The Course Project

Name: Vishwanath reddy saikam

Net id: wu3638

Part 1:

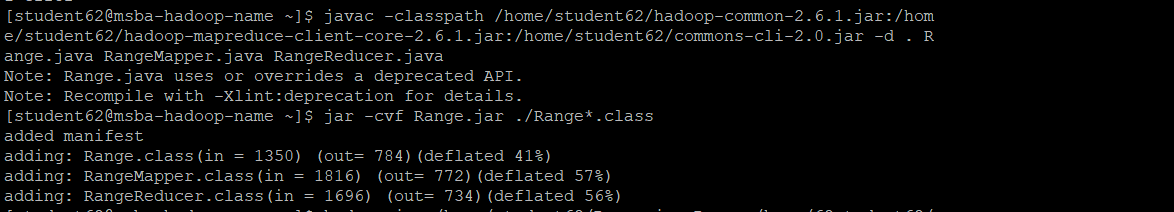
The first part is to develop a Mapper and Reducer application to calculate the *range* (the difference between max and min values) of *sky ceiling height* (meters) for *each observation month* from NCDC records (note: 99999 indicates missing value, and [01459] indicate good quality value).

1. Creating the mapper, reducer and job(main) file
2. Compiling the files to create .class files

javac -classpath /home/student62/hadoop-common-2.6.1.jar:/home/student62/hadoop-mapreduce-client-core-2.6.1.jar:/home/student62/commons-cli-2.0.jar -d . Range.java RangeMapper.java RangeReducer.java

1. And also creating the .jar file using the main class file

Jar -cvf Range.jar ./Range\*.class



1. Making a director In HDFS:

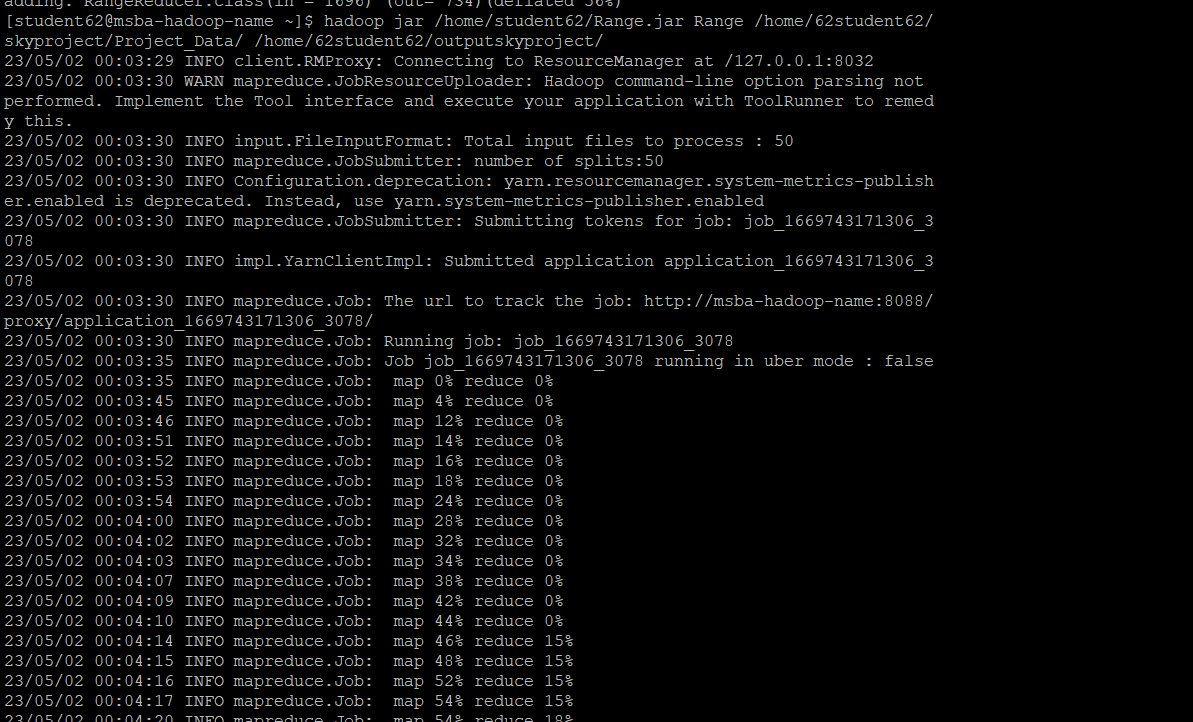
hdfs dfs -mkdir /home/62student62/skyproject

