REPORT

Group Members:-

Manan Shah – ms3452

Sneh Vora - sv992

Palak Pabani – pp872

Vinit Kumar Dobariya – vd363

Project Problem Statement:-

The aim is to process and analyze the weather data to address the following questions:

1. **Temperature Trends**: What are the monthly and seasonal variations in dry bulb temperature?
2. **Humidity and Dew Point**: How do humidity and dew point vary over time, and how do they correlate with temperature?
3. **Wind Patterns**: What are the dominant wind directions and speeds, and how do they vary spatially and temporally?

Dataset Overview:-

The dataset is sourced from Kaggle and titled *Hourly Weather Surface Brazil Southeast Region*. It provides hourly weather data collected across various stations in Brazil's southeastern region. This project focuses on the north.csv file, a 1.55 GB subset of the dataset, which contains meteorological observations for the northern region of Brazil. Each record corresponds to hourly measurements of weather parameters, with details about location and time.

Dataset Attributes:-

The file contains 26 columns, outlined below with their relevance to the analysis:

1. **index**: Row identifier, useful for maintaining the dataset's integrity during cleaning.
2. **Data**: Date of the observation.
3. **Hora**: Hour of the observation, important for time-series analysis.
4. **PRECIPITAÇÃO TOTAL, HORÁRIO (mm)**: Total hourly precipitation, a key indicator for rainfall patterns.
5. **PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORARIA (mB)**: Atmospheric pressure at the station, crucial for identifying weather conditions.
6. **PRESSÃO ATMOSFERICA MAX.NA HORA ANT. (AUT) (mB)** and **MIN. NA HORA ANT. (AUT) (mB)**: Max and min pressure values in the previous hour, useful for detecting pressure trends.
7. **RADIACAO GLOBAL (Kj/m²)**: Global radiation, indicating solar energy received, essential for analyzing sunlight variations.
8. **TEMPERATURA DO AR - BULBO SECO, HORARIA (°C)**: Hourly dry bulb air temperature, a primary metric for temperature analysis.
9. **TEMPERATURA DO PONTO DE ORVALHO (°C)**: Dew point temperature, indicating moisture content in the air.
10. **TEMPERATURA MÁXIMA/MÍNIMA NA HORA ANT. (AUT) (°C)**: Maximum and minimum air temperature in the previous hour.
11. **TEMPERATURA ORVALHO MAX./MIN. NA HORA ANT. (AUT) (°C)**: Max and min dew point temperature in the previous hour, reflecting moisture trends.
12. **UMIDADE REL. MAX./MIN. NA HORA ANT. (AUT) (%)**: Max and min relative humidity in the previous hour.
13. **UMIDADE RELATIVA DO AR, HORARIA (%)**: Hourly relative humidity, a key factor for weather conditions.

14. **VENTO, DIREÇÃO HORARIA (gr)** and **VENTO, VELOCIDADE HORARIA (m/s)**: Wind direction and speed, critical for studying wind patterns.

15. **VENTO, RAJADA MAXIMA (m/s)**: Maximum wind gust, useful for identifying extreme wind conditions.

16. **region, state, station, station_code**: Metadata indicating geographical location of the station.

17. **latitude, longitude, height**: Geographical coordinates and elevation of the weather station.

Detail Description of Jobs:-

1. Temperature Trends

Objective:

Analyze monthly and seasonal variations in the dry bulb temperature.

Algorithm:

1. **Mapper (AverageTemperatureMapper):**
   - Reads the dataset line by line.
   - Parses the TEMPERATURA DO AR - BULBO SECO column (dry bulb air temperature).
   - Extracts the station code (column 22) and temperature (column 9).
   - Outputs key-value pairs where the key is the station code, and the value is the temperature.

Key: Weather station code (column 22)
Value: Temperature (column 9)

   - Skips malformed records or invalid temperature values.

