WEATHER DATA ANALYSIS

JEEVA JOSE

ARAVIND KRISHNA A V

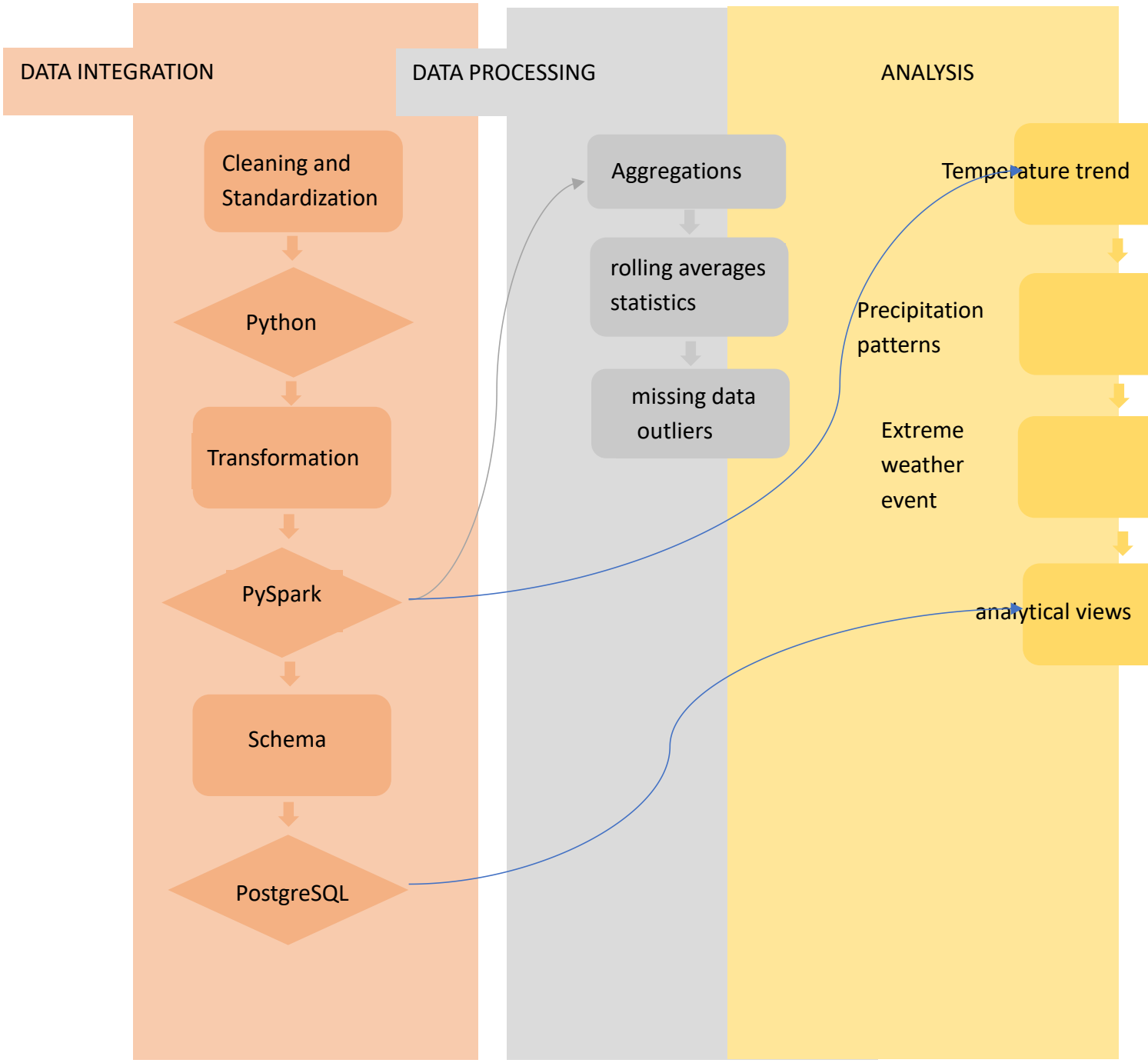
INTRODUCTION

Weather data is a vital resource for understanding the climate and environmental conditions that affect our daily lives. With the increasing availability of large-scale weather datasets, there is a growing need for efficient methods to clean, process, and analyze this data. The Weather Data Analysis project leverages modern data science techniques to unlock valuable insights from weather data, contributing to various fields such as meteorology, agriculture, urban planning, and disaster management.This document provides a detailed guide on how to integrate, process, and analyze weather data to derive meaningful insights into temperature trends, precipitation patterns, and extreme weather events. This project leverages various tools and techniques, including Python, pandas, PySpark, and PostgreSQL, to ensure efficient data handling and analysis.

OBJECTIVES

1. **Integrate and Standardize Data**: Combine weather data from various sources and ensure consistency.
2. **Large-Scale Data Processing**: Utilize PySpark for efficient data transformation and PostgreSQL for storage.
3. **Aggregate and Analyze Data**: Calculate daily/monthly patterns, rolling averages, and handle outliers.
4. **Trend and Pattern Analysis**: Identify temperature trends, precipitation patterns, and seasonal variations.
5. **Detect Extreme Weather Events**: Identify and analyze the frequency and impact of extreme weather events.
6. **Visualize and Report**: Create graphs, charts, dashboards, and analytical views in PostgreSQL.
7. **Document Methodologies**: Provide comprehensive documentation for data integration, processing, and analysis.

TECHNICAL ARCHITECTURE



DATA INTEGRATION

DATA SOURCES

The Global Weather Repository is a comprehensive dataset that provides daily updated weather information from various locations worldwide. It includes data on temperature, precipitation, humidity, wind speed, and other relevant weather parameters.

Data Cleaning

The data cleaning process involves handling missing values, checking for duplicates, and standardizing column names to ensure the dataset is accurate and consistent. Below are the steps taken to clean the weather data:

1. **Handling Missing Values**:
    - The dataset is checked for missing values in each column to identify any gaps in the data.
    - Used pandas methods for this..
