The Course Project

The course project includes 3 parts. The first part is to develop a Python **MRJob** 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.

You need to turn in 1) Python MRJob coding file, 2) the commands for executing the Python MRJob 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.

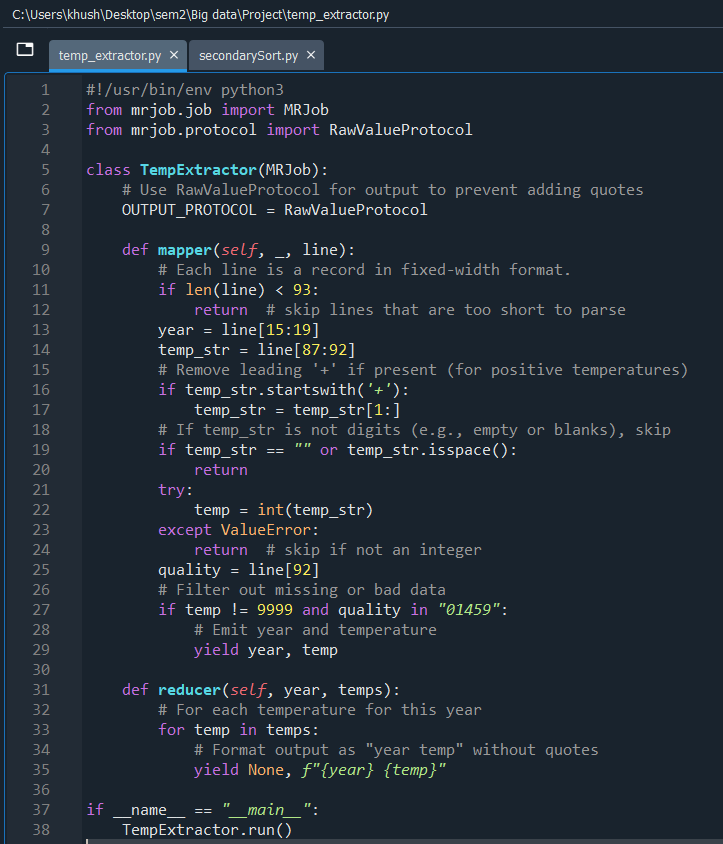
The original dataset for this project is available on Canvas.

# The Course Project

Url : 134.154.190.204

## Part 1: Develop a Python Mrjob Application

1. Creating the python file:



1. Transfer necessary files to hadoop:
2. temp\_extractor.py
3. Newprojectdata.zip

Moved the above file into a directory ncdc\_project

1. Extracting the NewProjectData.zip file



1. Creating directory in HDFS

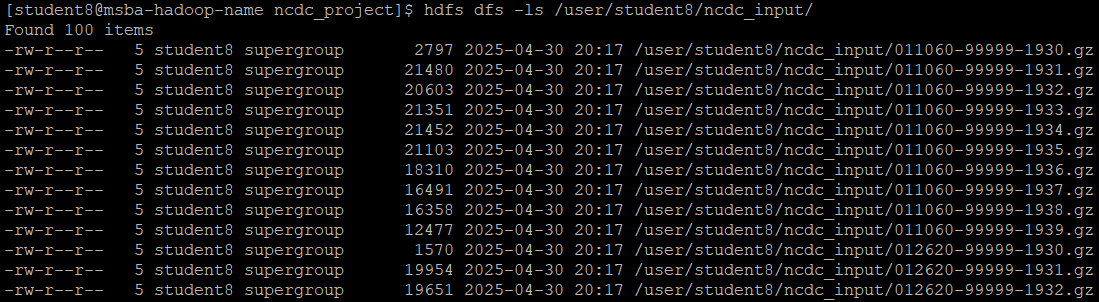
hdfs dfs -mkdir -p /user/student8/ncdc\_input