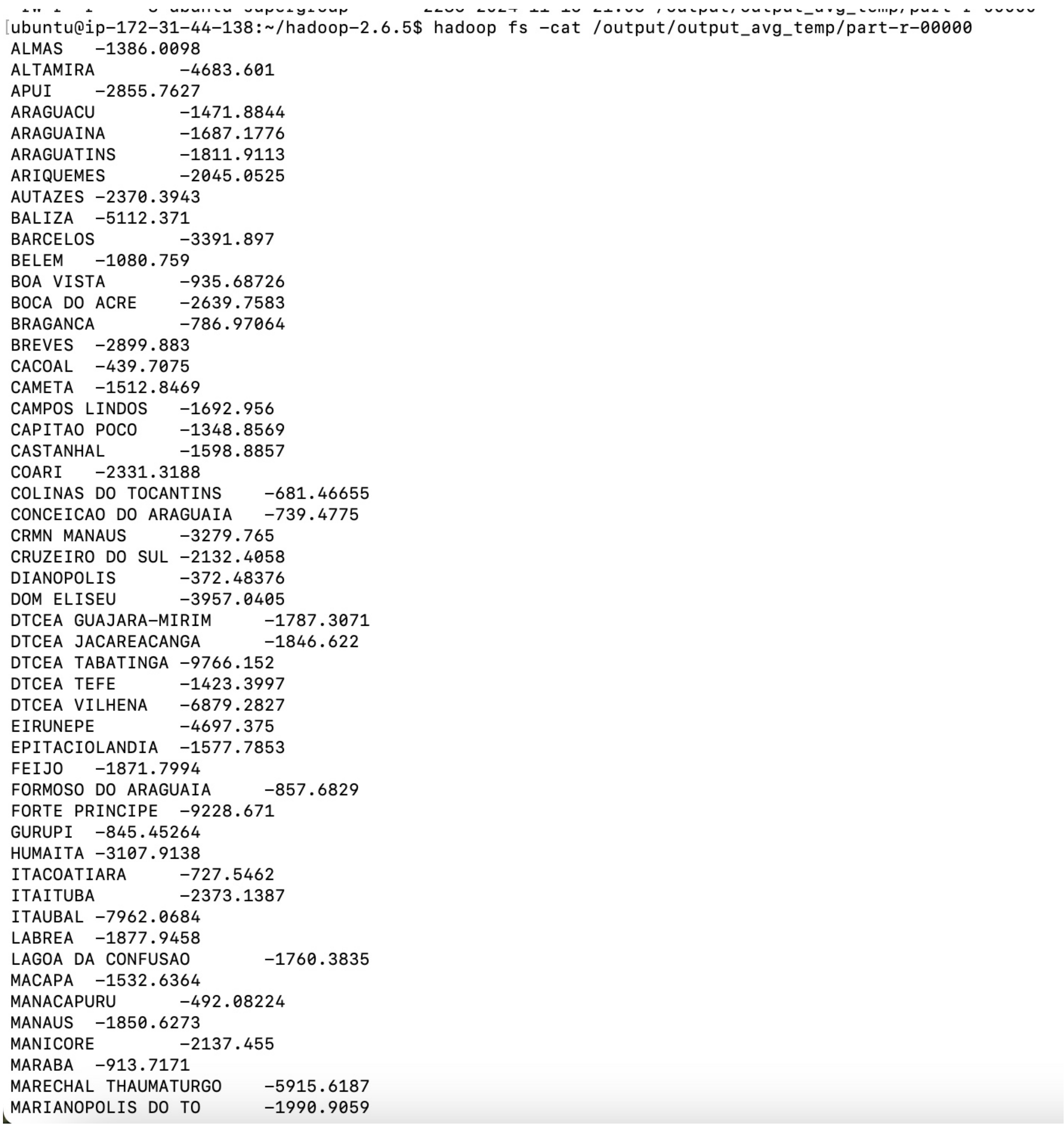
2. **Reducer (AverageTemperatureReducer):**
   - Receives a list of temperature values for each station.
   - Computes the average temperature by summing up the values and dividing by the count.

- Outputs the average temperature for each weather station.

Output Key: Weather station code
Output Value: Average temperature

Outcome:

Station-wise average temperatures can be aggregated further (e.g., by month or season) for a detailed analysis of temperature variations.



```
[ubuntu@ip-172-31-44-138:~/hadoop-2.6.5$ hadoop fs -cat /output/output_avg_temp/part-r-00000
ALMAS   -1386.0098
ALTAMIRA        -4683.601
APUI    -2855.7627
ARAGUACU        -1471.8844
ARAGUAINA       -1687.1776
ARAGUATINS      -1811.9113
ARIQUEMES       -2045.0525
AUTAZES -2370.3943
BALIZA  -5112.371
BARCELOS        -3391.897
BELEM   -1080.759
BOA VISTA       -935.68726
BOCA DO ACRE    -2639.7583
BRAGANCA        -786.97064
BREVES  -2899.883
CACOAL  -439.7075
CAMETA  -1512.8469
CAMPOS LINDOS   -1692.956
CAPITAO POCO    -1348.8569
CASTANHAL       -1598.8857
COARI   -2331.3188
COLINAS DO TOCANTINS    -681.46655
CONCEICAO DO ARAGUAIA   -739.4775
CRMN MANAUS     -3279.765
CRUZEIRO DO SUL -2132.4058
DIANOPOLIS      -372.48376
DOM ELISEU      -3957.0405
DTCEA GUAJARA-MIRIM     -1787.3071
DTCEA JACAREACANGA      -1846.622
DTCEA TABATINGA -9766.152
DTCEA TEFE      -1423.3997
DTCEA VILHENA   -6879.2827
EIRUNEPE        -4697.375
EPITACIOLANDIA  -1577.7853
FEIJO   -1871.7994
FORMOSO DO ARAGUAIA     -857.6829
FORTE PRINCIPE  -9228.671
GURUPI  -845.45264
HUMAITA -3107.9138
ITACOATIARA     -727.5462
ITAITUBA        -2373.1387
ITAUBAL -7962.0684
LABREA  -1877.9458
LAGOA DA CONFUSAO       -1760.3835
MACAPA  -1532.6364
MANACAPURU      -492.08224
MANAUS  -1850.6273
MANICORE        -2137.455
MARABA  -913.7171
MARECHAL THAUMATURGO    -5915.6187
MARIANOPOLIS DO TO      -1990.9059
```

2. Total Precipitation

Objective:

Analyze variations in humidity and dew point and their correlation with temperature.