2. **Checking for Duplicates**:
    - The dataset is examined for duplicate rows to maintain data integrity.
    - Any duplicate records are identified and removed to ensure uniqueness.
3. **Cleaning and Standardizing Column Names**:
    - Column names are cleaned and standardized by stripping whitespace, converting to lowercase, and replacing spaces with underscores.
    - This ensures consistency and readability in the dataset.

4. **Converting Date Columns**:
    - Date columns, such as 'last_updated', are converted to the appropriate datetime format.
    - This standardization allows for accurate time-based analysis and calculations.

Transforming Data using PySpark for Large-Scale Processing

To handle large-scale weather data efficiently, we use PySpark for data transformation. Below are the steps involved in transforming the data:

1. **Extract Date Information**:
    - Create a new column date_only by extracting the date from the last_updated column.
    - Extract the year, month, and day from the date_only column and store them in separate columns.
2. **Transformations**:
    - Utilize PySpark's withColumn method to perform the transformations.

DATA STORAGE

This PostgreSQL schema efficiently stores and analyzes weather data through seven interconnected tables. **locations** holds geographic details, while **weather_observations** records real-time conditions. Additional data like **weather_conditions** (sun, moon, and condition text) and **air_quality** (pollution levels) enhance insights. Aggregated data is stored in **daily_weather** and **monthly_weather** for trend analysis. **extreme_weather_events** logs significant weather occurrences

DATA PROCESSING

Data Aggregation

The data aggregation process involves grouping weather data by specific time periods and calculating statistical measures.

Daily Aggregation:

- Group data by year, month, and day.
- Calculate average temperature, total precipitation, average humidity, and average pressure for each day.

Monthly Aggregation:

- Group data by year and month.
- Calculate average temperature, total precipitation, average humidity, and average pressure for each month

Rolling Averages and Sums:

- **Temperature**: Calculate a 3-day rolling average for temperature in Fahrenheit.
- **Precipitation**: Calculate a 3-day rolling average and rolling sum for precipitation in inches.
- **Humidity**: Calculate a 3-day rolling average for humidity in percentage.