1. Moving all the .gz files to ncdc\_input

hdfs dfs -put \*.gz /user/student8/ncdc\_input/

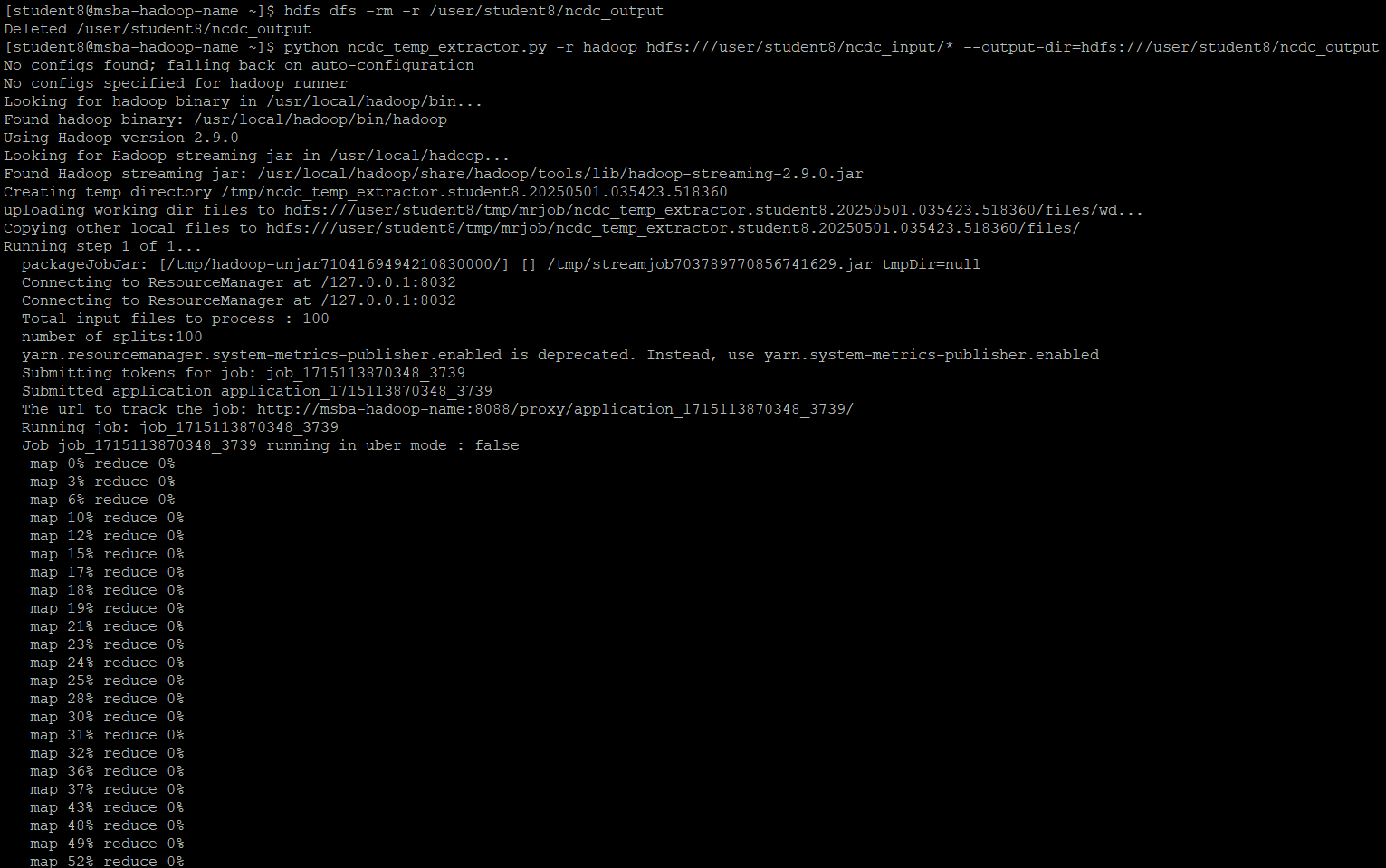
1. Checking all the files are moved or not:

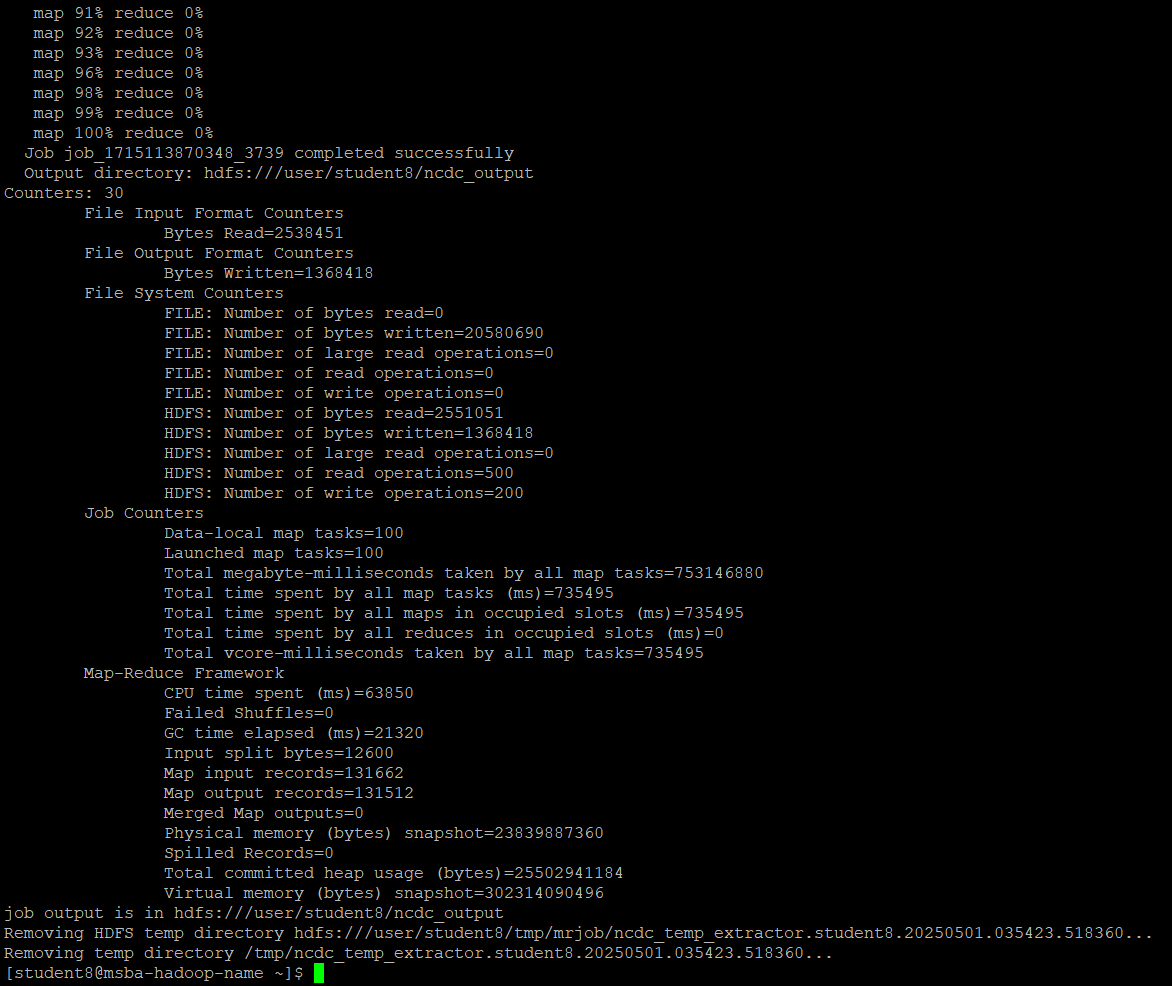
hdfs dfs -ls /user/student8/ncdc\_input/



1. Running the mrjob file

python ncdc\_temp\_extractor.py -r hadoop \ hdfs:///user/student8/ncdc\_input/\* \ --output-dir hdfs:///user/student8/ncdc\_output

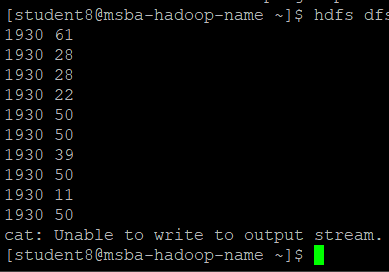




1. Checking the output

hdfs dfs -ls /user/student8/ncdc\_output/

hdfs dfs -cat /user/student8/ncdc\_output/part-00000 | head



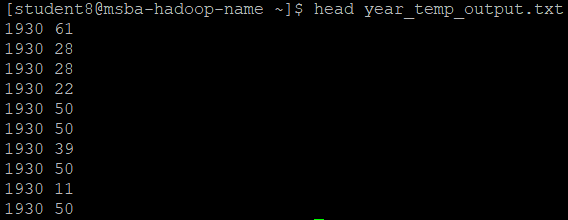
Now moving the file into text file:

hdfs dfs -getmerge /user/student8/ncdc\_output/ year\_temp\_output.txt

Checking the file is being created:



Checking if the files are stored properly:



**Part 2: highest and lowest temperatures for each year**

Make a pig file

Nano year\_minmax.pig

Store the below code

All command in one place:

-- Calculate min and max temperature per year using Pig

raw\_data = LOAD '/user/student8/input/year\_temp\_output.txt'

USING PigStorage(' ')

AS (year: chararray, temp: int);

-- (Optional cleanup: remove quotes from year if present)

-- clean\_data = FOREACH raw\_data GENERATE REPLACE(year, '\"', '') AS year, temp;