1. Copying from local to HDFS:

hdfs dfs -copyFromLocal /home/student62/Project\_Data /home/62student62/skyproject/Project\_Data

1. running the jar file to get the output we need

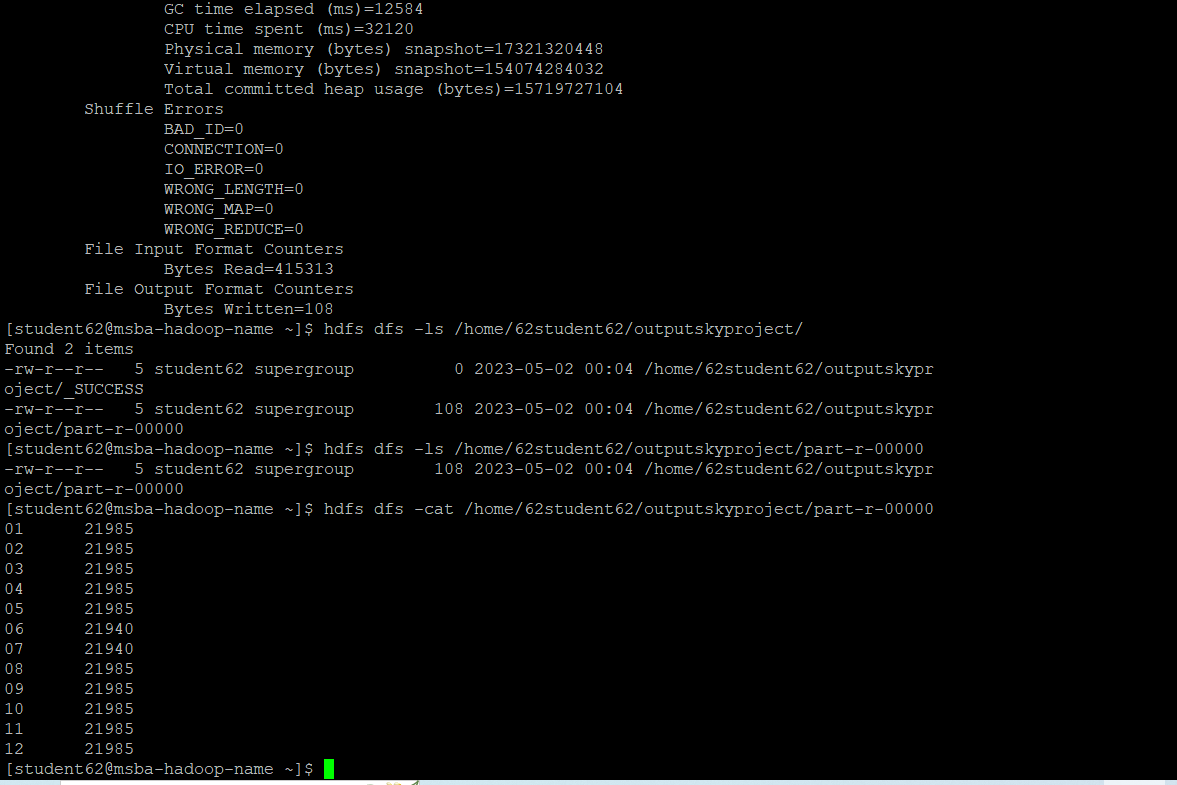
hadoop jar /home/student62/Range.jar Range /home/62student62/skyproject/Project\_Data /home/62student62/outputskyproject/



1. displaying the output files and the content of the file

hdfs dfs -ls /home/62student62/outputskyproject/

hdfs dfs -cat /home/62student62/outputskyproject/part-r-00000



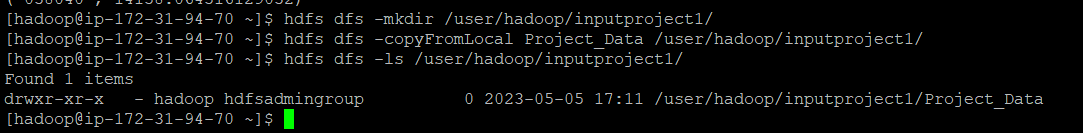
Part 2:

The second part is to develop a python application that can be implemented in PySpark to calculate the *average visibility distance* (meters) for *each USAF weather station ID* from NCDC records (note: 999999 indicates missing value, and [01459] indicate good quality value).

