***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.

1. **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.

A screenshot of a computer

Description automatically generated

**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

1. **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.

A screenshot of a computer

Description automatically generated

**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.

1. **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.

A screenshot of a computer

Description automatically generated

**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.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**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:-**

A graph with green and blue bars

Description automatically generated

**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:-**

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 |

**A graph of a graph

Description automatically generated with medium confidence**

**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.

**Instruction Document We followed:-**

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