Statistical Measures:

- Calculate the following statistics for temperature, precipitation, and humidity:
    - Mean, Sum, Count, Minimum, and Maximum values.

Handling Outliers

The process of detecting and handling outliers is crucial for maintaining data quality and accuracy. Below are the key steps taken to detect and replace outliers in the weather data:

1. **Detecting Outliers Using Z-Score**:
    - Calculate the mean and standard deviation of the target column.
    - Compute the Z-score for each data point in the column.
    - Identify outliers as data points with a Z-score greater than 3 (indicating they are more than three standard deviations away from the mean).
2. **Replacing Outliers with Median Value**:
    - Calculate the median value of the target column.
    - Replace the outliers with the median value to minimize their impact on the data.
3. **Outlier Detection and Replacement**:
    - Apply the above steps to detect and replace outliers in the temperature_fahrenheit and precip_in columns.
    - Ensure the cleaned data is saved for further analysis.

VISUALIZATION

Yearly and Monthly Analysis

Yearly Average Temperature and Precipitation:

- Grouped data by year to calculate and visualize the average temperature and precipitation trends.

Monthly Average Temperature and Precipitation:

- Grouped data by year and month to calculate and visualize the average temperature and precipitation trends

Extreme Weather Events Report

Detection

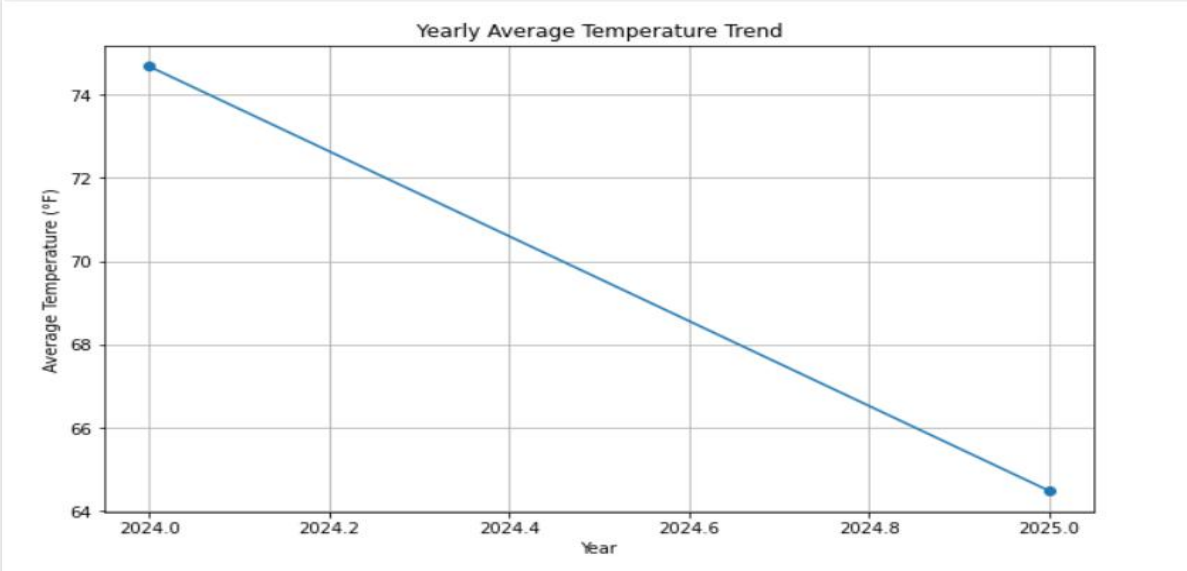
1. **Data Preparation**: Ensure date format.
2. **Thresholds**: Temperature > 100°F, Precipitation > 1 inch.
3. **Identify Events**: Detect and combine extreme temperature and precipitation events. Add event type.

