Spark-based Climate Change Analysis

Team: Madhav Reddy Betha, Karthik Pudi, Pooja Bavisetti,

March 12, 2024

Abstract

This project aims to analyze climate data from multiple stations using various meteorological parameters[1]. The dataset includes information such as temperature, weather conditions, precipitation, wind patterns, pressure, and snowfall. The analysis will focus on exploring the distribution of temperatures, calculating average temperatures for different time periods, and investigating trends in temperature anomalies. Additionally, it will analyze the frequency of different weather conditions and the patterns of precipitation to understand their variations by location or season, using features such as hourly and daily temperature, humidity, wind speed and direction, pressure, precipitation, sky conditions, and visibility. The project will analyze wind patterns' relationship with temperature, precipitation, and pressure anomalies. The dataset includes hourly and daily weather features such as sky conditions, visibility, temperature, humidity, wind speed and direction, pressure, and precipitation. Furthermore, the impact of elevation on temperature, precipitation, and other weather variables will be investigated. The study will provide insights into weather patterns and conditions at a single station, aiding in understanding local climate dynamics and trends over time.

0.1 Introduction

Climate change is one of the most pressing challenges facing our planet today, with far-reaching implications for ecosystems, economies, and human societies. Understanding the complex dynamics of climate is crucial for effective mitigation and adaptation strategies[2]. This project aims to analyze a comprehensive dataset encompassing various meteorological parameters across multiple stations to gain insights into climate patterns and trends.

The dataset includes a wide range of variables such as temperature, weather conditions, precipitation, wind patterns, pressure, and snowfall, collected over time[3]. By exploring the distribution of temperatures, calculating average temperatures for different time periods, and investigating trends in temperature anomalies, we can uncover patterns that may reveal the impact of climate change.

Additionally, analyzing the frequency of different weather conditions and the patterns of precipitation will provide valuable information on how these factors vary by location or season. Understanding wind patterns and their correlation with temperature and precipitation can further enhance our understanding of local climate dynamics.

Moreover, the study will delve into pressure data to identify trends or anomalies, shedding light on atmospheric conditions. The impact of elevation on temperature, precipitation, and other weather variables will also be investigated, providing insights into how topography influences local climate.

By analyzing the weather patterns and conditions at a single station, this project aims to understand the local climate dynamics. This analysis will contribute to a deeper understanding of climate patterns at the specific location, aiding in the development of informed strategies to address the challenges posed by climate change.

0.2 Architecture

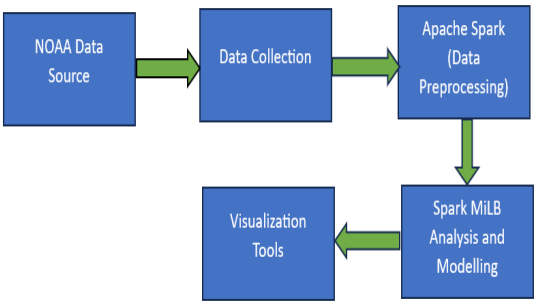

Figure 1: architecture of the weather analysis

The architecture of the project consists of the following components[4]:

Data Collection

Historical climate data will be obtained from reputable sources such as NOAA.[5] The data collection process involves accessing and retrieving the historical climate data from NOAA's databases or archives.

Data Preprocessing

Apache Spark[6] will be utilized for data preprocessing tasks, including cleaning, transforming, and preparing the data for analysis. This step ensures that the data is in a suitable format for further processing.

Analysis and Modeling

Spark's machine[7] learning libraries[8] will be leveraged for analyzing trends, patterns, and anomalies in the climate data. Various statistical and machine learning techniques will be applied to extract meaningful insights from the data.

Reporting and Visualization

The results of the analysis will be visualized using tools like Matplotlib[9] and Plotly[10] to provide stakeholders with a clear and intuitive understanding of the findings. Additionally, a detailed report summarizing the key findings and insights will be generated.

0.3 Initial Findings for the State of the Art

The analysis of climate data has yielded insightful findings regarding both temperature distribution and its relationship with relative humidity.

The histogram analysis of hourly dry bulb temperatures indicates a common range between 60 and 70 degrees Fahrenheit, with a skewed distribution towards cooler temperatures[11]. Additionally, a small peak around 90 degrees Fahrenheit suggests occasional periods of significantly warmer temperatures, potentially indicating anomalies in temperature patterns.

