Báo cáo Bài tập lớn môn Big Data

Introduction

Đây là báo cáo của nhóm sau quá trình thực hiện bài toán Tính nhiệt độ trung bình dựa trên dữ liệu về thay đổi nhiệt độ các bang miền Đông nước Mỹ giai đoạn 2006 - 2025.

Nhóm 1 bao gồm các thành viên:

- Trần Khắc Phúc Khánh
- Lưu Quang Linh
- Võ Duy Quang

Link github: https://github.com/KQL-Team/calculate-mean-temperature-hadoop

Data Collecting and Understanding

- Mục tiếu ban đầu của nhóm là thu thập dữ liệu thời tiết chia theo phút. Tuy nhiên, việc thu thập dữ liệu thực dưới dạng phút tỏ ra thiếu khả thi với các công cụ không trả phí, vì vậy nhóm quyết định thu thập dữ liệu theo giờ và thực hiện nội suy để lấy dữ liệu dưới dạng phút.
- Nguồn data được lấy từ meteostat API, danh sách data tập trung ở miền Đông nước Mỹ trong khoảng thời gian từ năm 2006 đến 2025, bao gồm 34 tiểu bang sau:
    1. Tennessee
    2. Kentucky
    3. Mississippi
    4. Louisiana
    5. New Mexico
    6. Iowa
    7. Missouri
    8. Illinois
    9. Indiana
    10. Ohio
    11. Arkansas
    12. North Dakota
    13. South Dakota
    14. Nebraska
    15. Kansas
    16. Oklahoma
    17. Texas
    18. Montana
    19. Wyoming
    20. Colorado
    21. Maine

22. New Hampshire
23. Massachusetts
24. Rhode Island
25. Connecticut
26. New York
27. New Jersey
28. Maryland
29. North Carolina
30. South Carolina
31. Georgia
32. Florida
33. Delaware
34. Virginia

Bằng cách lấy dữ liệu từ gần trung tâm từng tiểu bang.

```
# List of some U.S. regions with coordinates
regions = [
    # central and river plains
    ("Tennessee", 35.6000, -88.8000),
    ("Kentucky", 37.8393, -84.2700),
    ("Mississippi", 32.3547, -89.3985),
    ("Louisiana", 31.2000, -92.4000),
    ("New Mexico", 34.5199, -105.8701),
    ("Iowa", 41.6000, -93.6000),
    ("Missouri", 38.5739, -92.6038),
    ("Illinois", 39.8000, -89.6000),
    ("Indiana", 39.8000, -86.1000),
    ("Ohio", 40.0000, -83.0000),
    ("Arkansas", 34.7465, -92.2896),

    ("North Dakota", 46.8000, -100.8000),
    ("South Dakota", 44.2998, -99.4388),
    ("Nebraska", 41.5000, -99.7000),
    ("Kansas", 38.5000, -98.0000),
    ("Oklahoma", 35.5000, -97.5000),
    ("Texas", 35.2000, -101.8000),
    ("Montana", 47.1000, -104.7000),
    ("Wyoming", 42.1000, -104.2000),
    ("Colorado", 39.5501, -105.7821),

    # east coast
    ("Maine", 44.3106, -69.7795),
    ("New Hampshire", 43.2081, -71.5376),
    ("Massachusetts", 42.3601, -71.0589),
    ("Rhode Island", 41.8240, -71.4128),
    ("Connecticut", 41.7658, -72.6734),
```

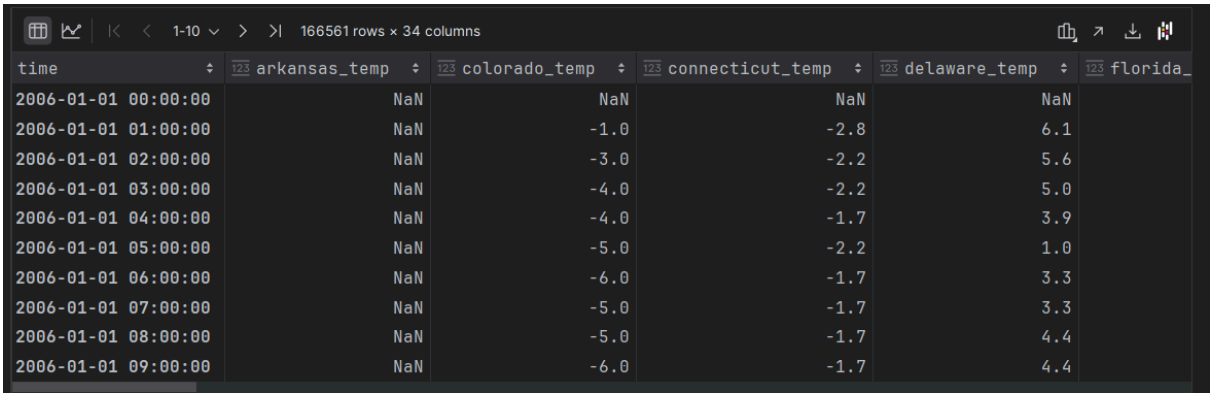
```
("New York", 40.7128, -74.0060),
("New Jersey", 40.2206, -74.7699),
("Maryland", 38.9784, -76.4922),
("North Carolina", 35.7796, -78.6382),
("South Carolina", 34.0007, -81.0348),
("Georgia", 33.7490, -84.3880),
("Florida", 30.4383, -84.2807),
("Delaware", 39.0000, -75.5000),
("Virginia", 37.5000, -78.7500)
```