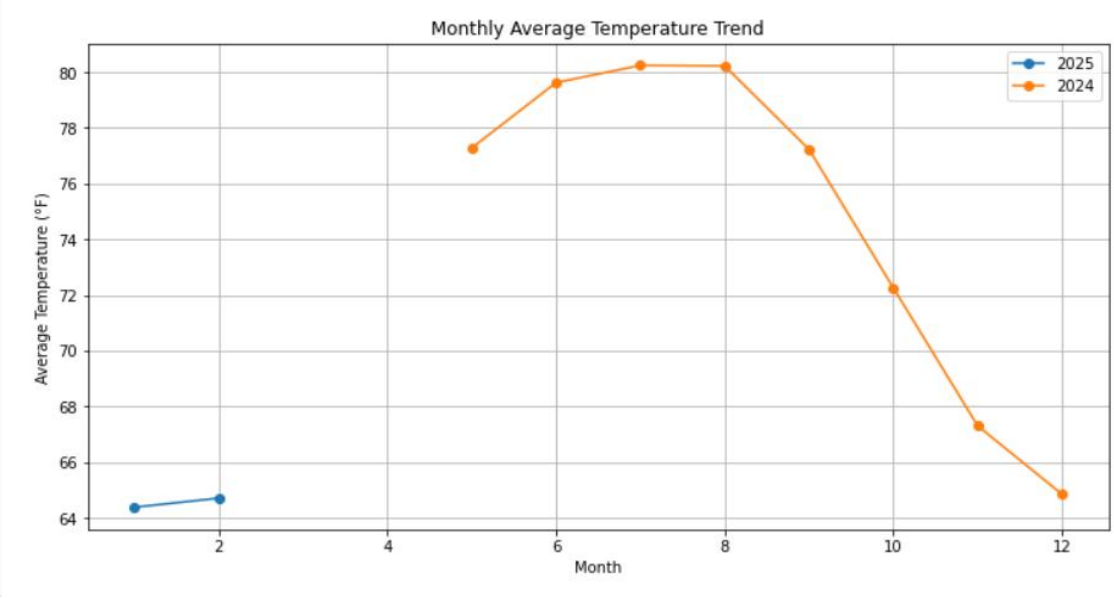
Analytical views

These PostgreSQL views provide structured insights into weather patterns. **avg_monthly_weather** aggregates monthly weather data, showing average temperature, precipitation, humidity, and wind speed for each location. **temperature_trends** calculates a **7-day rolling average temperature**, helping analyze temperature fluctuations over time. Both views enhance weather analysis by simplifying complex queries and improving data accessibility