1. First we have to create the python file, I have used groupby to sort the data to obtain weather station and visibility distance
2. Creating a directory and importing the ProjectData to HDFS and also we have to import the python file from computer to local hadoop

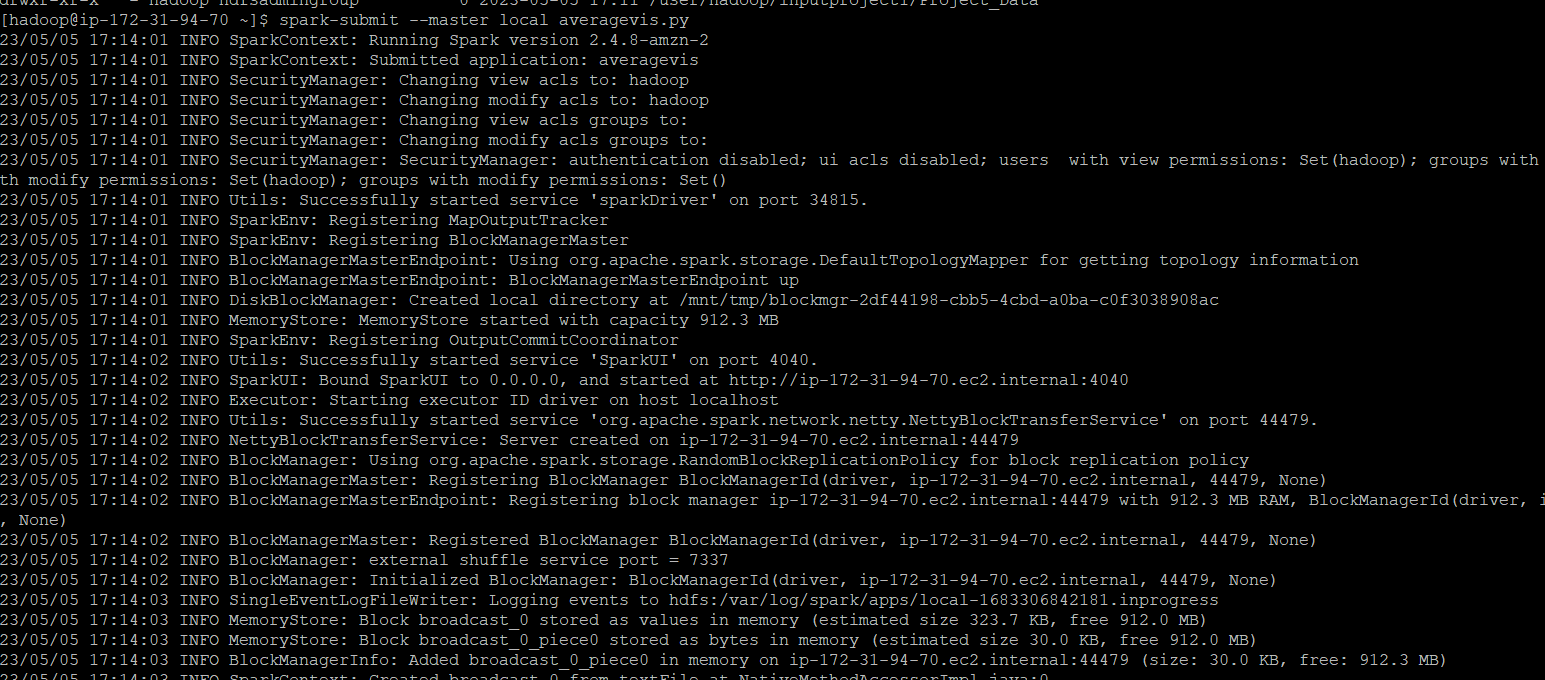
Hdfs dfs -mkdir /user/hadoop/inputproject1/