Algorithm (Extension of AverageTemperature and TotalPrecipitation):

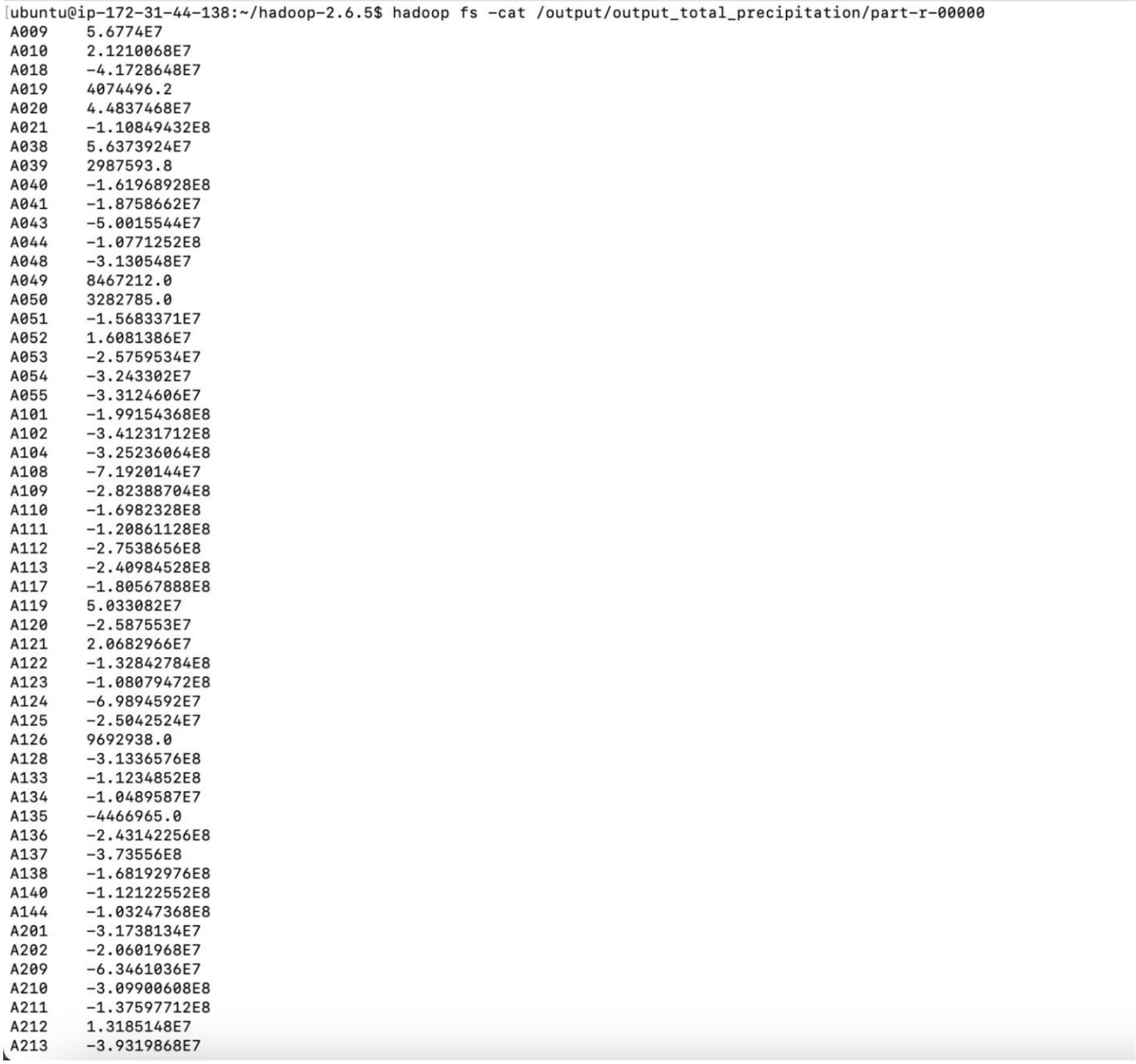
1. **Mapper:**
    - Modify or extend the AverageTemperatureMapper to include:
        - UMIDADE RELATIVA DO AR (relative humidity, column 13).
        - TEMPERATURA DO PONTO DE ORVALHO (dew point temperature, column 9).

Key: Weather station code (column 22)
Values: Humidity and dew point temperature

2. **Reducer:**
    - Modify or extend the AverageTemperatureReducer to calculate:
        - Average humidity.
        - Average dew point temperature.
    - Optionally correlate dew point with temperature using statistical measures (e.g., Pearson correlation).

Outcome:

Provides insights into humidity and dew point trends, highlighting their dependence on temperature.



```
ubuntu@ip-172-31-44-138:~/hadoop-2.6.5$ hadoop fs -cat /output/output_total_precipitation/part-r-00000
A009    5.6774E7
A010    2.1210068E7
A018    -4.1728648E7
A019    4074496.2
A020    4.4837468E7
A021    -1.10849432E8
A038    5.6373924E7
A039    2987593.8
A040    -1.61968928E8
A041    -1.8758662E7
A043    -5.0015544E7
A044    -1.0771252E8
A048    -3.130548E7
A049    8467212.0
A050    3282785.0
A051    -1.5683371E7
A052    1.6081386E7
A053    -2.5759534E7
A054    -3.243302E7
A055    -3.3124606E7
A101    -1.99154368E8
A102    -3.41231712E8
A104    -3.25236064E8
A108    -7.1920144E7
A109    -2.82388704E8
A110    -1.6982328E8
A111    -1.20861128E8
A112    -2.7538656E8
A113    -2.40984528E8
A117    -1.80567888E8
A119    5.033082E7
A120    -2.587553E7
A121    2.0682966E7
A122    -1.32842784E8
A123    -1.08079472E8
A124    -6.9894592E7
A125    -2.5042524E7
A126    9692938.0
A128    -3.1336576E8
A133    -1.1234852E8
A134    -1.0489587E7
A135    -4466965.0
A136    -2.43142256E8
A137    -3.73556E8
A138    -1.68192976E8
A140    -1.12122552E8
A144    -1.03247368E8
A201    -3.1738134E7
A202    -2.0601968E7
A209    -6.3461036E7
A210    -3.09900608E8
A211    -1.37597712E8
A212    1.3185148E7
A213    -3.9319868E7
```

3. Wind Patterns

Objective:

Determine dominant wind directions and speeds and analyze their spatial and temporal variations.

