The Course Project

Part 1:

Develop a Mapper and Reducer application to calculate the *average* of *wind direction* (degree) for *each observation month* *from each year* (e.g. 195001) from NCDC records (note: 999 indicates missing value, and [01459] indicate good quality value).

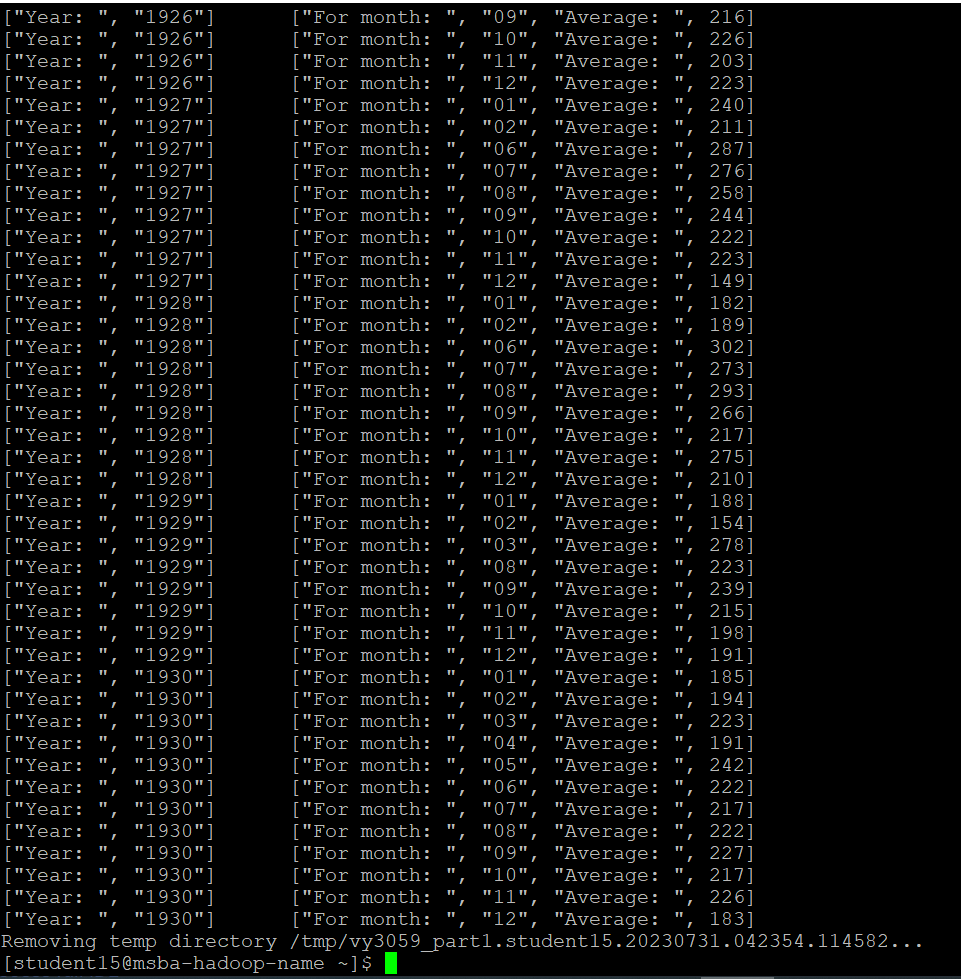
python vy3059\_part1.py -r local Project\_Data/\*.gz

A screen shot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated



python vy3059\_part1.py -r local Project\_Data/\*.gz --output-dir=./output\_Project1