- Trong quá trình thu thập, nhiều trạm quan sát không có sẵn dữ liệu với độ chia theo giờ (hourly) trong phần lớn thời gian mà nhóm muốn quan sát. Do vậy, nhóm tiến hành tìm tọa độ của các trạm quan sát khác trong cùng bang với dữ liệu đầy đủ hơn để fill vào những phần còn thiếu trong dữ liệu.

```python
# Collect data
all_data = []
for name, lat, lon in tqdm(regions, desc="Fetching weather"):
    stations = Stations().nearby(lat, lon).fetch(10)
    for station_id, row in stations.iterrows():
        if row['hourly_start'] <= start and row['hourly_end'] >= end:
            df = Hourly(station_id, start, end).fetch()
            break
    df.reset_index(inplace=True)
    df['region'] = name
    all_data.append(df[['time', 'region', 'temp']])

# Combine and save
df_all = pd.concat(all_data, ignore_index=True)
```

- Transform bảng dữ liệu để dữ liệu của mỗi bang nằm trong một cột nhằm thuận tiện cho quá trình xử lý. Dữ liệu sau khi transform gồm 35 cột (1 cột thời gian và 34 bang) cùng khoảng 166561 dòng và được lưu vào file CSV với kích thước 30MB

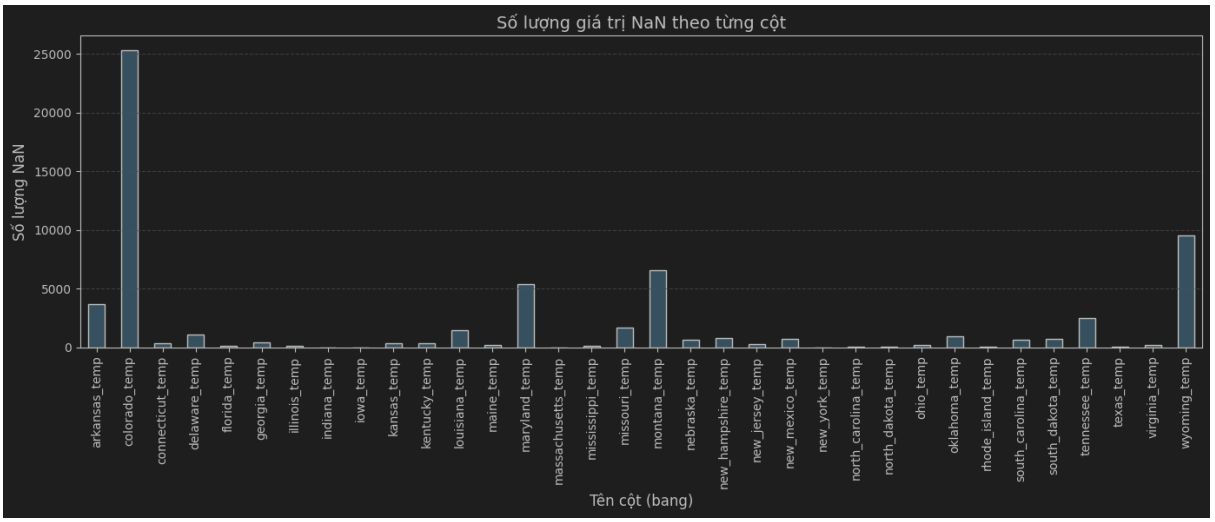


1-10 166561 rows × 34 columns

time	arkansas_temp	colorado_temp	connecticut_temp	delaware_temp	florida_
2006-01-01 00:00:00	NaN	NaN	NaN	NaN	
2006-01-01 01:00:00	NaN	-1.0	-2.8	6.1	
2006-01-01 02:00:00	NaN	-3.0	-2.2	5.6	
2006-01-01 03:00:00	NaN	-4.0	-2.2	5.0	
2006-01-01 04:00:00	NaN	-4.0	-1.7	3.9	
2006-01-01 05:00:00	NaN	-5.0	-2.2	1.0	
2006-01-01 06:00:00	NaN	-6.0	-1.7	3.3	
2006-01-01 07:00:00	NaN	-5.0	-1.7	3.3	
2006-01-01 08:00:00	NaN	-5.0	-1.7	4.4	
2006-01-01 09:00:00	NaN	-6.0	-1.7	4.4	

Data Preparation

1. Xử lý missing data



- Dựa theo thống kê số liệu trên, có thể thấy data NaN vẫn còn tồn đọng rải rác xuyên suốt dataset, nhất là ở hai bang Colorado và Wyoming, một phần do hai bang không thật sự có một trạm đo nhiệt độ tốt để có thể có dữ liệu đầy đủ hơn.
- Để fill vào các ô trống này, nhóm quyết định lấy trung bình nhiệt độ tại cùng thời điểm bị thiếu dữ liệu của 7 ngày trước và sau ngày hôm đó làm giá trị cho datapoint bị thiếu. Việc này đảm bảo dữ liệu có sự tương đồng nhất định với những ngày gần kề nó, một điều tương đối logic và thực tiễn trong đa số các trường hợp.
- Dưới đây là kết quả sau khi đã được xử lý giá trị NaN, có thể thấy từ một bảng dữ liệu có nhiều missing values, giờ đây đã thành một dataframe được lấp đầy các giá trị nhiệt độ.

5 rows 5 rows × 34 columns Static Output

time	arkansas_temp	colorado_temp	connecticut_temp	delaware_temp	florida_temp	geor
2006-01-01 00:00:00	11.6	-7.285714	-3.172066e-17	4.585714	19.4	
2006-01-01 01:00:00	10.4	-1.000000	-2.800000e+00	6.100000	19.4	
2006-01-01 02:00:00	9.4	-3.000000	-2.200000e+00	5.600000	17.8	
2006-01-01 03:00:00	8.0	-4.000000	-2.200000e+00	5.000000	16.1	
2006-01-01 04:00:00	6.8	-4.000000	-1.700000e+00	3.900000	15.0	

```
df.isna().sum().sum()
```