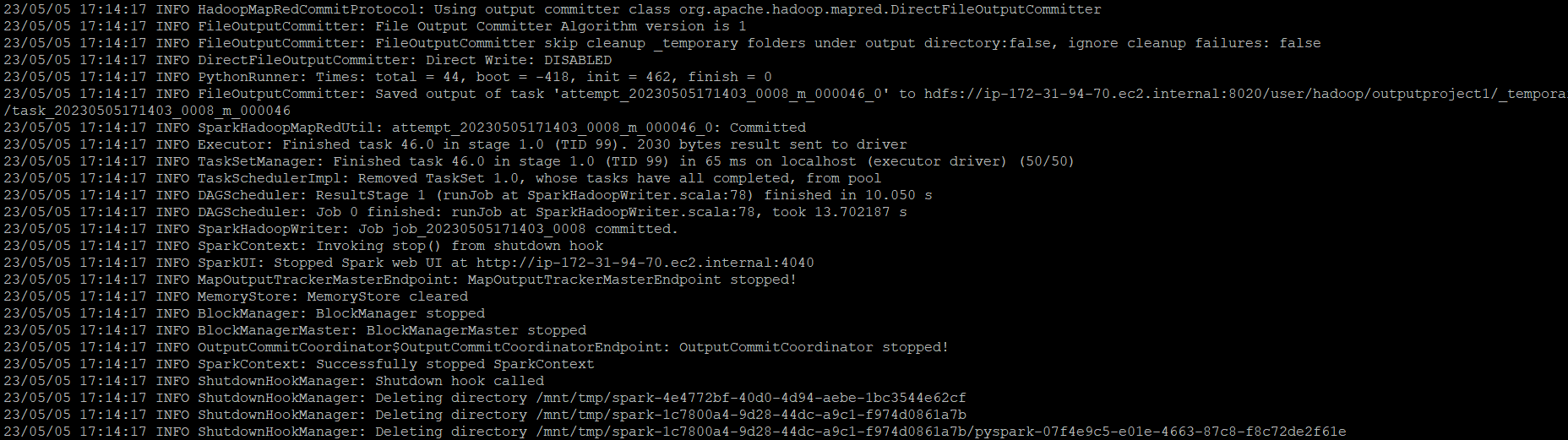
Hdfs dfs -copyFromLocal Project\_Data /user/hadoop/inputproject1/



3)we have to run the python file on local, in the python file we already provide the output path and input path

