**The Course Project- Jasvitha Buggana – student59- js3225**

The course project includes 3 parts. The first part is to develop a Python 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. The second part is to load the text file into Pig and get the highest and lowest temperatures for each year. The third part is to load the text file into Hive and get the average temperature for each year.

**Project Steps and Outputs:**

Created two Python files: Project\_Map.py and Project\_Reduce.py

Download CourseProjectData.zip and transfer the files using winscp.

* hdfs dfs -copyFromLocal /home/student59/CourseProjectData /home/59student59/CourseProjectData
* hadoop jar /home/student59/hadoop-streaming-2.9.0.jar \input /home/59student59/CourseProjectData \output /home/59student59/Project\_js3225 \mapper /home/student59/Project\_map.py \reducer /home/student59/Project\_Reduce.py \file /home/student59/Project\_map.py \file /home/student59/Project\_Reduce.py

A computer screen with white text

Description automatically generated

A computer screen with a black background

Description automatically generated

A computer screen shot of a person

Description automatically generated

* hdfs dfs -ls /home/59student59/Project\_js3225/

A black screen with white text

Description automatically generated

* hdfs dfs -copyToLocal /home/59student59/Project\_js3225/part-00000 /home/student59/js3225\_output.txt
* js\_3225\_output file will be generated.

A screenshot of a computer

Description automatically generated

* **In Pig:** pig -x local
* records = LOAD 'js3225\_output.txt' AS (year:chararray, temperature:int);
* grouped\_records = GROUP records BY year;
* maxtemp = FOREACH grouped\_records GENERATE group, MAX(records.temperature);
* DUMP maxtemp;

A screen shot of a computer screen

Description automatically generated

A computer screen shot of a computer code

Description automatically generated

A black screen with white text

Description automatically generated

* mintemp = FOREACH grouped\_records GENERATE group, MIN(records.temperature);
* DUMP mintemp;

A computer screen with white text

Description automatically generated

A black screen with white text

Description automatically generated

* **In hive:**
* DROP TABLE IF exists js3225table;
* CREATE TABLE js3225table (year STRING, temperature INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
* LOAD DATA LOCAL INPATH 'js3225\_output.txt' OVERWRITE INTO TABLE js3225table;
* SELECT year, AVG(temperature)

FROM js3225table

GROUP BY year;

A black screen with white text

Description automatically generated

A black screen with white text

Description automatically generated

You need to turn in 1) Python files (mapper and reducer), 2) the commands for executing the Python application in Hadoop, 3) the text file including Year and Temperature data created by you, 4) the screenshot of the text file being created, 5) the screenshot of the final Pig output showing the year and the highest and lowest temperatures, and 6) the screenshot of the final Hive output showing the year and average temperature.

The original dataset for this project is available on Canvas.