Executed at 2025.05.20 01:30:56 in 24ms

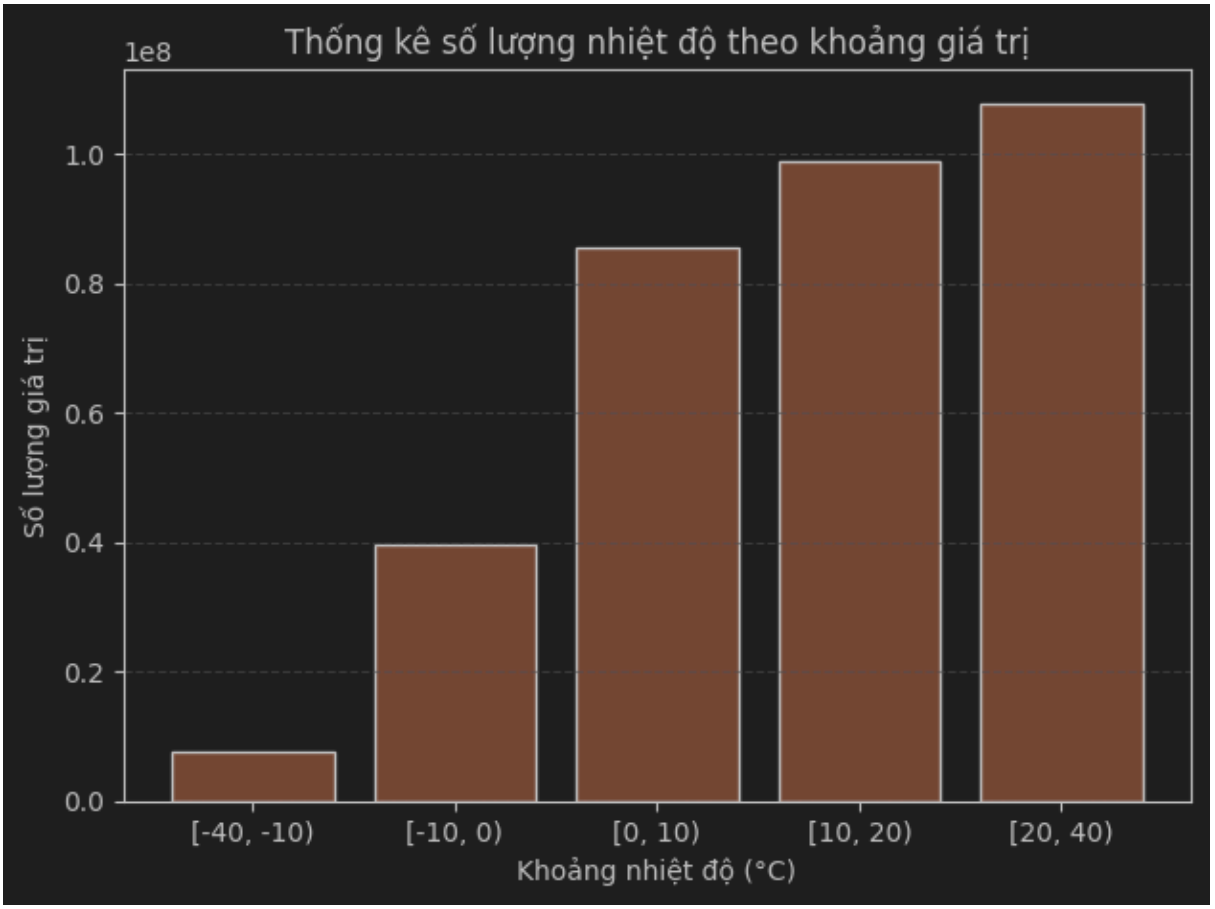
```
0
```

2. Nội suy dữ liệu để sinh data theo phút (Interpolation)

- Để tăng tính liên tục của chuỗi thời gian và tăng độ thực tế của dữ liệu, nhóm thực hiện nội suy dữ liệu từ độ chia giờ sang phút.
- Phương pháp nội suy sử dụng:
  - **Linear interpolation** cho nhiệt độ của tất cả các bang.
  - Nội suy được thực hiện theo trục thời gian để đảm bảo phù hợp với bản chất dữ liệu thời tiết time-series.
- Sau nội suy, data có 35 cột và 9993660 dòng và được chia theo phút cũng như lưu vào file CSV với kích thước gần 2GB.

time	arkansas_temp	colorado_temp	connecticut_temp	delaware_temp	florida_temp	geor
2006-01-01 00:00:00	11.6	-7.3	-0.0	4.6	19.4	
2006-01-01 00:01:00	11.6	-7.2	-0.0	4.6	19.4	
2006-01-01 00:02:00	11.6	-7.1	-0.1	4.6	19.4	
2006-01-01 00:03:00	11.5	-7.0	-0.1	4.7	19.4	
2006-01-01 00:04:00	11.5	-6.9	-0.2	4.7	19.4	

3. Dự đoán nhiệt độ trung bình



Thông qua biểu đồ cột thống kê số lượng giá trị trong từng khoảng, nhóm dự đoán nhiệt độ trung bình sẽ rời vào khoảng 10-20 độ C nhờ số lượng lớn nhiệt độ rơi vào 3 khoảng 0-10, 10-20 và 20-40.

Nhóm cũng nhận thấy rằng tỷ lệ nhiệt độ dưới -10°C là khá thấp, cho thấy những đợt giá rét sâu không thường xuyên xảy ra, và có thể chủ yếu tập trung ở một số bang như Colorado, New Hampshire hoặc Montana. Các nhiệt độ này sẽ không ảnh hưởng quá lớn đến nhiệt độ trung bình của toàn bộ miền Đông nước Mỹ.

Ngôn ngữ sử dụng

Python cho tác vụ thu thập data, tiền xử lý.

- Các thư viện được sử dụng: pandas, numpy, matplotlib, pandas, tqdm, sys, meteostat.
- Các file được sử dụng: crawl_temperature_data.ipynb, data-preprocessing.ipynb.

Java + Hadoop MapReduce cho tác vụ xử lý task chính.

- Các file được sử dụng: Các file mapper, reducer và driver.