Algorithm:

1. **Mapper (MaxWindSpeedMapper):**
    - Reads the dataset line by line.

- Parses the VENTO, VELOCIDADE HORARIA column (hourly wind speed).
- Extracts the station code (column 22) and wind speed (column 19).
- Outputs key-value pairs where the key is the station code, and the value is the wind speed.

Key: Weather station code (column 22)
Value: Wind speed (column 19)

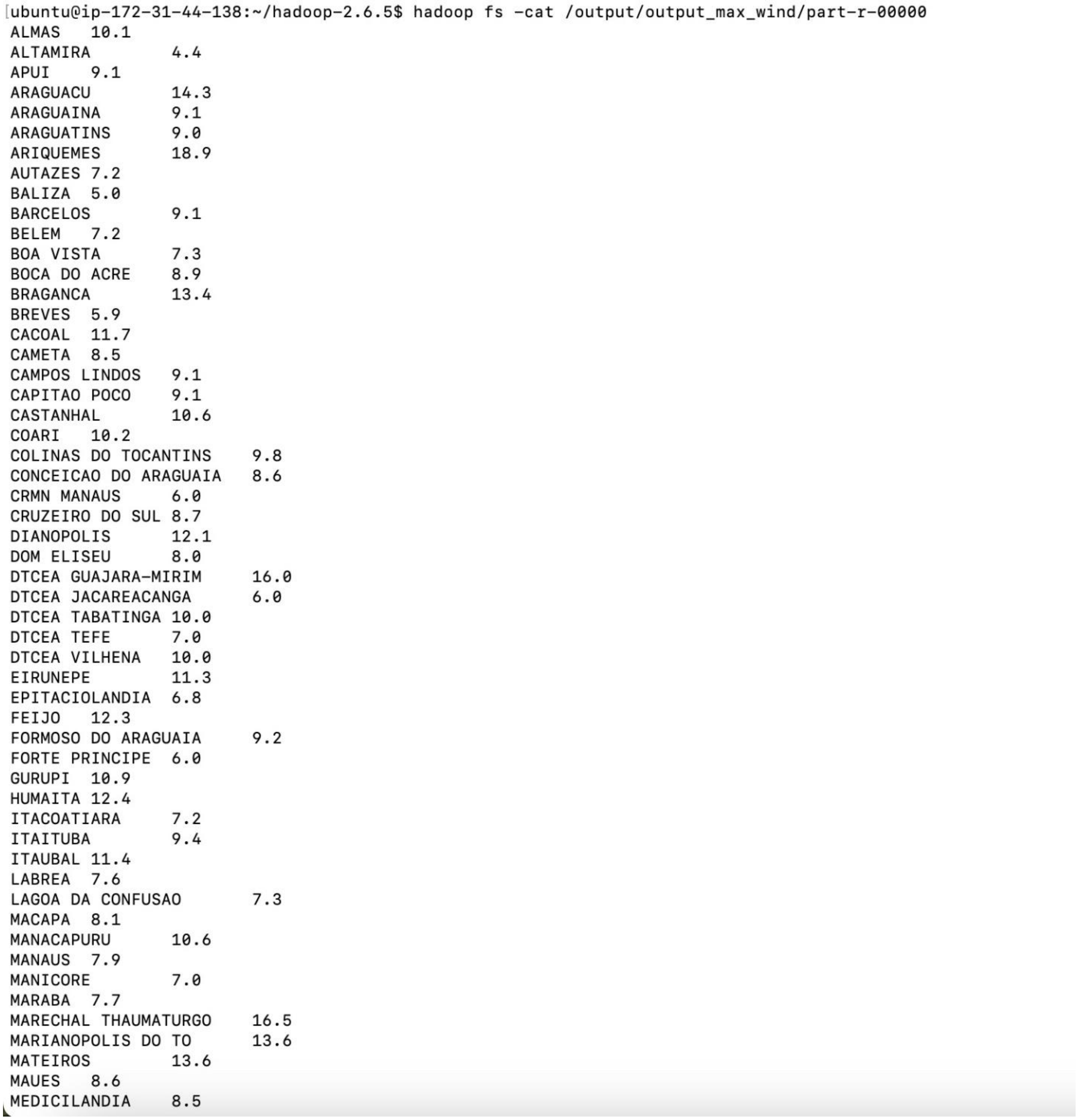
- Skips malformed records or invalid wind speed values.

2. **Reducer (MaxWindSpeedReducer):**
   - Receives a list of wind speed values for each station.
   - Finds the maximum wind speed by comparing all values.
   - Outputs the maximum wind speed for each weather station.

Output Key: Weather station code
Output Value: Maximum wind speed

Outcome:

Identifies the maximum wind speed at each station. Further analysis can aggregate results to determine dominant wind patterns.



