BIG DATA ANALYSIS

Hadoop MapReduce for Climate Data Analytics

NAME - SAHIL
ROLL NO. -107121086
ELECTRICAL AND ELECTRONICS ENGINEERING

Task 3

(a) Let’s assume a solution which proposes to use a MapReduce job that does the following:

- The Map task(s) send(s) (,) pairs to the Reducer. Temperatures are converted to degrees Celsius. The output will be: (6, -5.6) (0, -4.4) ...
- For each key/value pair, the Reduce task subtracts the minimum temperature from the maximum temperature, converts it to degrees, and writes the result to a file.

What is wrong with this approach? Can you propose an alternative solution?

Solution: Some of issues regarding given approach tha i felt are listed below:

1. Preprocessing is required to filter out the data for Central Park (USW00094728).
2. Mapper function will execute single line at a time i.e if Tmax and Tmin are in different lines so we will not able to create output key/values pairs (Tmax,Tmin) for reducer.
3. For resolving 2nd issue I did Preprocessing again on filtered dataset and this time i clubbed Tmax and Tmin for a day in one line.
4. Some of Tmin values corresponding to a particular Tmax value can be club together for different days which may create some confusion during ploting of output result.

Alternative Solution:

Mapper (TemperatureMapper class):

- **Input Types:** This Mapper takes LongWritable (representing the byte offset) as the input key, Text (representing a line of text) as the input value, and outputs Text as the output key and Text as the output value.
- **Map Method:** The map method is implemented to parse each line of input, which is assumed to be in CSV format. It extracts relevant information such as the station, date, temperature type (TMAX or TMIN), and temperature value.
- **Filtering**: It filters data for a specific station (USW00094728) and temperature types (TMAX or TMIN).
- **Output:** For the selected station and temperature types, it emits key/value pairs where the key is the date and the value is a concatenation of temperature type and temperature value.

Reducer (TemperatureReducer class):

- **Input Types:** This Reducer takes Text as the input key and an iterable of Text as the input values. It outputs Text as both the output key and output value.
- **reduce Method:** The reduce method processes the grouped data by date. It iterates through the values for each date, which represent temperature information.

- **Calculations:** It calculates the temperature difference (diff) between TMAX and TMIN for each date. Additionally, it normalizes the temperature values by dividing by 10 if they are greater than 10 or less than -10.
- **Output:** The result is emitted as key/value pairs where the key is the date, and the value is the temperature difference or "Not Found" if either TMAX or TMIN is missing for that date.

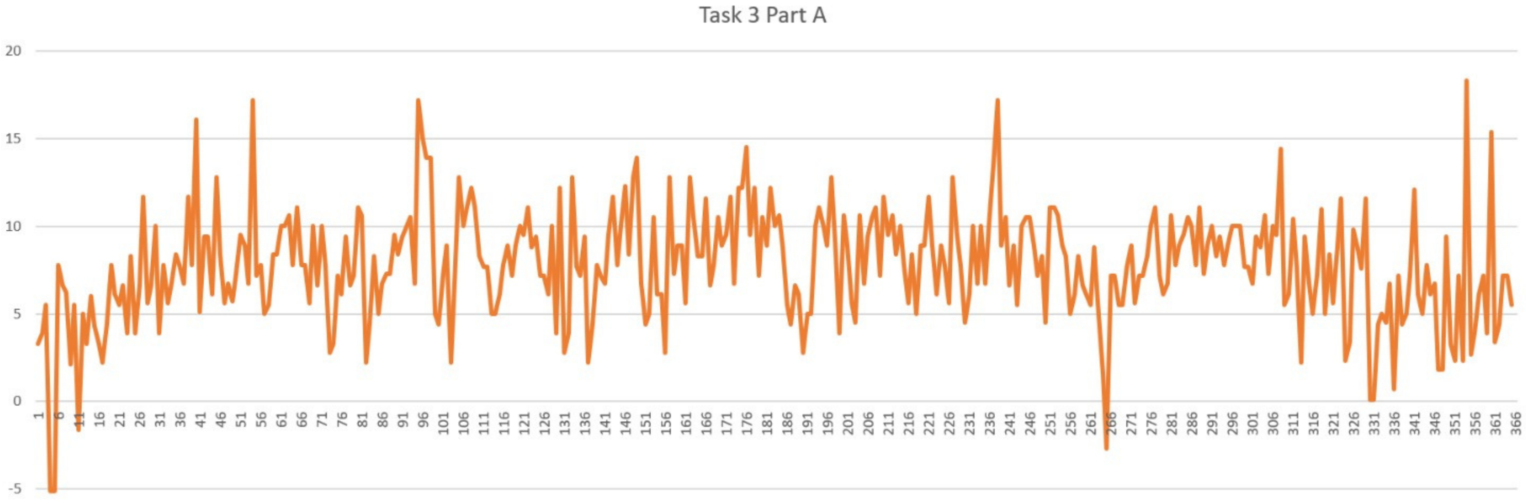
Some Additional Points:

- Preprocessing that I have mentioned above in Issue 1 and Issue 3 is done using **python** on **Google Collab** (code is mention below).
- I have written code for the approach that is given in question and that code is talking dataset after 2nd preprocessing.
- I have followed the same alternative approach for b part of this task.

Python Code:

```
import pandas as pd
from google.colab import drive
drive.mount('/content/drive')
import pandas as pd
file_path = '/content/drive/MyDrive/CSOE18_BDA_Assigment_2 (MR)_Moodle/CSOE18_BDA_Assigment_2 (MR)_Moodle/ncdc-2013-sorted.csv.gz'
df = pd.read_csv(file_path, compression='gzip')
columns_to_remove = [4,5,6,7]
df.drop(columns=df.columns[columns_to_remove], inplace=True)
import csv
string_to_keep = 'USW00094728'
df.loc[df['AE000041196'] != string_to_keep, 'AE000041196'] = pd.NA
df.dropna(subset=['AE000041196'], inplace=True)
s1 = 'TMAX'
s2 = 'TMIN'
df = df[df['TMAX'].isin([s1, s2])]
tmax_rows = df[df['TMAX'] == 'TMAX']
tmin_rows = df[df['TMAX'] == 'TMIN']
dates_with_both = set(tmax_rows['20130101']).intersection(set(tmin_rows['20130101']))
df = df[df['20130101'].isin(dates_with_both)]
df.loc[df['250'].abs() > 10, '250'] /= 10
df = df.pivot(index='20130101', columns='TMAX', values='250')
file_path = '/content/drive/MyDrive/CSOE18_BDA_Assigment_2 (MR)_Moodle/CSOE18_BDA_Assigment_2 (MR)_Moodle/output4.txt'
with open(file_path, 'w', newline='') as file:
    writer = csv.writer(file)
    for row in df.values:
        writer.writerow(row)
```

Output Plot


Task 3 Part A
20
15
10
5
0
-5
1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96 101 106 111 116 121 126 131 136 141 146 151 156 161 166 171 176 181 186 191 196 201 206 211 216 221 226 231 236 241 246 251 256 261 266 271 276 281 286 291 296 301 306 311 316 321 326 331 336 341 346 351 356 361 366