Map Reduce

Nhóm quyết định ứng dụng Map Reduce cho ba hướng tính toán để hiểu rõ data hơn: tính trung bình của toàn bộ dữ liệu nhiệt độ xuyên suốt các bang trong giai đoạn 2006-2005, tính trung bình nhiệt độ từng bang ở miên đông nước Mỹ trong giai đoạn 2006-2005 và cuối cùng là tính trung bình nhiệt độ từng năm của miền đông nước Mỹ.

1. Mapper

- **MeanTemperatureMapper**: Đây là file Java thực hiện quá trình ánh xạ tất cả các giá trị nhiệt độ đầu vào thành cặp dạng `("temp", value)`. Mục đích là để tính trung bình toàn bộ các giá trị nhiệt độ trong tập dữ liệu.

```java
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Mapper;

import java.io.IOException;

public class MeanTemperatureMapper extends Mapper<LongWritable, Text, Text, FloatWritable> {
    boolean isHeader = true;

    @Override
    protected void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
        String line = value.toString();

        if (isHeader && line.contains("time")) {
            isHeader = false;
            return;
        }

        String[] parts = line.split(",");
        for (int i = 1; i < parts.length; i++) {
            try {
                float temp = Float.parseFloat(parts[i]);
                context.write(new Text("temp"), new FloatWritable(temp));
            } catch (NumberFormatException ignored) {
            }
        }
    }
}
```

- **MeanTemperatureYearlyMapper**: Tương tự như Mapper trên, nhưng lần này mỗi giá trị nhiệt độ được ánh xạ thành cặp `(<year>, value)`, tương ứng với từng năm. Điều này cho phép tính trung bình nhiệt độ theo từng năm.

```
import java.io.IOException;

public class MeanTemperatureYearlyMapper extends Mapper<LongWritable, Text, Text, FloatWritable> {

    @Override
    protected void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {

        String line = value.toString();

        // Skip header
        if (key.get() == 0 && line.toLowerCase().contains("time")) {
            return;
        }

        String[] parts = line.split(",");
        if (parts.length < 2) return; // skip malformed lines

        String timeStr = parts[0].trim();  // e.g. "2006-01-01 00:00:00"
        String year = timeStr.split("-")[0];  // Extract "2006"

        for (int i = 1; i < parts.length; i++) {
            try {
                float temp = Float.parseFloat(parts[i].trim());
                context.write(new Text(year), new FloatWritable(temp));
            } catch (NumberFormatException ignored) {
            }
        }
    }
}
```

- **Mean TemperatureStateMapper**: Trong file này các giá trị nhiệt độ được xử lý dưới dạng `(<state>, value)`, tương ứng với từng bang. File này dùng để tính trung bình nhiệt độ theo từng bang.

```
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Mapper;

import java.io.IOException;

public class MeanTemperatureStateMapper extends Mapper<LongWritable, Text, Text, FloatWritable> {
    private static final String HEADER_PREFIX = "time";
    private String[] headers;
    private boolean isHeaderParsed = false;

    @Override
    protected void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {

        String line = value.toString().trim();

        if (!isHeaderParsed && line.toLowerCase().startsWith(HEADER_PREFIX)) {
            headers = line.split(",");
            isHeaderParsed = true;
            return;
        }

        if (!isHeaderParsed || headers == null) return;

        String[] parts = line.split(",");
        if (parts.length != headers.length) return;

        for (int i = 1; i < parts.length; i++) {
            try {
                float temp = Float.parseFloat(parts[i]);
                String state = headers[i].replace("_temp", "").trim().toLowerCase();

                context.write(new Text(state), new FloatWritable(temp));
            } catch (NumberFormatException ignored) {}
        }
    }
}
```

2. Reducer

- **MeanTemperatureReducer**: Thực hiện tổng hợp tất cả các cặp `("temp", value)` được gửi từ Mapper, tính toán trung bình nhiệt độ trên toàn bộ tập dữ liệu và xuất ra kết quả cuối cùng.

```java
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Reducer;

import java.io.IOException;

public class MeanTemperatureReducer extends Reducer<Text, FloatWritable, Text, FloatWritable> {

    @Override
    protected void reduce(Text key, Iterable<FloatWritable> values, Context context)
            throws IOException, InterruptedException {

        double sum = 0.0;
        int count = 0;

        for (FloatWritable val : values) {
            sum += val.get();
            count += 1;
        }

        float mean = (float)(sum / count);


        context.write(new Text(key.toString() + "_TotalSum"), new FloatWritable((float) sum));
        context.write(new Text(key.toString() + "_TotalCount"), new FloatWritable(count));
        context.write(new Text(key.toString() + "_Mean"), new FloatWritable(mean));
    }
}
```

- **MeanTemperatureYearlyReducer**: Nhận các cặp `(<year>, value)` từ Mapper, tiến hành tính trung bình nhiệt độ theo từng năm và ghi kết quả ra output.

```java
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Reducer;

import java.io.IOException;

public class MeanTemperatureYearlyReducer extends Reducer<Text, FloatWritable, Text, FloatWritable> {

    @Override
    protected void reduce(Text key, Iterable<FloatWritable> values, Context context)
            throws IOException, InterruptedException {

        double sum = 0.0;
        int count = 0;

        for (FloatWritable val : values) {
            sum += val.get();
            count++;
        }

        float mean = (float)(sum / count);

        context.write(new Text(key.toString() + "_MeanTemperature"), new FloatWritable(mean));
    }
}
```

- **MeanTemperatureByStateReducer**: Tổng hợp các giá trị `(<state>, value)` từ Mapper, sau đó tính toán trung bình nhiệt độ cho từng bang và xuất kết quả.

```java
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Reducer;

import java.io.IOException;

public class MeanTemperatureStateReducer extends Reducer<Text, FloatWritable, Text, FloatWritable> {

    @Override
    protected void reduce(Text key, Iterable<FloatWritable> values, Context context)
            throws IOException, InterruptedException {

        double sum = 0.0;
        int count = 0;

        for (FloatWritable val : values) {
            sum += val.get();
            count++;
        }

        float mean = (float)(sum / count);

        context.write(new Text(key.toString() + "_MeanTemperature"), new FloatWritable(mean));
    }
}
```