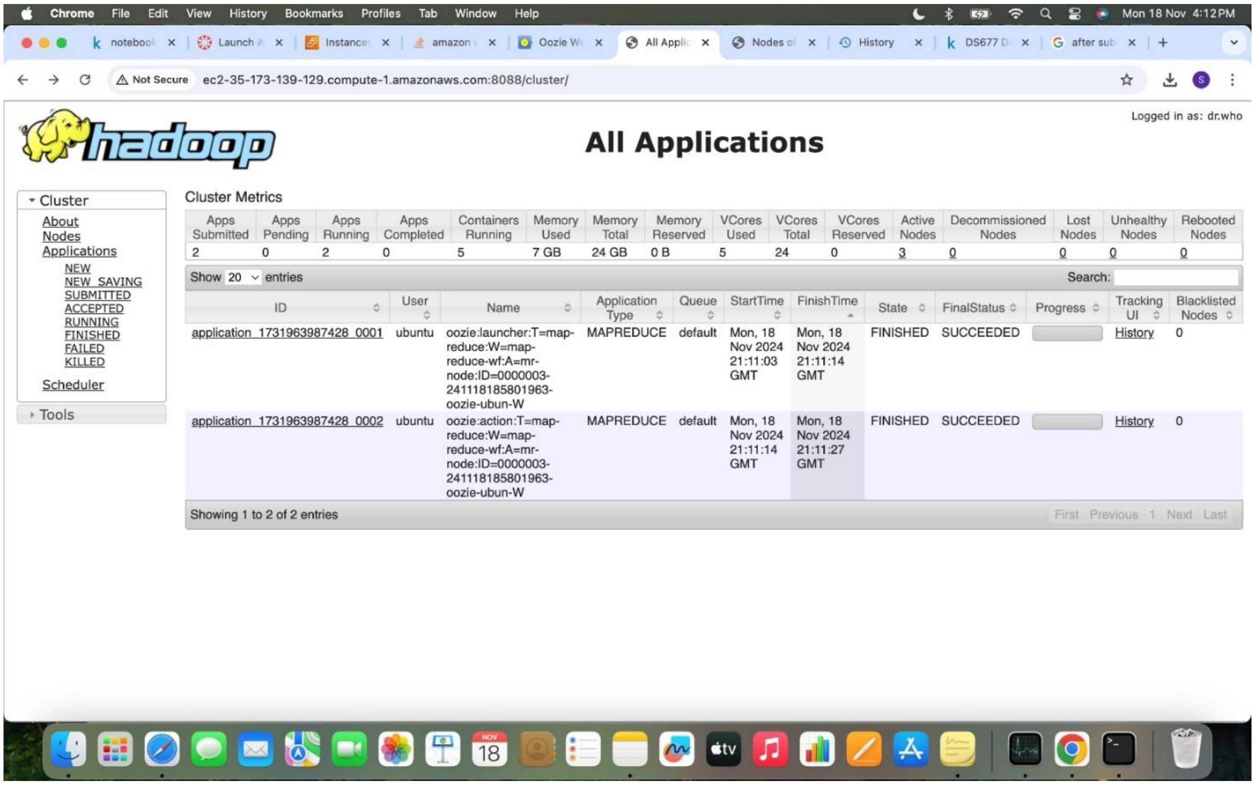
```
ubuntu@ip-172-31-44-138:~/hadoop-2.6.5$ hadoop fs -cat /output/output_max_wind/part-r-00000
ALMAS   10.1
ALTAMIRA        4.4
APUI    9.1
ARAGUACU        14.3
ARAGUAINA       9.1
ARAGUATINS      9.0
ARIQUEMES       18.9
AUTAZES 7.2
BALIZA  5.0
BARCELOS        9.1
BELEM   7.2
BOA VISTA       7.3
BOCA DO ACRE    8.9
BRAGANCA        13.4
BREVES  5.9
CACOAL  11.7
CAMETA  8.5
CAMPOS LINDOS   9.1
CAPITAO POCO    9.1
CASTANHAL       10.6
COARI   10.2
COLINAS DO TOCANTINS    9.8
CONCEICAO DO ARAGUAIA   8.6
CRMN MANAUS     6.0
CRUZEIRO DO SUL 8.7
DIANOPOLIS      12.1
DOM ELISEU      8.0
DTCEA GUAJARA-MIRIM     16.0
DTCEA JACAREACANGA      6.0
DTCEA TABATINGA 10.0
DTCEA TEFE      7.0
DTCEA VILHENA   10.0
EIRUNEPE        11.3
EPITACIOLANDIA  6.8
FEIJO   12.3
FORMOSO DO ARAGUAIA     9.2
FORTE PRINCIPE  6.0
GURUPI  10.9
HUMAITA 12.4
ITACOATIARA     7.2
ITAITUBA        9.4
ITAUBAL 11.4
LABREA  7.6
LAGOA DA CONFUSAO       7.3
MACAPA  8.1
MANACAPURU      10.6
MANAUS  7.9
MANICORE        7.0
MARABA  7.7
MARECHAL THAUMATURGO    16.5
MARIANOPOLIS DO TO      13.6
MATEIROS        13.6
MAUES   8.6
MEDICILANDIA    8.5
```

Oozie Installation and Workflow Execution:-

We successfully installed Oozie and configured it for our project. As part of this process, we created the necessary workflow.xml and job.properties files, which have been included in our project zip file.

Despite numerous attempts, we encountered challenges in successfully running the Oozie jobs for our project. However, as a part of the setup verification, we were able to execute a prebuilt example successfully, demonstrating that the installation was completed correctly.

For reference, a screenshot of the Oozie installation process is also attached.