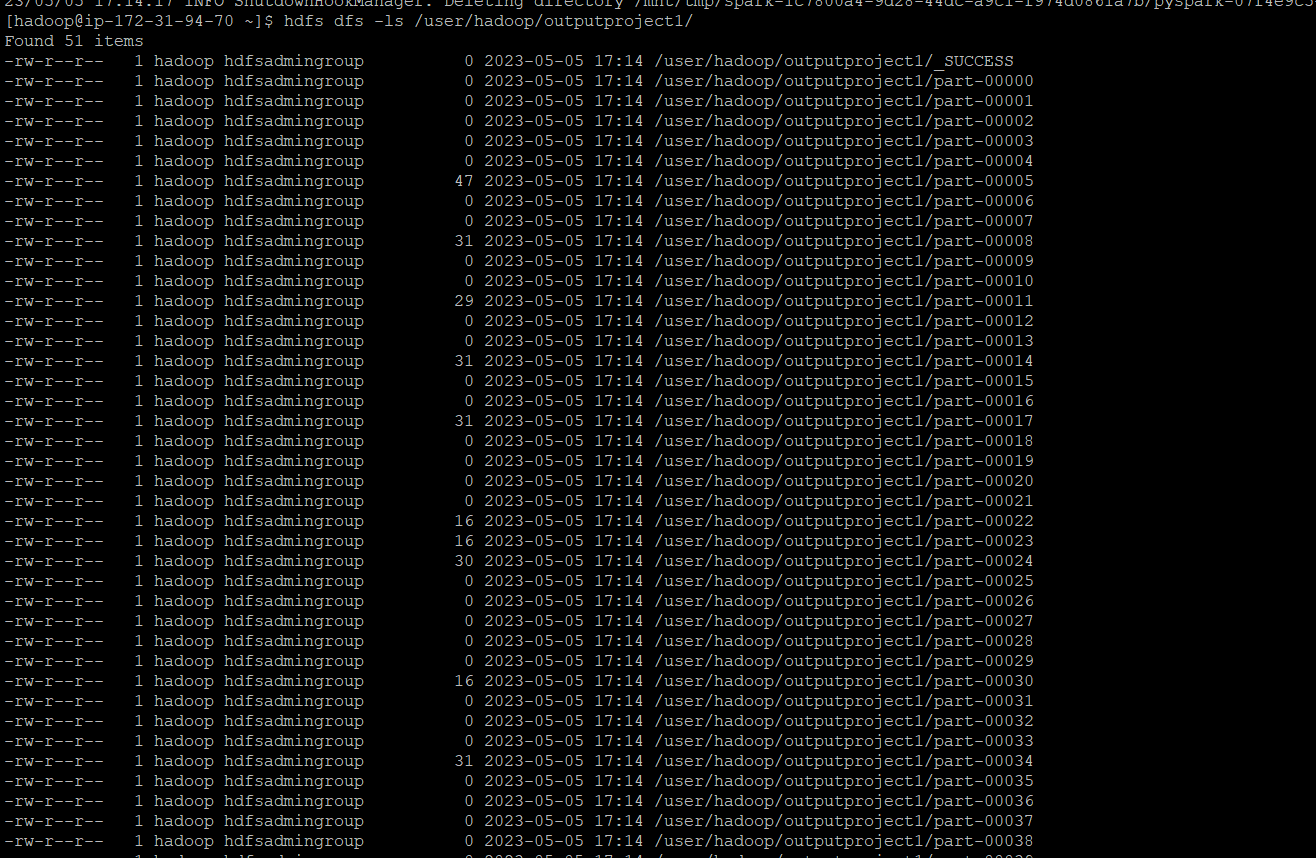
Spark-submit –master local averagevis.py



There were no errors

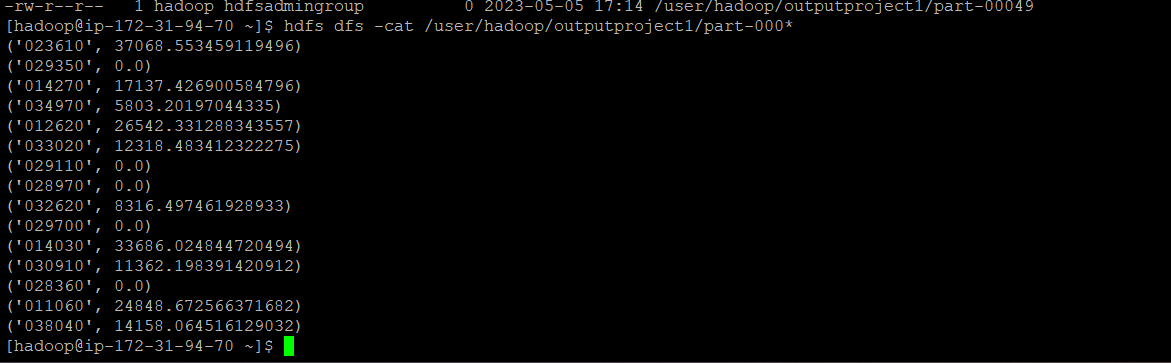
4)displaying the output files:

Hdfs dfs -ls /user/hadoop/outputproject1/



5)concatenating all files and displaying them

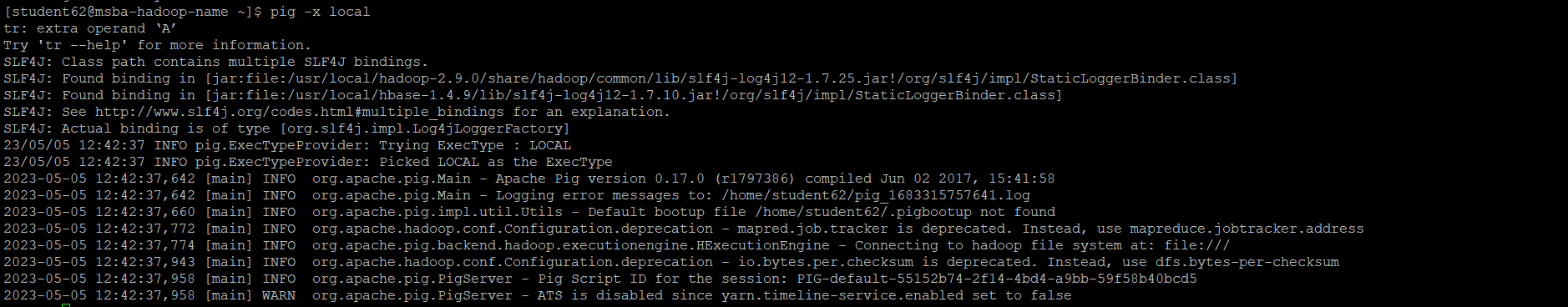
Hdfs dfs -cat /user/hadoop/outputproject1/part-000\*



Part 3:

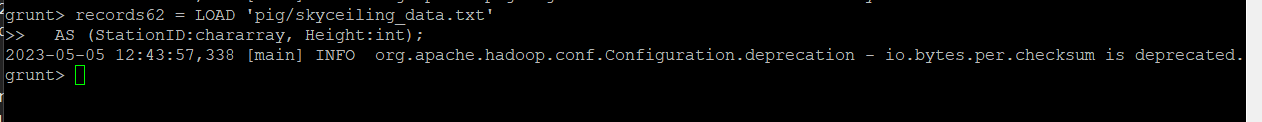
The third part is to load the text file into Pig and get the range of sky ceiling height for each USAF weather station ID.

pig –x local

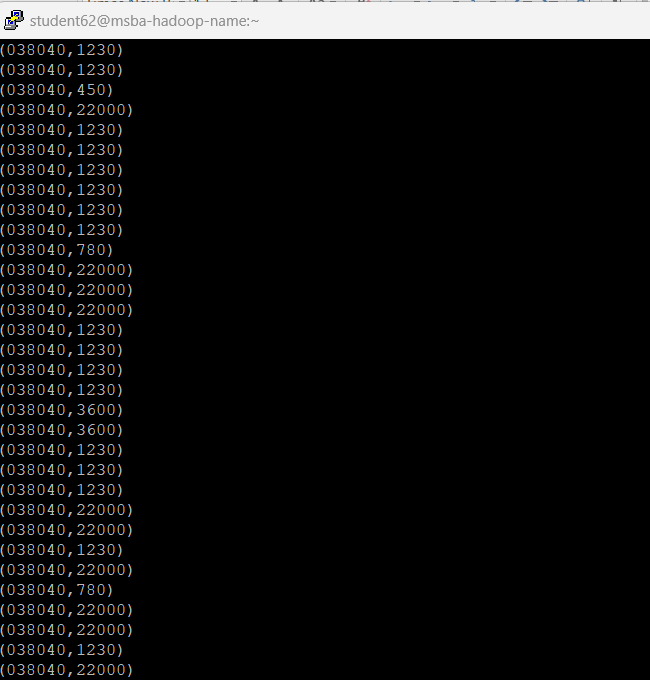


records62 = LOAD 'pig/skyceiling\_data.txt'

AS (StationID:chararray, Height:int);