3. Driver

Cả ba task trên đều sẽ sử dụng một file Driver điều khiển cho job của nó, file này sẽ thiết lập để sử dụng các Mapper và Reducer cho phép xử lý dữ liệu nhiệt độ cho toàn bộ dữ liệu hay theo từng năm hoặc từng bang.

```java
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Reducer;

import java.io.IOException;

public class MeanTemperatureYearlyReducer extends Reducer<Text, FloatWritable, Text, FloatWritable> {

    @Override
    protected void reduce(Text key, Iterable<FloatWritable> values, Context context)
            throws IOException, InterruptedException {

        double sum = 0.0;
        int count = 0;

        for (FloatWritable val : values) {
            sum += val.get();
            count++;
        }

        float mean = (float)(sum / count);

        context.write(new Text(key.toString() + "_MeanTemperature"), new FloatWritable(mean));
    }
}
```

```java
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class MeanTemperatureStateDriver {
    public static void main(String[] args) throws Exception {
        if (args.length != 2) {
            System.err.println("Usage: MeanTemperature <input path> <output path>");
            System.exit(-1);
        }

        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "Mean Temperature");

        job.setJarByClass(MeanTemperatureStateDriver.class);
        job.setMapperClass(MeanTemperatureStateMapper.class);
        job.setReducerClass(MeanTemperatureStateReducer.class);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(FloatWritable.class);
        job.setNumReduceTasks(1);

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

Running Hadoop

