The Course Project

Apurva Vaze

NetId:os6218

The course project includes 4 parts.

**Part 1**:

The first part is to develop a Mapper and Reducer application to calculate average visibility distance (meters) for each USAF weather station ID from NCDC records (note: 999999 indicates missing value, and [01459] indicate good quality value).

Step 1:  
Created python files for mapper and reducer avg\_visibility\_distance\_map.py and avg\_visibility\_distance\_reduce.py using Spyder text editor.

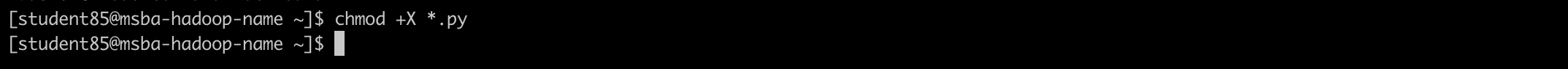
Text

Description automatically generated

Text

Description automatically generated

Step 2:  
Moved mapper and reducer python files to local path /home/student85 using cyberduck and changed its permission.



Step 3:

Created a new folder named as “ProjectData” on local system /home/student85 and copied all provided files to that folder. Also, created new directory “ProjectDataInput”on HDFS to copy files from local system.

Graphical user interface, application, table

Description automatically generated

hdfs dfs -mkdir /home/85student85/ProjectDataInput



hdfs dfs -copyFromLocal /home/student85/ProjectData/ /home/85student85/ProjectDataInput/



Text

Description automatically generated

Step 4:

Executed map reduce jobs using below command:

hadoop jar hadoop-streaming-2.7.3.jar -file /home/student85/avg\_visibility\_distance\_map.py -mapper /home/student85/avg\_visibility\_distance\_map.py -file /home/student85/avg\_visibility\_distance\_reduce.py -reducer /home/student85/avg\_visibility\_distance\_reduce.py -input /home/85student85/ProjectDataInput/ProjectData/ -output /home/85student85/projectDataOutput;

Text

Description automatically generated

Text

Description automatically generated

hdfs dfs -ls /home/85student85/projectDataOutput

Text

Description automatically generated

Text

Description automatically generated

**Part 2:**

The second part is to develop a Mapper and Reducer application to retrieve USAF weather station ID and sky ceiling height (meters) from NCDC records (note: 99999 indicates missing value, and [01459] indicate good quality value) and then write the USAF weather station ID and sky ceiling height data into a text file.

Step 1:  
Created python files for mapper and reducer Get\_stationID\_skyht\_map.py and Get\_stationID\_skyht\_reduce.py using Spyder text editor.

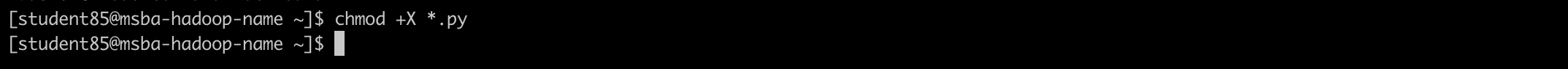
Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Step 2:  
Moved mapper and reducer python files to local path /home/student85 using cyberduck and changed its permission.



Note: Input files are already available on HDFS under folder /home/85student85/ProjectDataInput/ProjectData/ (this is already done in part 1).  
  
  
Step 3:  
Executed map reduce jobs using below command:

hadoop jar hadoop-streaming-2.7.3.jar -file /home/student85/Get\_stationID\_skyht\_map.py -mapper /home/student85/Get\_stationID\_skyht\_map.py -file /home/student85/Get\_stationID\_skyht\_reduce.py -reducer /home/student85/Get\_stationID\_skyht\_reduce.py -input /home/85student85/ProjectDataInput/ProjectData/ -output /home/85student85/projectDataOutput\_part2;

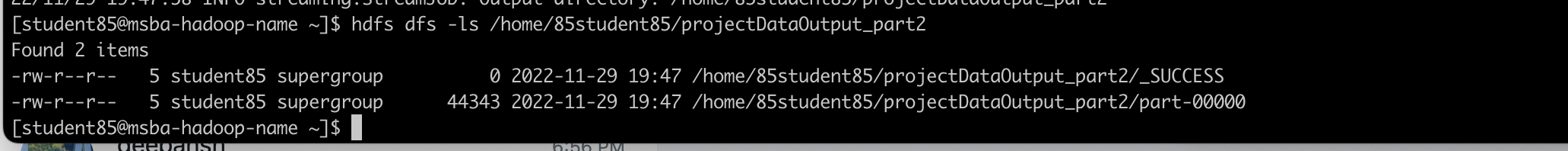
Text

Description automatically generated

Text

Description automatically generated

hdfs dfs -ls /home/85student85/projectDataOutput\_part2



hdfs dfs -cat /home/85student85/projectDataOutput\_part2/part-00000

Shape, rectangle

Description automatically generated

Step 4:

