Temperature analysis

The project includes 3 parts. The first part is to develop a Mapper and Reducer application to retrieve Year and Temperature from original NCDC records 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.

Part 1:

Compile Java files:

javac -classpath /home/student4/hadoop-common-2.6.1.jar:/home/student4/hadoop-mapreduce-client-core-2.6.1.jar:/home/student4/commons-cli-2.0.jar -d . MaxTemperature.java MaxTemperatureMapper.java MaxTemperatureReducer.java

Create a Jar file:

jar -cvf max.jar MaxTemperature\*.class

A close up of text on a white background

Description automatically generated

Copy the data from local to HDFS:

hdfs dfs -copyFromLocal Data /home/student4/

Command for setting location:

export HADOOP\_CLASSPATH=/home/student4/

Run a Jar file on Hadoop:

hadoop jar max.jar MaxTemperature /home/student4/Data /home/student4/output120

A screenshot of a cell phone

Description automatically generated

Display the output on screen:

hdfs dfs -text /home/student4/output120/part-r-00000

A screenshot of a cell phone

Description automatically generated

Write Year and Temperature to text file from output120 to local and change name:

hdfs dfs -copyToLocal /home/student4/output120/part-r-00000 /home/student4/Temperature.txt

Copy the data from local to HDFS:

hdfs dfs -copyFromLocal Temperature.txt /home/student4/

A screenshot of a computer

Description automatically generated

A screenshot of a social media post

Description automatically generated

Part2: pig

Run pig:

pig -x local

Command for reading txt file to pig:

records = LOAD 'Temperature.txt' AS (year:int, temperature:int);

DUMP records;

A close up of a logo

Description automatically generated

Commands in pig:

grouped\_records = GROUP records BY year;

DUMP grouped\_records;

A close up of an animal

Description automatically generated

Command to get highest temperature in pig:

max = FOREACH grouped\_records GENERATE group,

MAX(records.temperature);

DUMP max;

A screenshot of a social media post

Description automatically generated

Command to get lowest temperature in pig:

min = FOREACH grouped\_records GENERATE group,

MIN(records.temperature);

DUMP min;

A screenshot of a social media post

Description automatically generated

Part 3: Hive

Commends for activating Hive:

mv metastore\_db metastore\_db.old

schematool -dbType derby -initSchema

hive

Add jar file:

ADD JAR max.jar;

Create mean temperature table:

DROP TABLE IF EXISTS temperature\_mean;

CREATE TABLE temperature\_mean (year STRING, temperature INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t';

Populate Hive with the data:

LOAD DATA LOCAL INPATH 'Temperature.txt'

OVERWRITE INTO TABLE temperature\_mean;

Run the query:

SELECT year, AVG(temperature)

FROM temperature\_mean

GROUP BY year;

A screenshot of a social media post

Description automatically generated