```
\Windows\System32>hadoop jar C:\Users\Admin\calculate_mean_temperature.jar MeanTemperatureDriver /user/input/us_eastern_regions_minutely_temperature_2006_to_2025.csv /user/output/temp_avg
025-05-19 22:27:03,261 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
025-05-19 22:27:03,765 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
025-05-19 22:27:03,789 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/Admin/.staging/job_1747668290848_0001
025-05-19 22:27:04,461 INFO input.FileInputFormat: Total input files to process : 1
025-05-19 22:27:04,945 INFO mapreduce.JobSubmitter: number of splits:14
025-05-19 22:27:05,057 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1747668290848_0001
025-05-19 22:27:05,057 INFO mapreduce.JobSubmitter: Executing with tokens: []
025-05-19 22:27:05,172 INFO conf.Configuration: resource-types.xml not found
025-05-19 22:27:05,173 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
025-05-19 22:27:05,544 INFO impl.YarnClientImpl: Submitted application application_1747668290848_0001
025-05-19 22:27:05,585 INFO mapreduce.Job: The url to track the job: http://DESKTOP-FTFDCHT:8088/proxy/application_1747668290848_0001/
025-05-19 22:27:05,587 INFO mapreduce.Job: Running job: job_1747668290848_0001
025-05-19 22:27:12,742 INFO mapreduce.Job: Job job_1747668290848_0001 running in uber mode : false
025-05-19 22:27:39,889 INFO mapreduce.Job:  map 21% reduce 0%
025-05-19 22:27:46,055 INFO mapreduce.Job:  map 28% reduce 0%
025-05-19 22:28:29,465 INFO mapreduce.Job:  map 63% reduce 0%
025-05-19 22:28:30,475 INFO mapreduce.Job:  map 65% reduce 0%
025-05-19 22:28:35,551 INFO mapreduce.Job:  map 69% reduce 0%
025-05-19 22:28:50,970 INFO mapreduce.Job:  map 84% reduce 0%
025-05-19 22:28:51,989 INFO mapreduce.Job:  map 85% reduce 0%
025-05-19 22:28:53,006 INFO mapreduce.Job:  map 86% reduce 0%
025-05-19 22:29:07,220 INFO mapreduce.Job:  map 86% reduce 17%
025-05-19 22:29:08,237 INFO mapreduce.Job:  map 95% reduce 17%
025-05-19 22:29:12,293 INFO mapreduce.Job:  map 97% reduce 17%
025-05-19 22:29:13,310 INFO mapreduce.Job:  map 97% reduce 29%
025-05-19 22:29:14,319 INFO mapreduce.Job:  map 98% reduce 29%
025-05-19 22:29:19,365 INFO mapreduce.Job:  map 100% reduce 31%
025-05-19 22:29:25,437 INFO mapreduce.Job:  map 100% reduce 34%
025-05-19 22:29:31,504 INFO mapreduce.Job:  map 100% reduce 41%
025-05-19 22:29:37,588 INFO mapreduce.Job:  map 100% reduce 49%
025-05-19 22:29:43,658 INFO mapreduce.Job:  map 100% reduce 57%
025-05-19 22:29:49,709 INFO mapreduce.Job:  map 100% reduce 65%
025-05-19 22:29:55,772 INFO mapreduce.Job:  map 100% reduce 69%
025-05-19 22:30:01,850 INFO mapreduce.Job:  map 100% reduce 72%
025-05-19 22:30:06,913 INFO mapreduce.Job:  map 100% reduce 74%
025-05-19 22:30:12,990 INFO mapreduce.Job:  map 100% reduce 77%
025-05-19 22:30:19,063 INFO mapreduce.Job:  map 100% reduce 80%
025-05-19 22:30:25,132 INFO mapreduce.Job:  map 100% reduce 82%
025-05-19 22:30:31,206 INFO mapreduce.Job:  map 100% reduce 85%
025-05-19 22:30:37,273 INFO mapreduce.Job:  map 100% reduce 88%
025-05-19 22:30:43,347 INFO mapreduce.Job:  map 100% reduce 91%
025-05-19 22:30:49,429 INFO mapreduce.Job:  map 100% reduce 94%
025-05-19 22:30:55,485 INFO mapreduce.Job:  map 100% reduce 96%
025-05-19 22:31:01,547 INFO mapreduce.Job:  map 100% reduce 99%
025-05-19 22:31:02,560 INFO mapreduce.Job:  map 100% reduce 100%
025-05-19 22:31:03,582 INFO mapreduce.Job: Job job_1747668290848_0001 completed successfully
```

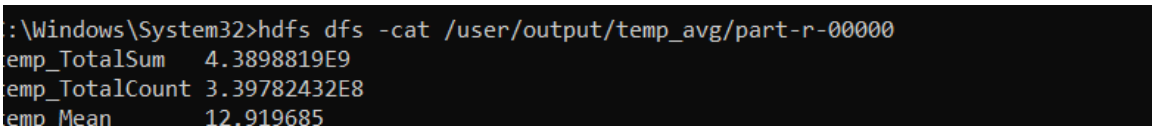
```
:\Windows\System32>hadoop jar C:\Users\Admin\calculate_yearly_mean_temperature.jar MeanTemperatureYearlyDriver /user/input/us_eastern_regions_minutely_temperature_2006_to_2025.csv /user/output/temp_avg_yearly
025-05-20 02:05:15,448 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
025-05-20 02:05:15,933 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
025-05-20 02:05:15,959 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/Admin/.staging/job_1747681353611_0001
025-05-20 02:05:16,663 INFO input.FileInputFormat: Total input files to process : 1
025-05-20 02:05:16,738 INFO mapreduce.JobSubmitter: number of splits:14
025-05-20 02:05:16,859 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1747681353611_0001
025-05-20 02:05:16,859 INFO mapreduce.JobSubmitter: Executing with tokens: []
025-05-20 02:05:16,991 INFO conf.Configuration: resource-types.xml not found
025-05-20 02:05:16,992 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
025-05-20 02:05:17,398 INFO impl.YarnClientImpl: Submitted application application_1747681353611_0001
025-05-20 02:05:17,444 INFO mapreduce.Job: The url to track the job: http://DESKTOP-FTFDCHT:8088/proxy/application_1747681353611_0001/
025-05-20 02:05:17,444 INFO mapreduce.Job: Running job: job_1747681353611_0001
025-05-20 02:05:26,653 INFO mapreduce.Job: Job job_1747681353611_0001 running in uber mode : false
025-05-20 02:05:34,866 INFO mapreduce.Job:  map 0% reduce 0%
025-05-20 02:05:53,666 INFO mapreduce.Job:  map 7% reduce 0%
025-05-20 02:05:59,810 INFO mapreduce.Job:  map 11% reduce 0%
025-05-20 02:06:05,904 INFO mapreduce.Job:  map 15% reduce 0%
025-05-20 02:06:12,007 INFO mapreduce.Job:  map 19% reduce 0%
025-05-20 02:06:17,104 INFO mapreduce.Job:  map 20% reduce 0%
025-05-20 02:06:18,135 INFO mapreduce.Job:  map 23% reduce 0%
025-05-20 02:06:23,225 INFO mapreduce.Job:  map 26% reduce 0%
025-05-20 02:06:24,239 INFO mapreduce.Job:  map 27% reduce 0%
025-05-20 02:06:29,322 INFO mapreduce.Job:  map 29% reduce 0%
025-05-20 02:06:35,420 INFO mapreduce.Job:  map 32% reduce 0%
025-05-20 02:06:41,508 INFO mapreduce.Job:  map 36% reduce 0%
025-05-20 02:06:47,597 INFO mapreduce.Job:  map 39% reduce 0%
025-05-20 02:06:52,694 INFO mapreduce.Job:  map 40% reduce 0%
025-05-20 02:06:53,716 INFO mapreduce.Job:  map 42% reduce 0%
025-05-20 02:06:54,734 INFO mapreduce.Job:  map 43% reduce 0%
025-05-20 02:07:18,274 INFO mapreduce.Job:  map 44% reduce 0%
025-05-20 02:07:19,291 INFO mapreduce.Job:  map 45% reduce 0%
025-05-20 02:07:21,328 INFO mapreduce.Job:  map 49% reduce 0%
025-05-20 02:07:24,379 INFO mapreduce.Job:  map 49% reduce 12%
025-05-20 02:07:25,393 INFO mapreduce.Job:  map 50% reduce 12%
025-05-20 02:07:27,421 INFO mapreduce.Job:  map 52% reduce 12%
025-05-20 02:07:29,454 INFO mapreduce.Job:  map 52% reduce 14%
025-05-20 02:07:30,469 INFO mapreduce.Job:  map 53% reduce 14%
025-05-20 02:07:33,510 INFO mapreduce.Job:  map 56% reduce 14%
025-05-20 02:07:36,534 INFO mapreduce.Job:  map 57% reduce 14%
025-05-20 02:07:39,558 INFO mapreduce.Job:  map 59% reduce 14%
025-05-20 02:07:42,582 INFO mapreduce.Job:  map 61% reduce 14%
025-05-20 02:07:45,624 INFO mapreduce.Job:  map 63% reduce 14%
025-05-20 02:07:48,655 INFO mapreduce.Job:  map 65% reduce 14%
025-05-20 02:07:51,679 INFO mapreduce.Job:  map 67% reduce 14%
025-05-20 02:07:54,702 INFO mapreduce.Job:  map 68% reduce 14%
025-05-20 02:07:57,744 INFO mapreduce.Job:  map 70% reduce 14%
```