Copy output data file from HDFS to local and saved into text format as weather\_data.txt

hdfs dfs -copyToLocal /home/85student85/projectDataOutput\_part2/part-00000 /home/student85/weather\_data.txt

cat /home/student85/weather\_data.txt

Shape

Description automatically generated with medium confidence

Background pattern

Description automatically generated

Part 3:

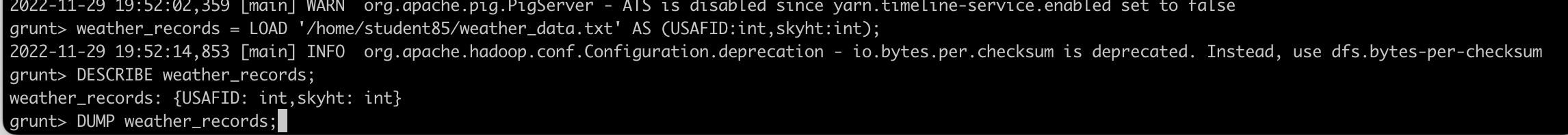
The third part is to load the text file into Pig and get the average sky ceiling height for each USAF weather station ID.

Step 1:  
Connect to Pig.  
pig -x local

Text

Description automatically generated

Step 2:  
Load data weather\_data into pig using below command.  
weather\_records = LOAD '/home/student85/weather\_data.txt' AS (USAFID:int,skyht:int);  
DESCRIBE weather\_records;  
DUMP weather\_records;



Shape

Description automatically generated with medium confidence

Step 3:

DEFINE avg\_by\_group(X, group\_key, avg\_field) RETURNS Y {  
A = GROUP $X by $group\_key;  
$Y = FOREACH A GENERATE group, AVG($X.$avg\_field);  
};

avg\_temp = avg\_by\_group(weather\_records,USAFID,skyht);  
DUMP avg\_temp;

A black screen with white text

Description automatically generated with low confidence

Text

Description automatically generated

Step 4:

Exit from pig using quit command.

Part 4:

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

Step 1:  
Connect to hive

Text

Description automatically generated

Step 2:  
DROP TABLE IF EXISTS weather\_records85;

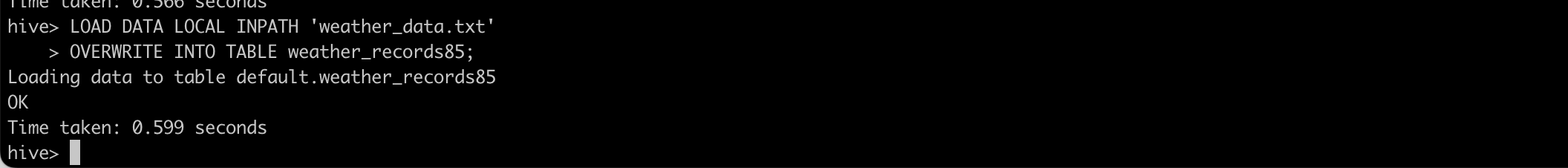


Step 3:  
CREATE TABLE weather\_records85 (USAFID int,skyht INT)   
ROW FORMAT DELIMITED   
FIELDS TERMINATED BY '\t';

A picture containing graphical user interface

Description automatically generated

Step 4:  
LOAD DATA LOCAL INPATH 'weather\_data.txt'  
OVERWRITE INTO TABLE weather\_records85;

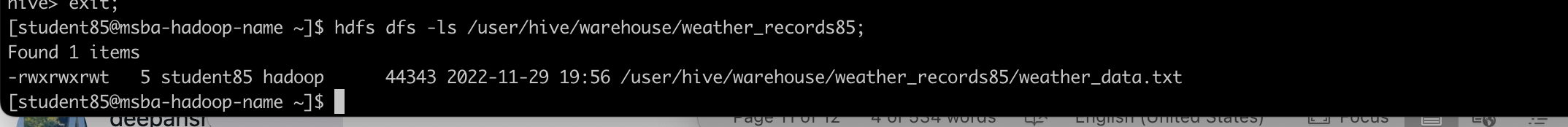


select \* from weather\_records85;

Shape

Description automatically generated

Step 5:  
Exit from hive.  
hdfs dfs -ls /user/hive/warehouse/weather\_records85;



Step 6:  
Once again connect to hive to run below query.  
To display column names highest\_height, lowest\_ht given in below query we have to set this value as true  
set hive.cli.print.header=true;

Text

Description automatically generated

Step 7:  
SELECT USAFID as USAFID, max(skyht) as highest\_height, min(skyht) as lowest\_height FROM weather\_records85 GROUP BY USAFID;

Text

Description automatically generated