-- grouped\_data = GROUP clean\_data BY year;

grouped\_data = GROUP raw\_data BY year;

min\_max\_temp = FOREACH grouped\_data GENERATE

group AS year,

MIN(raw\_data.temp) AS min\_temp,

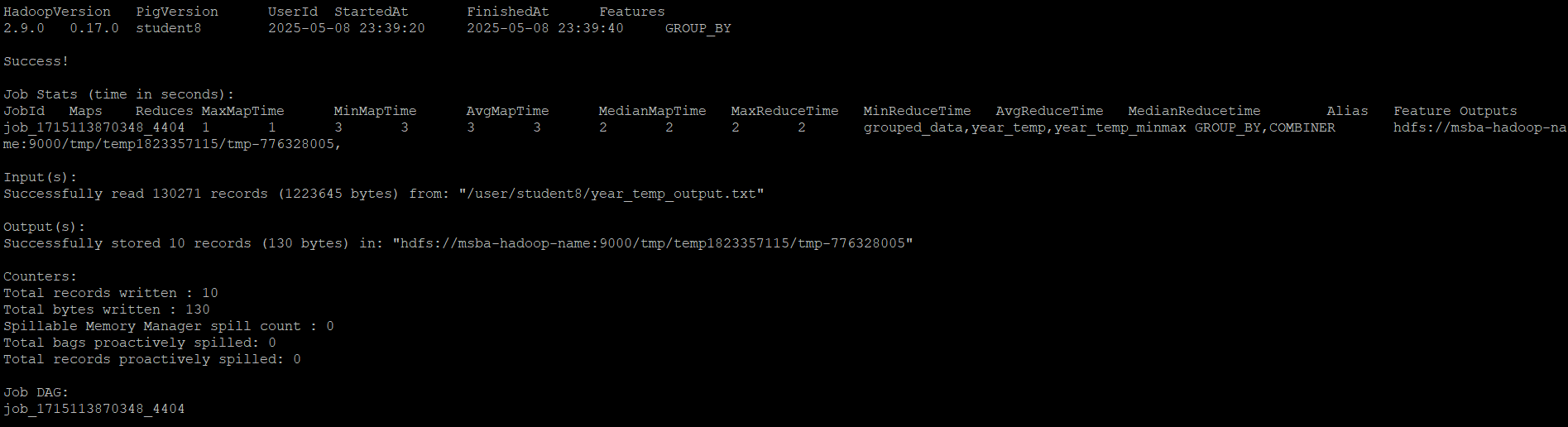
MAX(raw\_data.temp) AS max\_temp;

STORE min\_max\_temp INTO '/user/student8/output/year\_minmax' USING PigStorage(',');

Run this code:

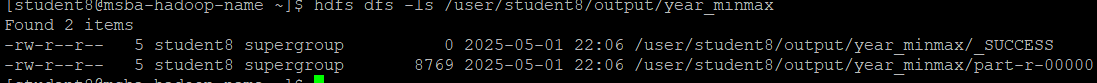
pig year\_minmax.pig

Output:



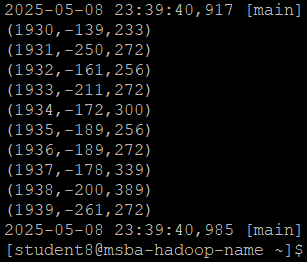
To verify the output directory:

hdfs dfs -ls /user/student8/output/year\_minmax



Verify the data now:

hdfs dfs -cat /user/student8/output/year\_minmax/part-r-00000 | head -20

4

**Part 3: Average temperature for each year**

Initialize New Metastore Schema

schematool -dbType derby -initSchema

Start hive

hive

Create Table with Proper Syntax

CREATE TABLE year\_temp (

year STRING,

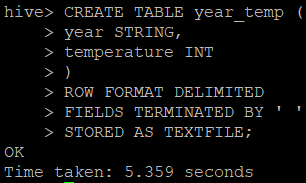
temperature INT

)

ROW FORMAT DELIMITED

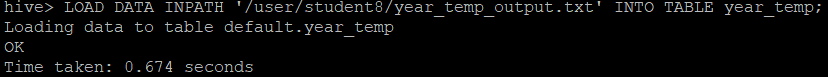
FIELDS TERMINATED BY ' '

STORED AS TEXTFILE;



Load Data from HDFS

LOAD DATA INPATH '/user/student8/year\_temp\_output.txt' INTO TABLE year\_temp;



Calculate Average Temperatures

SELECT year, AVG(temperature) AS avg\_temp

FROM year\_temp

GROUP BY year

ORDER BY year;

****