```
C:\Windows\System32>hadoop jar C:\Users\Admin\calculate_state_mean_temperature.jar MeanTemperatureStateDriver /user/input/us_eastern_regions_minutely_temperature_2006_to_2025.csv /user/output/temppp
2025-05-20 17:30:10,628 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-05-20 17:30:11,094 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-20 17:30:11,109 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/Admin/.staging/job_1747734615612_0003
2025-05-20 17:30:11,279 INFO input.FileInputFormat: Total input files to process : 1
2025-05-20 17:30:11,337 INFO mapreduce.JobSubmitter: number of splits:14
2025-05-20 17:30:11,454 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1747734615612_0003
2025-05-20 17:30:11,455 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-20 17:30:11,579 INFO conf.Configuration: resource-types.xml not found
2025-05-20 17:30:11,580 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-05-20 17:30:11,637 INFO impl.YarnClientImpl: Submitted application application_1747734615612_0003
2025-05-20 17:30:11,683 INFO mapreduce.Job: The url to track the job: http://DESKTOP-FTFDCHT:8088/proxy/application_1747734615612_0003/
2025-05-20 17:30:11,686 INFO mapreduce.Job: Running job: job_1747734615612_0003
2025-05-20 17:30:19,863 INFO mapreduce.Job: Job job_1747734615612_0003 running in uber mode : false
2025-05-20 17:30:19,866 INFO mapreduce.Job:  map 0% reduce 0%
2025-05-20 17:30:31,195 INFO mapreduce.Job:  map 7% reduce 0%
2025-05-20 17:30:32,255 INFO mapreduce.Job:  map 36% reduce 0%
2025-05-20 17:30:41,439 INFO mapreduce.Job:  map 37% reduce 0%
2025-05-20 17:30:43,477 INFO mapreduce.Job:  map 73% reduce 0%
2025-05-20 17:30:52,678 INFO mapreduce.Job:  map 74% reduce 0%
2025-05-20 17:30:56,408 INFO mapreduce.Job:  map 81% reduce 0%
2025-05-20 17:30:57,432 INFO mapreduce.Job:  map 95% reduce 0%
2025-05-20 17:30:59,461 INFO mapreduce.Job:  map 96% reduce 0%
2025-05-20 17:31:05,527 INFO mapreduce.Job:  map 97% reduce 31%
2025-05-20 17:31:10,590 INFO mapreduce.Job:  map 98% reduce 31%
2025-05-20 17:31:16,668 INFO mapreduce.Job:  map 99% reduce 31%
2025-05-20 17:31:19,726 INFO mapreduce.Job:  map 100% reduce 31%
2025-05-20 17:31:29,856 INFO mapreduce.Job:  map 100% reduce 97%
2025-05-20 17:31:30,869 INFO mapreduce.Job:  map 100% reduce 100%
2025-05-20 17:31:31,267 INFO mapreduce.Job: Job job_1747734615612_0003 completed successfully
2025-05-20 17:31:31,329 INFO mapreduce.Job: Counters: 51
       File System Counters
              FILE: Number of bytes read=788216636
              FILE: Number of bytes written=1186304105
              FILE: Number of read operations=0
              FILE: Number of large read operations=0
              FILE: Number of write operations=0
              HDFS: Number of bytes read=1831605142
              HDFS: Number of bytes written=1201
              HDFS: Number of read operations=47
              HDFS: Number of large read operations=0
              HDFS: Number of write operations=2
              HDFS: Number of bytes read erasure-coded=0
       Job Counters
              Killed map tasks=1
              Launched map tasks=15
              Launched reduce tasks=1
              Data-local map tasks=15
```

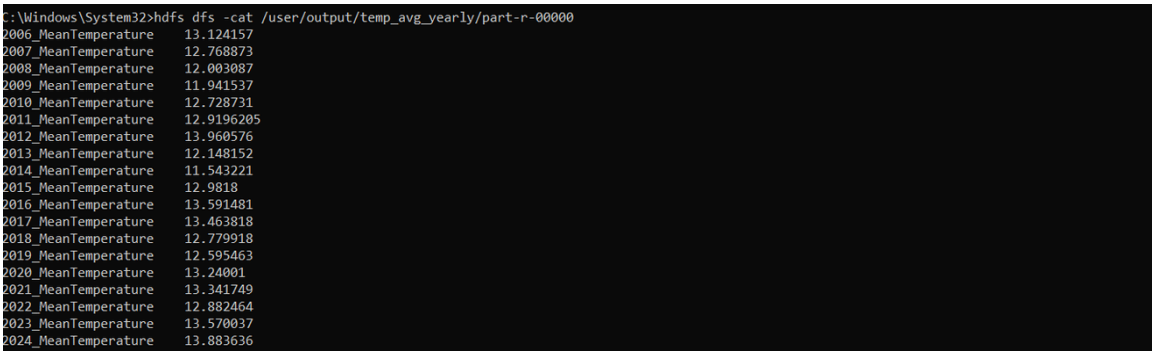
```
              HDFS: Number of bytes read erasure-coded=0
       Job Counters
              Killed map tasks=1
              Launched map tasks=15
              Launched reduce tasks=1
              Data-local map tasks=15
              Total time spent by all maps in occupied slots (ms)=234454
              Total time spent by all reduces in occupied slots (ms)=46567
              Total time spent by all map tasks (ms)=234454
              Total time spent by all reduce tasks (ms)=46567
              Total vcore-milliseconds taken by all map tasks=234454
              Total vcore-milliseconds taken by all reduce tasks=46567
              Total megabyte-milliseconds taken by all map tasks=240080896
              Total megabyte-milliseconds taken by all reduce tasks=47684608
       Map-Reduce Framework
              Map input records=9993602
              Map output records=25046134
              Map output bytes=344016017
              Map output materialized bytes=394108369
              Input split bytes=2142
              Combine input records=0
              Combine output records=0
              Reduce input groups=34
              Reduce shuffle bytes=394108369
              Reduce input records=25046134
              Reduce output records=34
              Spilled Records=75138402
              Shuffled Maps =14
              Failed Shuffles=0
              Merged Map outputs=14
              GC time elapsed (ms)=4335
              CPU time spent (ms)=0
              Physical memory (bytes) snapshot=0
              Virtual memory (bytes) snapshot=0
              Total committed heap usage (bytes)=6936854528
       Shuffle Errors
              BAD_ID=0
              CONNECTION=0
              IO_ERROR=0
              WRONG_LENGTH=0
              WRONG_MAP=0
              WRONG_REDUCE=0
       File Input Format Counters
              Bytes Read=1831603000
       File Output Format Counters
              Bytes Written=1201
```

Result

- Nhiệt độ trung bình trên toàn miền đông nước Mỹ giai đoạn 2006-2025



- Nhiệt độ trung bình từng năm trên toàn miền đông nước Mỹ từ 2006-2025



- Nhiệt độ trung bình từng bang trên miền đông nước Mỹ từ 2006-2025