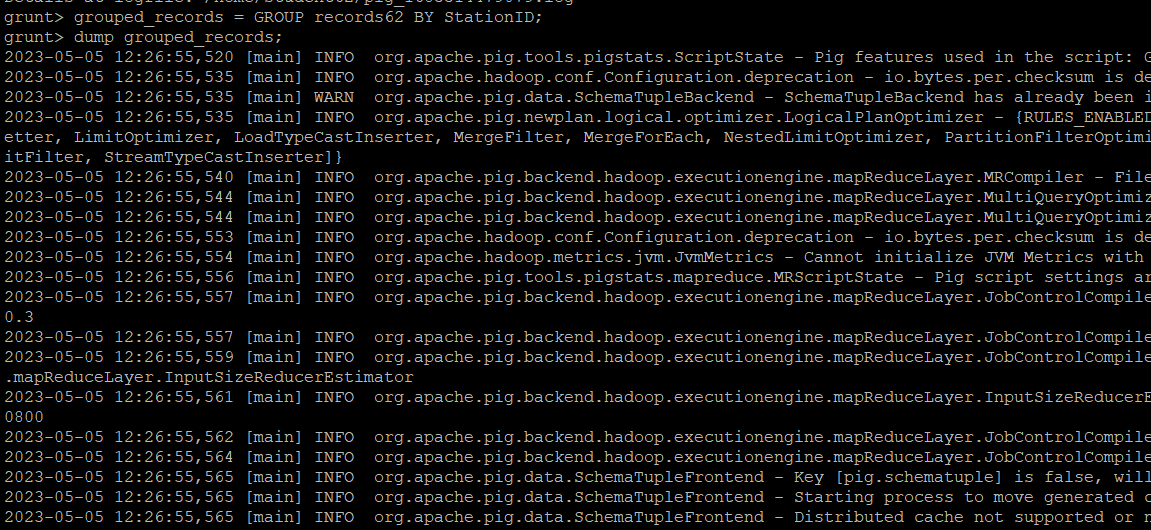


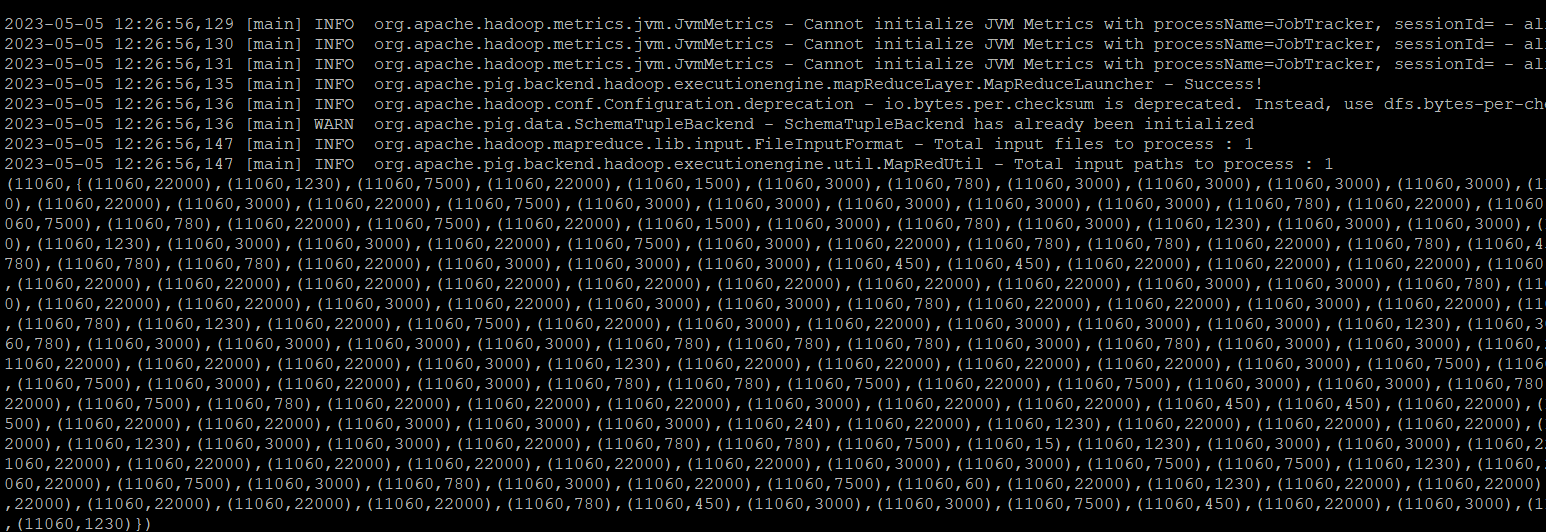
dump records;



grouped\_records = GROUP records BY StationID;

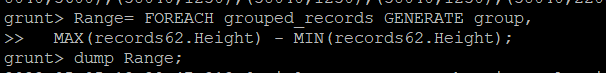
DUMP grouped\_records;