A computer screen with text and numbers

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

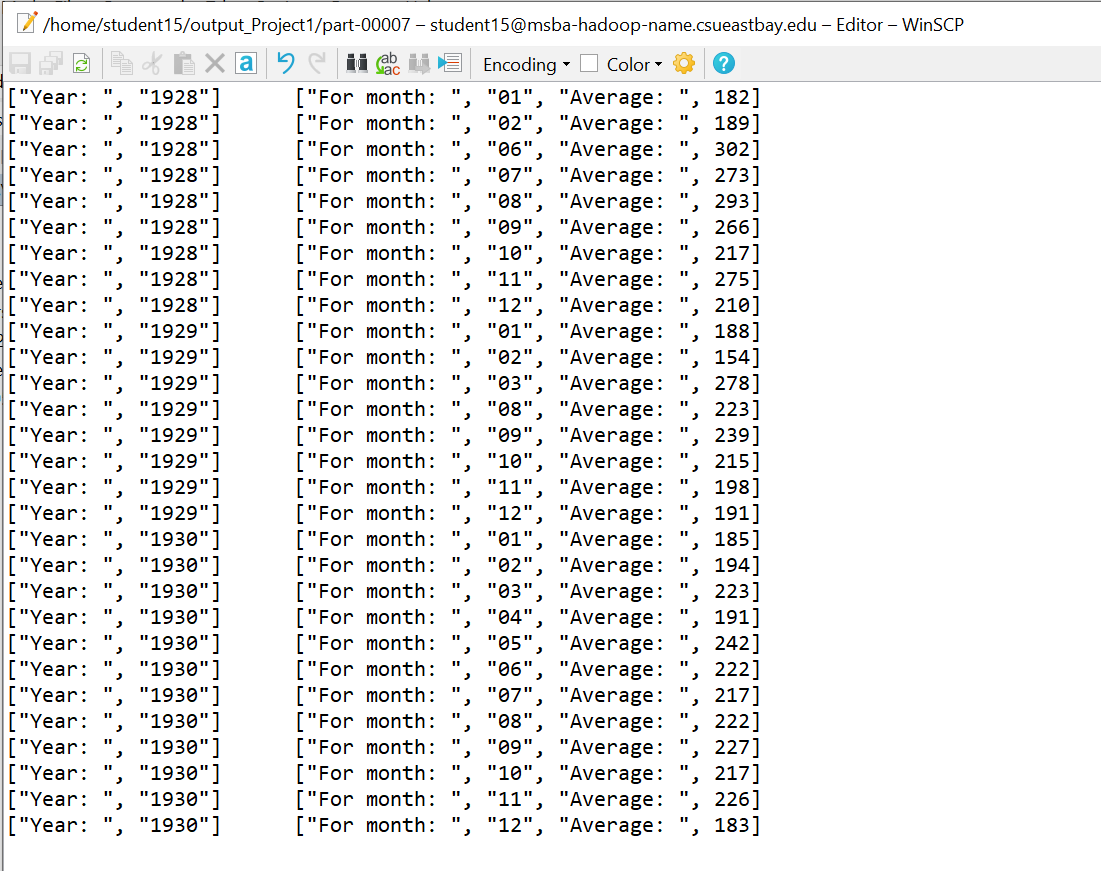
Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

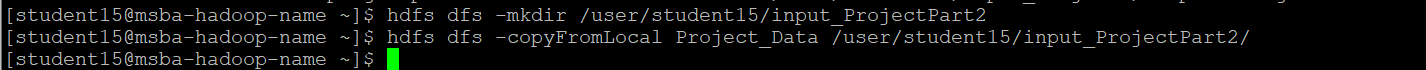


Part 2:

Develop a python application that can be implemented in PySpark to calculate the *range* (the difference between max and min values) of *sky ceiling height* (meters) for *each USAF weather station ID* from NCDC records (note: 99999 indicates missing value, and [01459] indicate good quality value).

hdfs dfs -mkdir /user/student15/input\_ProjectPart2

hdfs dfs -copyFromLocal Project\_Data /user/student15/input\_ProjectPart2/



spark-submit --master local vy3059\_part2.py

A computer screen with many small colored lines

Description automatically generated

hdfs dfs -ls /user/student15/output\_ProjectPart2-1/

A black screen with many small colored lines

Description automatically generated

hdfs dfs -cat /user/student15/output\_ProjectPart2-1/\*

A screen shot of a computer

Description automatically generated

Part 3:

Develop a Mapper and Reducer application to retrieve *USAF weather station ID* and *visibility distance (meters)* from NCDC records (note: 999999 indicates missing value, and [01459] indicate good quality value) and then *write* the USAF weather station ID and visibility distance data into *a text file*.

python vy3059\_part3.py -r local Project\_Data/\*.gz

A screenshot of a computer

Description automatically generated

python vy3059\_part3.py -r local Project\_Data/\*.gz > VisibilityAndStation\_data.txt

A computer screen with numbers and letters

Description automatically generated

A screenshot of a computer

Description automatically generated

Part 4:

Load the text file into Pig and get the range of *visibility distance* for each USAF weather station ID.