```
arkansas_MeanTemperature        16.32122
colorado_MeanTemperature        -2.186649
connecticut_MeanTemperature     9.841461
delaware_MeanTemperature        12.667599
florida_MeanTemperature 19.03456
georgia_MeanTemperature 15.550001
illinois_MeanTemperature        11.037681
indiana_MeanTemperature 10.78844
iowa_MeanTemperature    9.7102165
kansas_MeanTemperature  12.845024
kentucky_MeanTemperature        11.994065
louisiana_MeanTemperature       17.833996
maine_MeanTemperature   6.323285
maryland_MeanTemperature        12.514952
massachusetts_MeanTemperature   9.61259
mississippi_MeanTemperature     17.040539
missouri_MeanTemperature        12.834514
montana_MeanTemperature 6.302548
nebraska_MeanTemperature        8.596544
new_hampshire_MeanTemperature   7.2075934
new_jersey_MeanTemperature      11.2944565
new_mexico_MeanTemperature      8.921865
new_york_MeanTemperature        11.82072
north_carolina_MeanTemperature  14.943919
north_dakota_MeanTemperature    5.718866
ohio_MeanTemperature    10.202611
oklahoma_MeanTemperature        15.583606
rhode_island_MeanTemperature    9.955936
south_carolina_MeanTemperature  17.006777
south_dakota_MeanTemperature    8.268676
tennessee_MeanTemperature       14.731546
texas_MeanTemperature   12.656521
virginia_MeanTemperature        12.301749
wyoming_MeanTemperature 8.229129
```

Interpretation & Conclusion

Qua quá trình thực hiện bài toán, nhóm nhận thấy có sự gia tăng đáng kể về mức nhiệt độ trung bình cao nhất, thấp nhất, nhiệt độ ghi nhận qua các pha El Nino và La Nina cũng có sự gia tăng nhất định ở nhiều bang, chỉ ra xu hướng không thể đảo ngược của biến đổi khí hậu và nóng lên toàn cầu.

Nhìn chung với nhiệt độ trung bình chung ở khoảng 12.92 độ C khá mát mẻ, phản ánh sự đa dạng về khí hậu trong khu vực - từ vùng lạnh giá phía Bắc đến vùng nhiệt đới ẩm phía Nam

Nhiệt độ trung bình năm dao động từ 11.5 độ C đến 13.9 độ C, có xu hướng tăng nhẹ về sau, nhất là giai đoạn 2020-2024, phản ánh một phần ảnh hưởng của

dịch Covid-19 và biến đổi khí hậu toàn cầu.

Từ bảng nhiệt độ trung bình từ các bang, có thể nhận thấy các bang phía Nam như Florida, Louisiana hay Georgia có khí hậu ấm áp, cận nhiệt đới. Trong khi đó các bang Bắc-Tây như Colorado, Montana, New Hampshire có khí hậu lạnh hơn rõ rệt, có thể ảnh hưởng bởi độ cao và vĩ độ.

Các đặc điểm khí hậu thông qua kết quả tính toán trên hoàn toàn phù hợp với đặc điểm cũng như vị trí địa lý của các bang trên miền Đông Hoa Kì, phản ánh rằng các tính toán thông qua Hadoop và quá trình thu thập dữ liệu của nhóm được thực hiện chính xác và hoàn thiện.

Future works

Trong tương lai, nhóm đề xuất một số hướng phát triển mở rộng như sau:

1. **Phân tích chuyên sâu theo mùa và tháng**

   Hiện tại dữ liệu mới được xử lý theo từng năm hoặc toàn bộ. Việc phân tích theo từng mùa (xuân, hạ, thu, đông) hoặc tháng sẽ giúp hiểu rõ hơn về xu hướng nhiệt độ trong ngắn hạn và các đặc trưng thời tiết định kỳ.

2. **So sánh dữ liệu với các vùng khác**

   Mở rộng phạm vi nghiên cứu ra các vùng khác của Hoa Kỳ (miền Tây, Trung Tây, miền Nam) để so sánh đặc điểm khí hậu, tìm ra sự khác biệt về nhiệt độ và ảnh hưởng của vị trí địa lý.

3. **Kết hợp thêm các yếu tố khí hậu khác**

   Bổ sung thêm các thông số như: độ ẩm, lượng mưa, tốc độ gió,... để phân tích đa biến, từ đó đánh giá toàn diện hơn về tình hình thời tiết và môi trường.

4. **Ứng dụng học máy (Machine Learning)**

   Áp dụng các mô hình học máy để dự đoán xu hướng nhiệt độ trong tương lai, phát hiện các mẫu bất thường (anomaly detection) hoặc cảnh báo sớm biến đổi khí hậu tại từng bang.

5. **Tối ưu hóa hiệu năng xử lý dữ liệu lớn**

   Nâng cấp hệ thống xử lý bằng cách tích hợp Apache Spark, sử dụng bộ nhớ thay vì ghi đĩa như Hadoop MapReduce truyền thống, nhằm giảm thời gian chạy và tăng hiệu suất.

Team Report

- Crawl data và làm sạch: Trần Khắc Phúc Khánh + Lưu Quang Linh
- Data Interpolation và Data Preprocessing: Trần Khắc Phúc Khánh
- Code Jar file và chạy Hadoop + MapReduce: Trần Khắc Phúc Khánh + Võ Duy Quang
- Viết báo cáo: Võ Duy Quang + Lưu Quang Linh + Trần Khắc Phúc Khánh
- Làm Slides: Võ Duy Quang + Lưu Quang Linh