Regarding the relationship between temperature and relative humidity, there is a weak negative correlation, implying that as temperature increases, relative humidity tends to decrease, and vice versa. However, this correlation is not strong, indicating exceptions to this trend. The scatter plot of temperature against relative humidity shows considerable variability, suggesting that factors other than temperature also influence relative humidity.

These initial findings highlight the complexity of climate data analysis and emphasize the importance of further exploration to understand the underlying factors driving temperature and humidity patterns.

0.4 Methodology

Data pre-processing will be performed to clean and format the data for analysis. Spark will be used for data processing, and statistical analysis and machine learning techniques will be applied to the data[12].

0.5 Planned Achievements and Expected Results

The primary goal of this project is to develop a robust and scalable framework for analyzing climate change data using Apache Spark. The expected results include:

- Identification of trends and patterns in climate data.
- Development of predictive models for future climate scenarios.
- Visualization of the analysis results for easy interpretation and decision-making.

By achieving these goals, this project aims to contribute to the understanding of climate change and provide valuable insights for decision-makers and researchers in the field.

0.6 Plan of Work and Timeline

The work will be split into the following phases:

Data Collection and Preprocessing (2 weeks)

Exploratory Data Analysis (EDA) (2 weeks)

Machine Learning Modeling (4 weeks)

Reporting and Visualization (2 weeks)

Finalizing the Project and Documentation (1 week)

This timeline provides a structured approach to completing the project within a reasonable timeframe, allowing for each phase to be completed thoroughly and efficiently.

0.7 Conclusion

This project aims to leverage the capabilities of Apache Spark to gain insights into climate change patterns and trends. By developing a scalable and efficient framework [13], this project aims to contribute to the ongoing efforts in understanding and mitigating the impacts of climate change.

0.8 GitHub Account

GitHub Repository Link: `https://github.com/poojab2813/603-project`

0.9 References

Data Sources

National Oceanic and Atmospheric Administration (NOAA)

Apache Spark

Apache Spark website

Data Preprocessing

Foster, Ian, et al. "Big Data: Principles and best practices of scalable computing." (2017)

Machine Learning Libraries for Spark

Spark MLlib documentation

Data Visualization Libraries

Matplotlib website
Plotly website

Additional References

Intergovernmental Panel on Climate Change (IPCC)
Rahmstorf, Stefan, and Dim Coumou. "Increase of risk of severe convective precipitation events in Europe with global warming." Environmental Research Letters 6.4 (2011)

Bibliography

[1] J. Zhang and H. Wang, "Analysis of clarans algorithm for weather data based on spark," in *International Symposium on Biometrics and Security Technologies (ISBAST)*. IEEE, 2019, p. 15.

[2] A. G. BySarah E. Cornell, "Is climate change the most important challenge of our times?" in *in 2019 International Symposium on Biometrics and Security Technologies (ISBAST)*. IEEE, 2019, p. 15.

[3] W. P. c. A. Y. H. d. E. G.-O. e. L. A. M. f. C. F. A. g. P. S. h. C. K. i. d. G. J. H. j. d. M. d. C. a. Francisco J. Tapiador a, F.J. Turk b, "Global precipitation measurement: Methods, datasets and applications," in *in IEEE Access*. IEEE, 2018, pp. 70–97.

[4] L. Y. b. D. H. L. Joseph C. Lam a, C.L. Tsang a, "Weather data analysis and design implications for different climatic zones in china," in *in IEEE Access*. IEEE, 2005, pp. 70–97.

[5] "National oceanic and atmospheric administration (noaa)," https://www.ncei.noaa.gov/pub/data/cdo/samples/LCD_sample_csv.csv.

[6] "Apache spark website," https://spark.apache.org/.

[7] M. A. E. B. G. L. A. P. Tafti, "Big data machine learning using apache spark mllib." IEEE, 2017.

[8] "Spark mllib documentation," https://spark.apache.org/mllib/.

[9] "Matplotlib website," https://matplotlib.org/.

[10] "Plotly website," https://plotly.com/python/.

[11] R. Paull, "Effect of temperature and relative humidity on fresh commodity quality." IEEE, 1999.

[12] M. Bowles, "Machine learning with spark and python." IEEE, 1999.

[13] D. Z. . M. B. Khadija Aziz, "Leveraging resource management for efficient performance of apache spark." IEEE, 2019.