All Applications

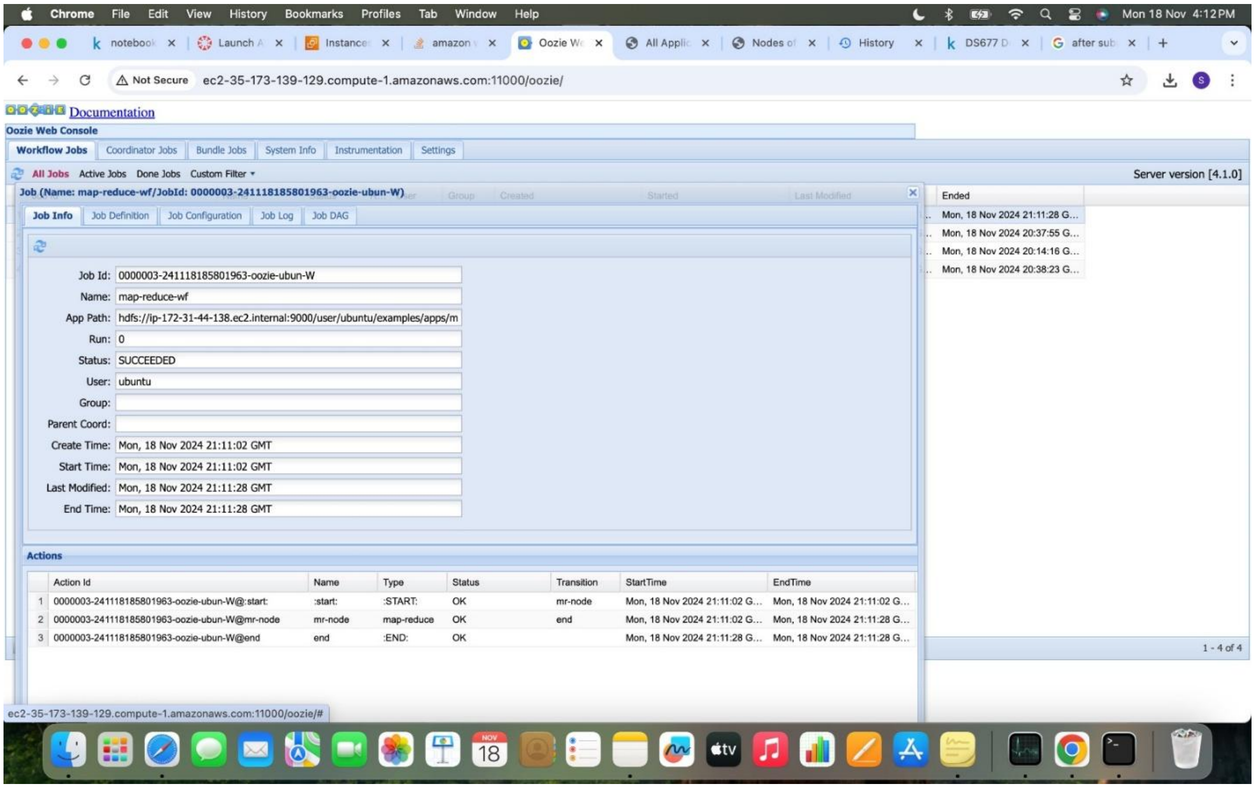
Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
2	0	2	0	5	7 GB	24 GB	0 B	5	24	0	3	0	0	0	0

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes
application_1731963987428_0001	ubuntu	oozie:launcher:T=map-reduce:W=map-reduce-wf:A=mr-node:ID=0000003-241118185801963-oozie-ubun-W	MAPREDUCE	default	Mon, 18 Nov 2024 21:11:03 GMT	Mon, 18 Nov 2024 21:11:14 GMT	FINISHED	SUCCEEDED		History	0
application_1731963987428_0002	ubuntu	oozie:action:T=map-reduce:W=map-reduce-wf:A=mr-node:ID=0000003-241118185801963-oozie-ubun-W	MAPREDUCE	default	Mon, 18 Nov 2024 21:11:14 GMT	Mon, 18 Nov 2024 21:11:27 GMT	FINISHED	SUCCEEDED		History	0

Showing 1 to 2 of 2 entries First Previous 1 Next Last



Job (Name: map-reduce-wf/JobId: 0000003-241118185801963-oozie-ubun-W)