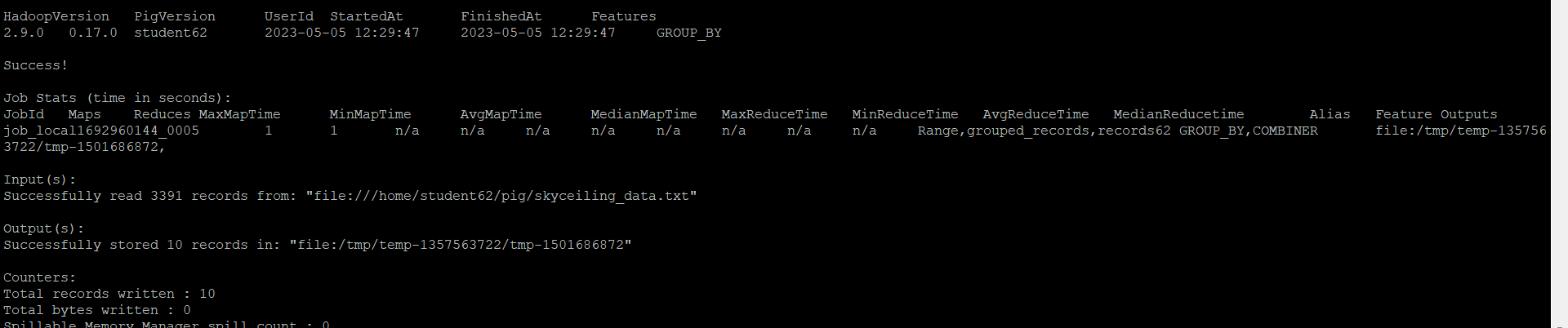


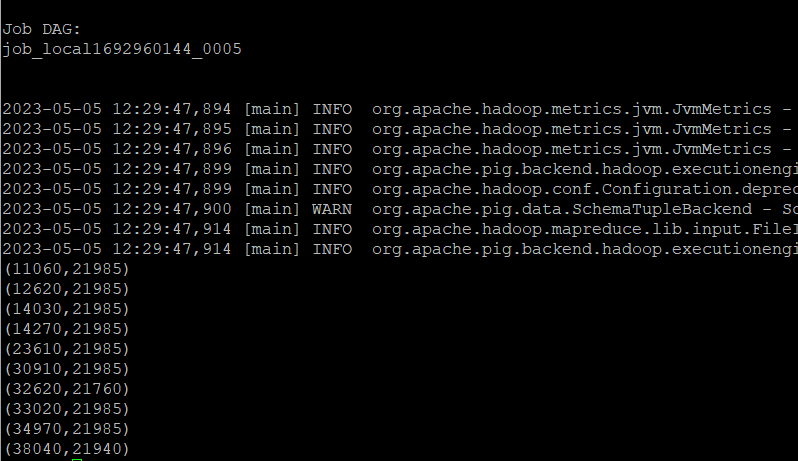
Range= FOREACH grouped\_records GENERATE group,

MAX(records62.Height) - MIN(records62.Height);



DUMP Range;

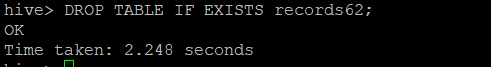




Part 4:

The fourth part is to load the text file into Hive and get the average sky ceiling height for each USAF weather station ID.

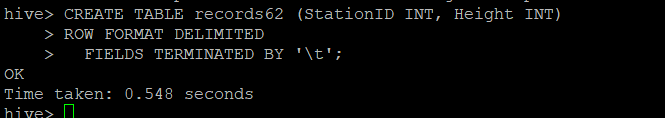
DROP TABLE IF EXISTS records62;



CREATE TABLE records62 (StationID INT, Height INT)

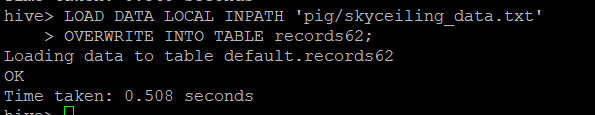
ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t';



LOAD DATA LOCAL INPATH 'pig/skyceiling\_data.txt'

OVERWRITE INTO TABLE records62;



SELECT StationID, AVG(Height)

FROM records62

GROUP BY StationID;