RESULTS

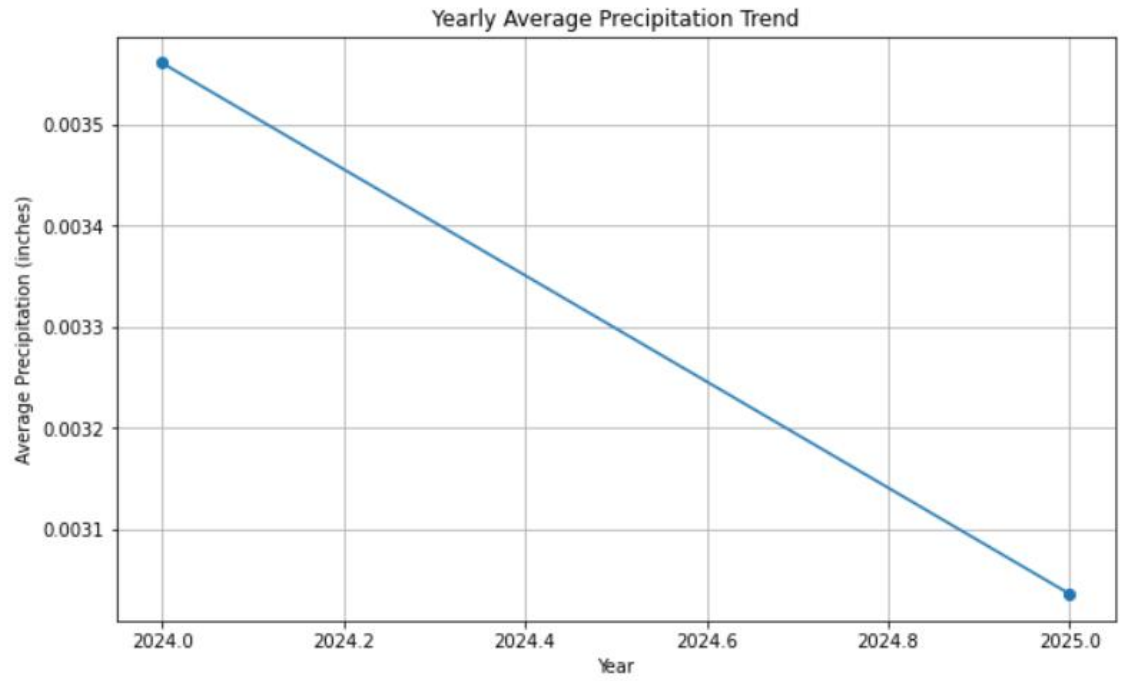
Temperature trend analysis

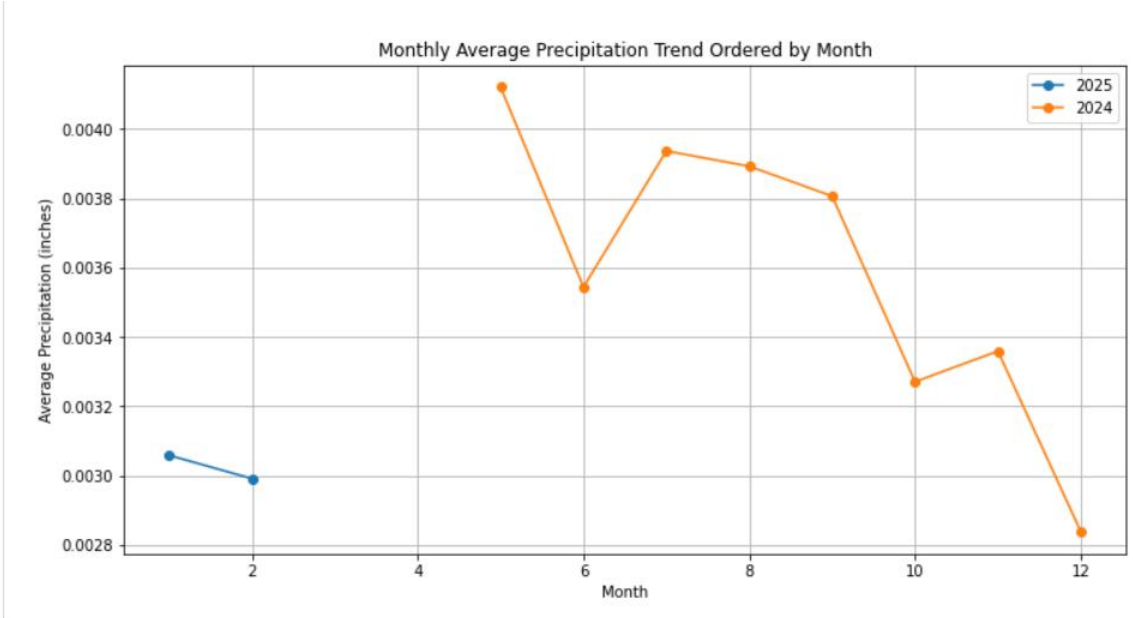




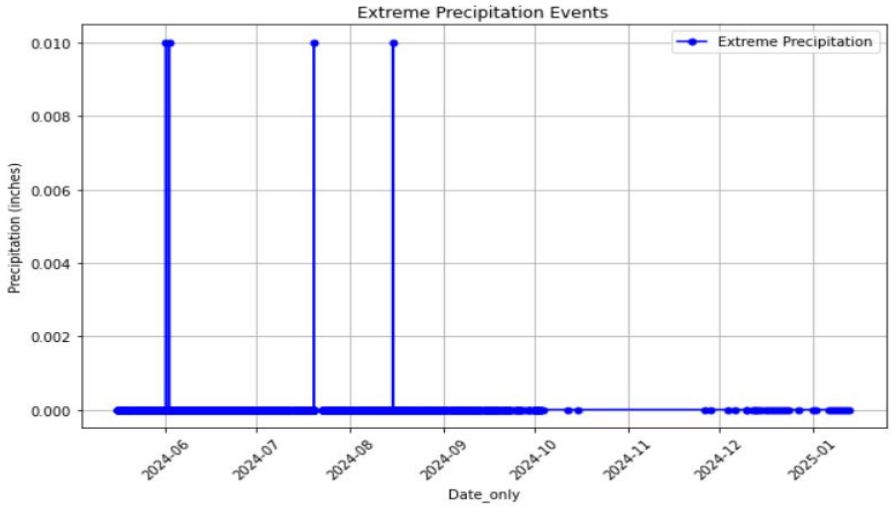
WEATHER DATA ANALYSIS

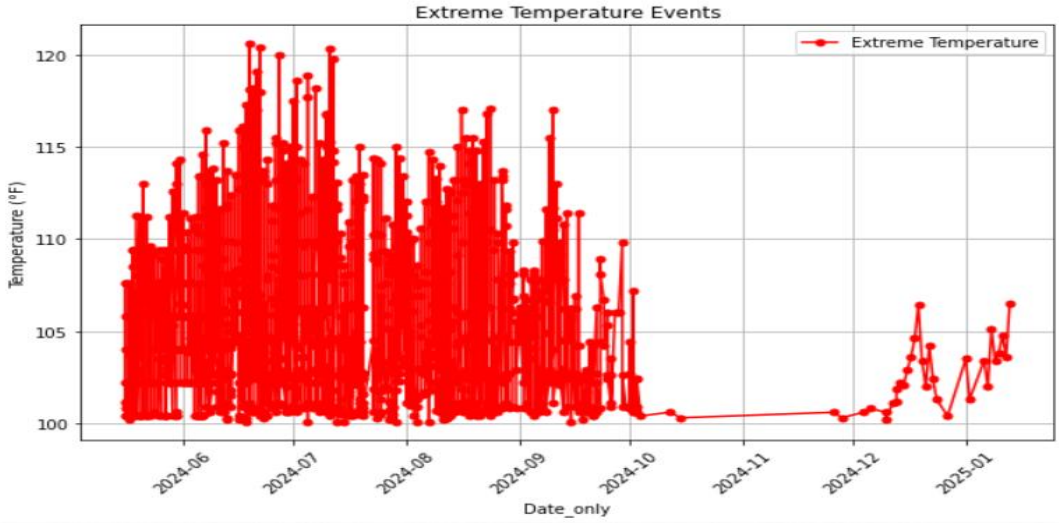
Precipitation patterns





Extreme weather event detection





CONCLUSION

The Weather Data Analysis project has successfully implemented a comprehensive data pipeline to collect, clean, standardize, and analyze weather data. Key achievements include:

Data Integration and Cleaning: Weather data was efficiently integrated from multiple sources, cleaned, and standardized to ensure consistency and accuracy.

Data Processing and Aggregation: The data was aggregated on a daily and monthly basis, enabling the calculation of relevant statistical measures such as rolling averages and sums.

Outlier Detection and Handling: Outliers were detected and replaced with median values to maintain data integrity.

Trend and Pattern Analysis: Temperature and precipitation trends were analyzed on both yearly and monthly scales, providing valuable insights into weather patterns.

Extreme Weather Events Detection: Extreme temperature and precipitation events were identified and visualized, highlighting patterns and potential impacts.

-Visualization and Reporting: Clear and informative visualizations were created to effectively communicate findings and insights.

FUTURE SCOPE

The project scope includes collecting, cleaning, and standardizing weather data, aggregating data, calculating statistics, handling outliers, analyzing trends and extreme weather events, designing PostgreSQL schemas, and creating visualizations with documentation. Real-time weather forecasting, detailed climate change analysis, and developing new data collection tools are out of scope, ensuring focused and efficient execution.