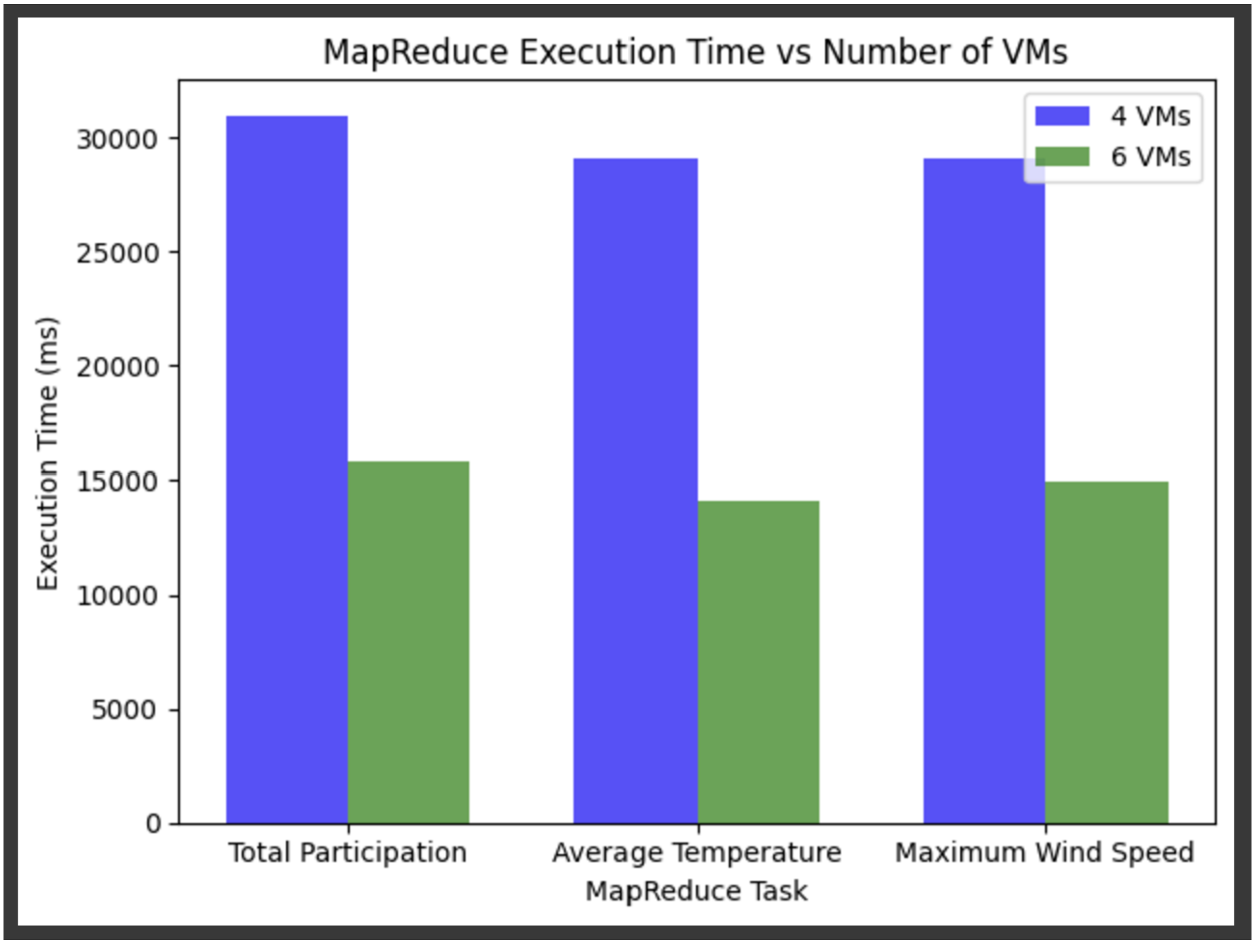
Job Id:	0000003-241118185801963-oozie-ubun-W
Name:	map-reduce-wf
App Path:	hdfs://ip-172-31-44-138.ec2.internal:9000/user/ubuntu/examples/apps/m
Run:	0
Status:	SUCCEEDED
User:	ubuntu
Group:	
Parent Coord:	
Create Time:	Mon, 18 Nov 2024 21:11:02 GMT
Start Time:	Mon, 18 Nov 2024 21:11:02 GMT
Last Modified:	Mon, 18 Nov 2024 21:11:28 GMT
End Time:	Mon, 18 Nov 2024 21:11:28 GMT

Ended
Mon, 18 Nov 2024 21:11:28 G...
Mon, 18 Nov 2024 20:37:55 G...
Mon, 18 Nov 2024 20:14:16 G...
Mon, 18 Nov 2024 20:38:23 G...

Actions

	Action Id	Name	Type	Status	Transition	StartTime	EndTime
1	0000003-241118185801963-oozie-ubun-W@:start:	:start:	:START:	OK	mr-node	Mon, 18 Nov 2024 21:11:02 G...	Mon, 18 Nov 2024 21:11:02 G...
2	0000003-241118185801963-oozie-ubun-W@mr-node	mr-node	map-reduce	OK	end	Mon, 18 Nov 2024 21:11:02 G...	Mon, 18 Nov 2024 21:11:28 G...
3	0000003-241118185801963-oozie-ubun-W@end	end	:END:	OK		Mon, 18 Nov 2024 21:11:28 G...	Mon, 18 Nov 2024 21:11:28 G...

<u>A performance measurement plot that compares the Map Reduce execution time in response to an increasing number of VMs used for processing the entire data set:-</u>

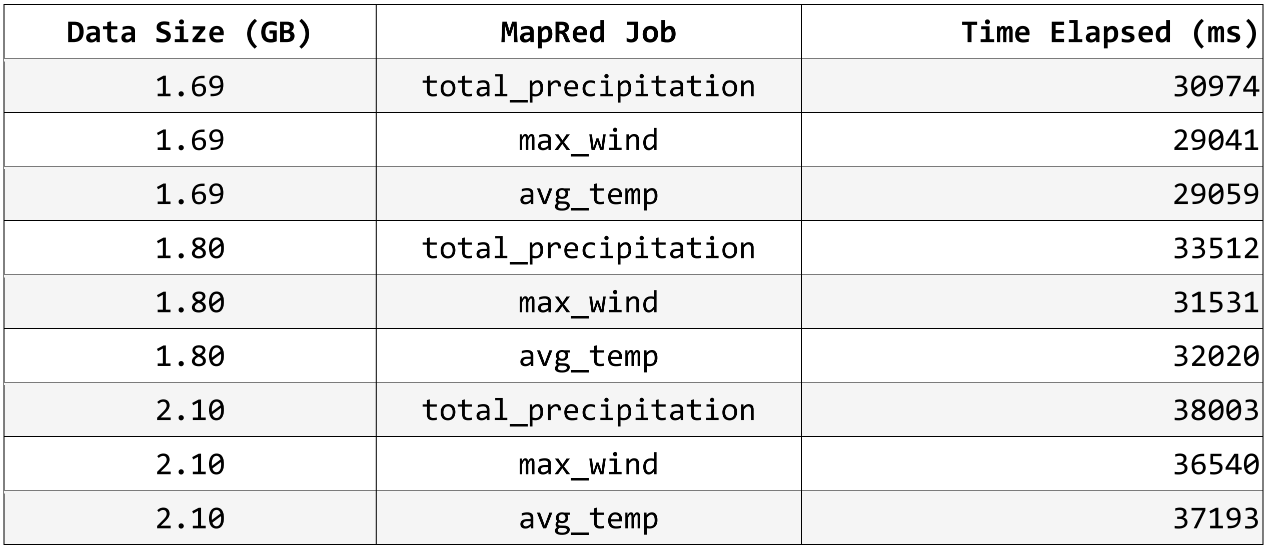


<u>A performance measurement plot that compares the workflow execution time in response to increasing data size and an in-depth discussion on the observed performance comparison results:-</u>

