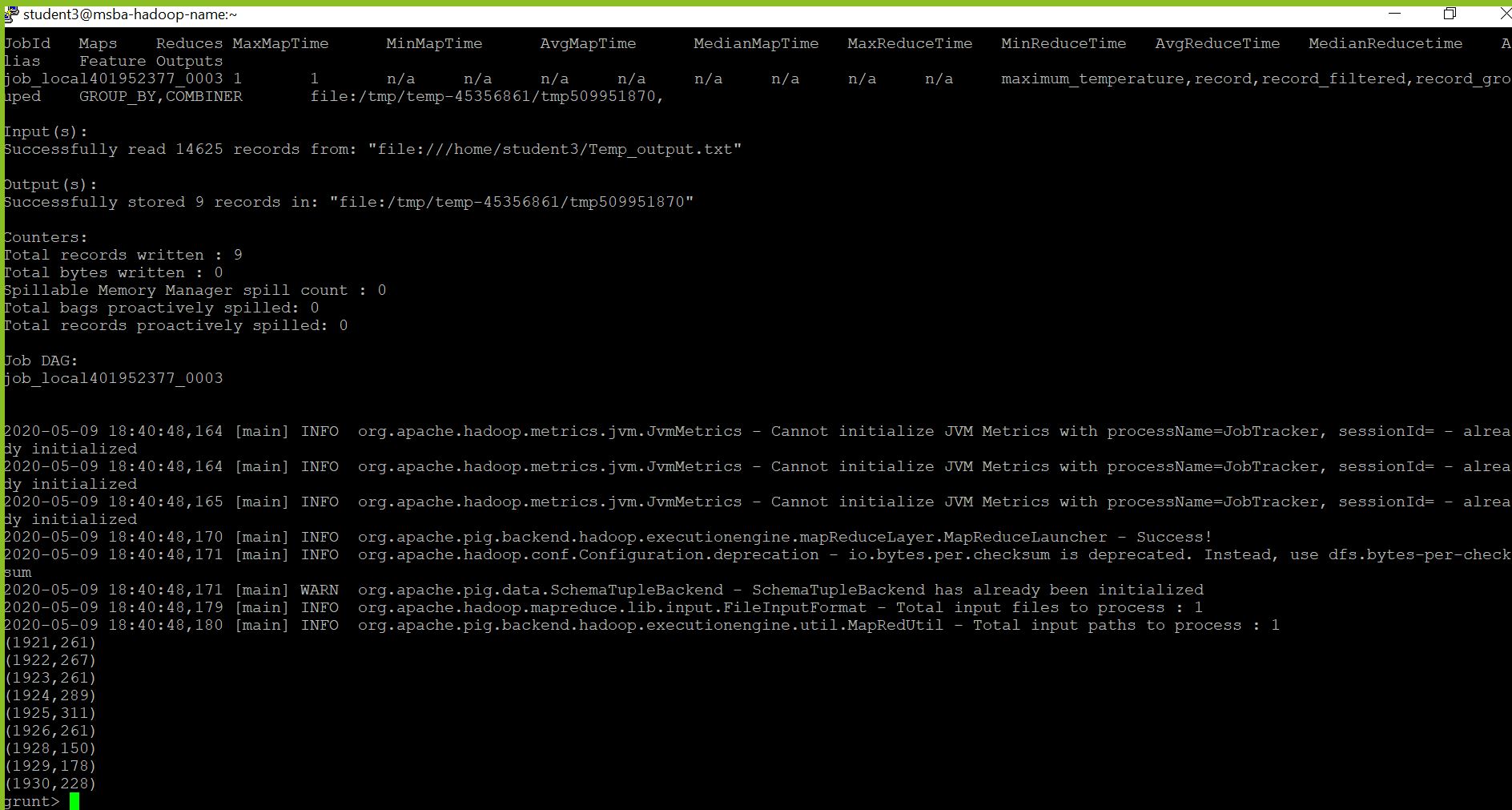
**Commands used:**

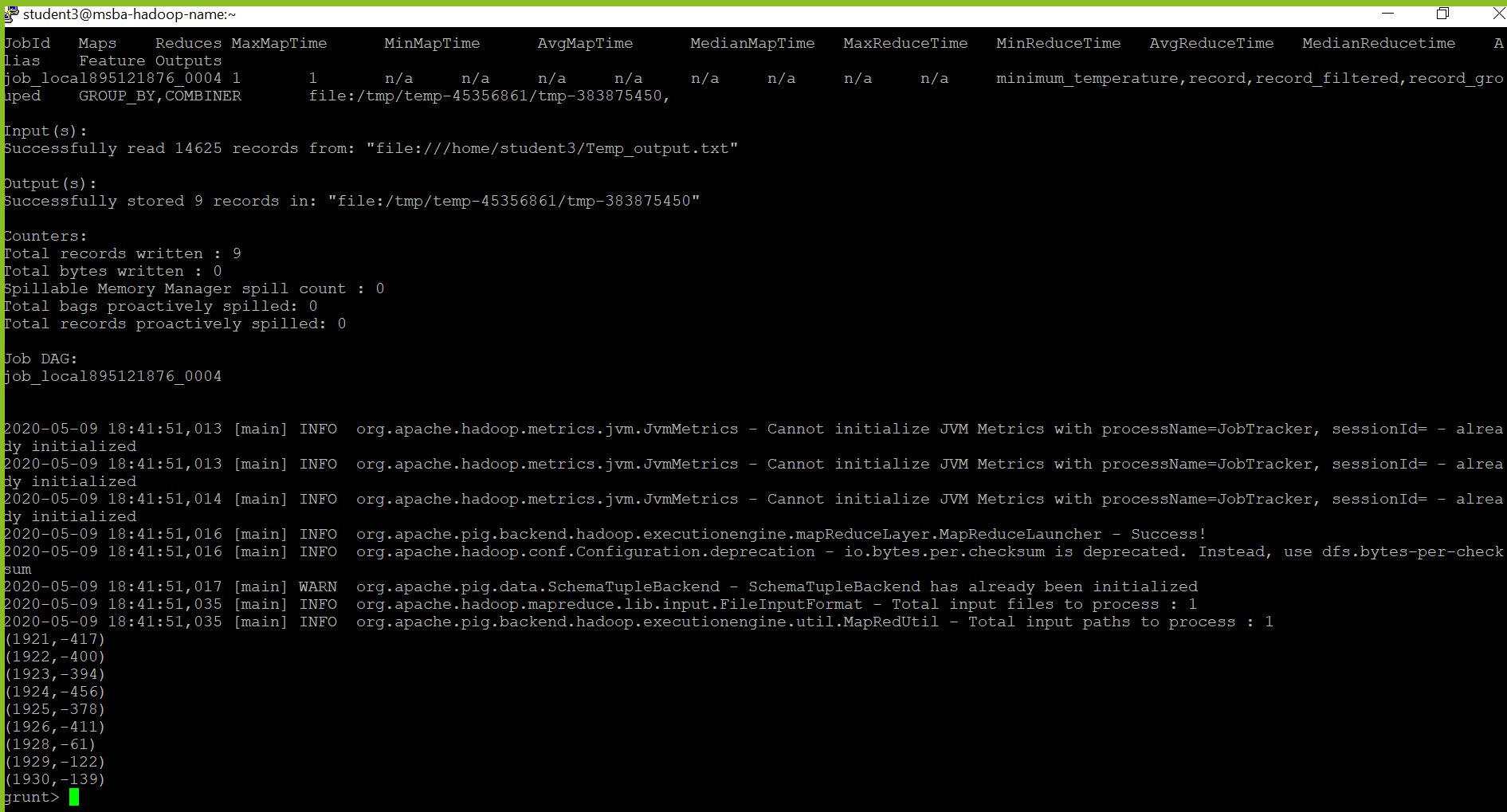
Now going into pig,

* pig -x local
* record = LOAD 'Temp\_output.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;

**Result Screenshots:**







**Part:3**

**Commands used:**

In hive,

* ls -l | grep meta
* mv metastore\_db metastore\_db.old
* schematool -dbType derby -initSchema
* hive
* Drop table if exists records03\_student3;
* CREATE TABLE records03\_student3 (year STRING, temperature INT)
* ROW FORMAT DELIMITED
* FIELDS TERMINATED BY '\t';
* load data local inpath 'Temp\_output.txt' overwrite into table records03\_student3;
* SELECT year, round(AVG(temperature),2) FROM records03\_student3 WHERE temperature != 9999 GROUP BY year;

**Results screenshot:**