pig -x local

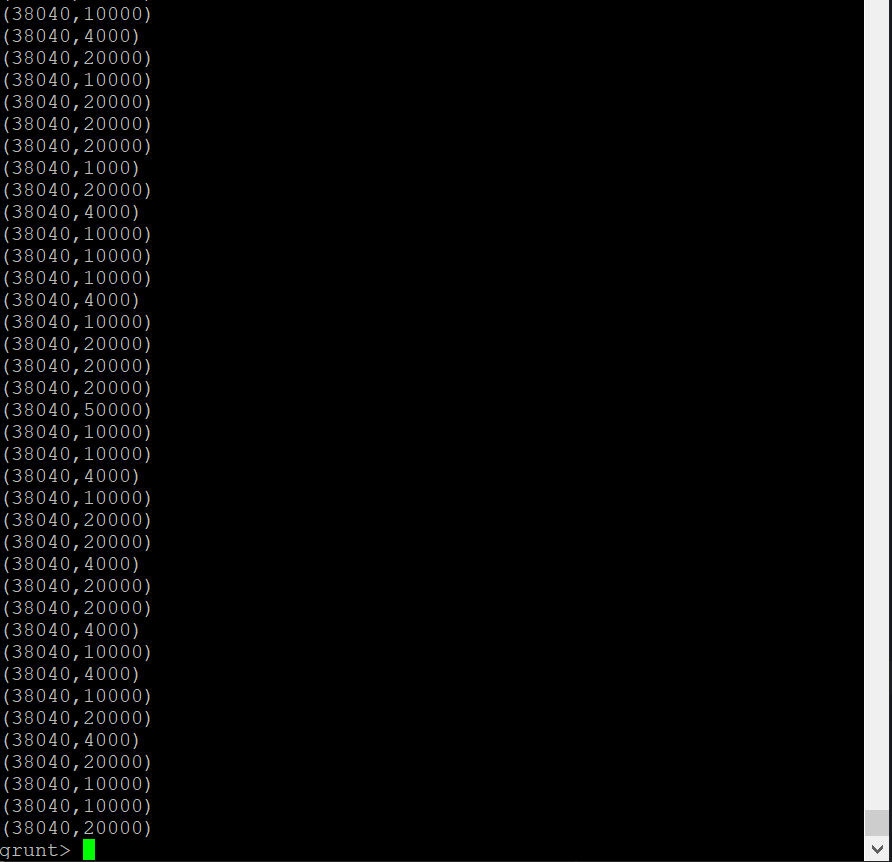
Visibility\_records = LOAD 'VisibilityAndStation\_data.txt'

AS (station\_id:int, visibility\_dist:int);

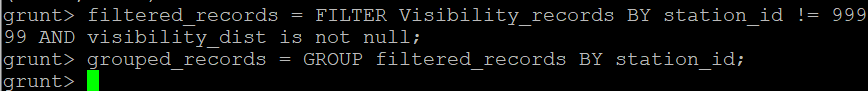
A screen shot of a computer

Description automatically generated

DUMP Visibility\_records;



grouped\_records = GROUP Visibility\_records BY station\_id;

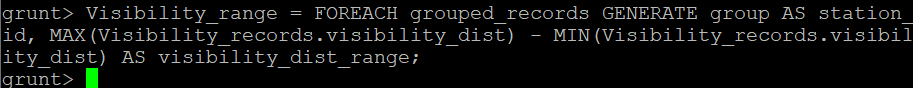


DUMP grouped\_records;

A screenshot of a computer screen

Description automatically generated

Visibility\_range = FOREACH grouped\_records GENERATE group AS station\_id, MAX(Visibility\_records.visibility\_dist) - MIN(Visibility\_records.visibility\_dist) AS visibility\_dist\_range;



DUMP Visibility\_range;

A screen shot of a computer

Description automatically generated

STORE Visibility\_range INTO 'output\_Project4';

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

Part 5:

Load the text file into Hive and get the average *visibility distance* for each USAF weather station ID.

DROP TABLE IF EXISTS VisibilityAndStation1000;

CREATE TABLE VisibilityAndStation1000 (station\_id INT, visibility\_dist INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t';

A computer screen shot of a computer

Description automatically generated

LOAD DATA LOCAL INPATH 'VisibilityAndStation\_data.txt'

OVERWRITE INTO TABLE VisibilityAndStation1000;

A black screen with white text

Description automatically generated

SELECT station\_id, AVG(visibility\_dist)

FROM VisibilityAndStation1000

WHERE station\_id != 99999 AND visibility\_dist IS NOT NULL

GROUP BY station\_id;

A computer screen shot of a program

Description automatically generated