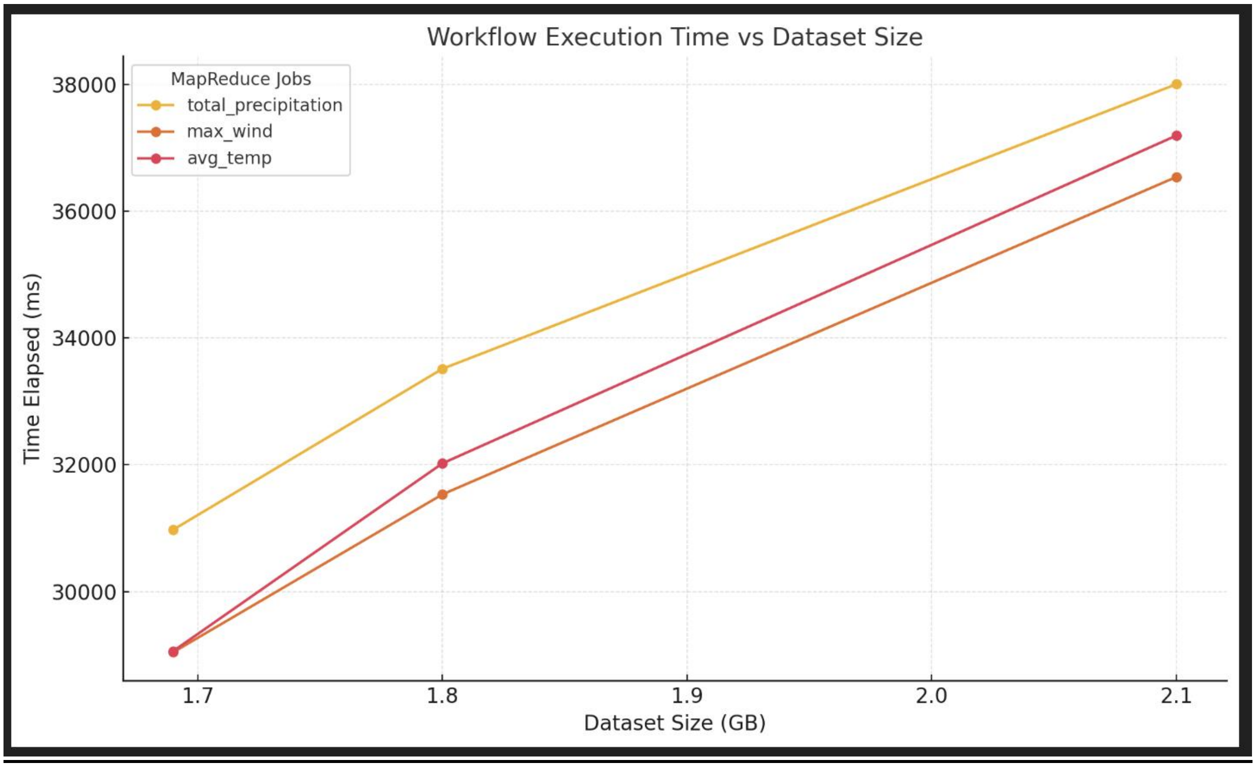
The primary goal of this project is to measure processing times and identify performance trends while performing an incremental increase in dataset size.

Incremental Increases in Dataset Size

After each MapReduce job, we recorded the dataset size and the time it took to perform each job.



Data Size (GB)	MapRed Job	Time Elapsed (ms)
1.69	total_precipitation	30974
1.69	max_wind	29041
1.69	avg_temp	29059
1.80	total_precipitation	33512
1.80	max_wind	31531
1.80	avg_temp	32020
2.10	total_precipitation	38003
2.10	max_wind	36540
2.10	avg_temp	37193



Performance Optimization:-

1. **Job Configuration**:
   - Optimize memory allocation and JVM settings for tasks.
   - Use compression for intermediate outputs to reduce I/O overhead.
2. **Parallel Processing**:
   - Run independent jobs concurrently to utilize resources better.
   - Configure appropriate reducer counts for task distribution.
3. **Resource Scaling**:
   - Add more VMs or increase node capacity to improve job execution.
   - Leverage dynamic resource allocation for priority workflows.
4. **Data Handling**:
   - Use efficient file formats (e.g., Parquet, ORC).
   - Partition datasets for parallel processing and balance input splits.
5. **Error Management**:
   - Implement retries and errors handling mechanisms.
   - Centralize log collection for easier debugging.

These optimizations improve performance, reduce execution time, and ensure scalability for workflows handling large datasets.

Error Handling and Troubleshooting:-

Challenges:

- **Oozie Environment Setup**: Encountered issues with Oozie environment installation due to incorrect commands. *Solution*: Ensured Oozie installs with proper commands, and all required Oozie components were configured correctly.

- **Data Format Issues**: Input data errors caused job failures.
  Solution: Cleaned and validated input data before processing.
- **Resource Limitations**: Job failures due to memory constraints.
  Solution: Adjusted memory settings in the job configuration.

<u>Instruction Document We followed:-</u>

- Kumar Ranjan Oozie Installation
- ChatGPT
- Gemini
- Example of Word Count from Canvas