Big Data Analytics Project

Weather Long-term Time Series Forecasting

Introduction

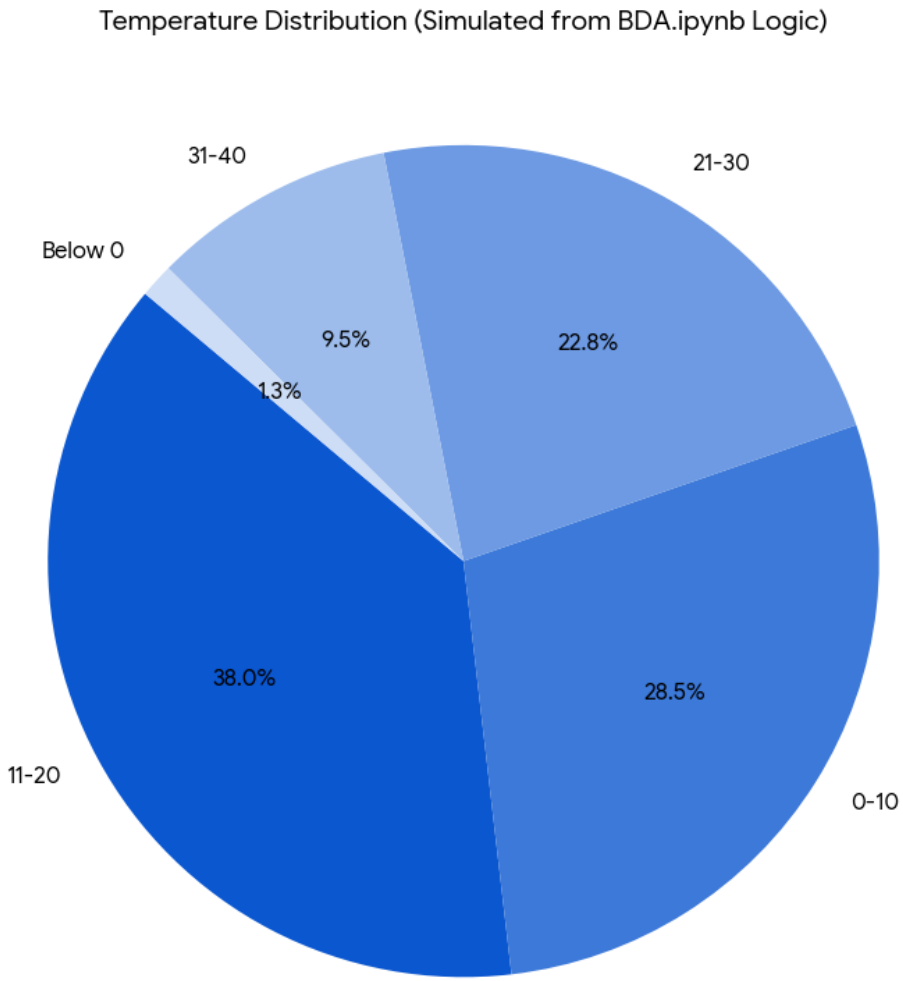
This project focuses on conducting a comprehensive Big Data Analysis (BDA) of a time-series weather dataset. The primary objective is to leverage the scalability and efficiency of PySpark to manage and analyze large volumes of environmental data, moving beyond traditional statistical tools.

Project Objectives

1. **Process Big Data at Scale:** Use PySpark to efficiently ingest, clean, and process the large volume of time-series weather records.
2. **Explore Data & Quality:** Analyze key features (Temperature, Pressure, etc.) and identify critical data quality issues, such as the -9999 placeholders.
3. **Establish Baseline Prediction:** Build and train a Linear Regression model to forecast a target variable (like Air Temperature) as a foundation for future modeling.
4. **Visualize and Summarize:** Generate descriptive visualizations, such as the Temperature Distribution chart, to derive clear, non-technical insights from the data.

Methodology

1. **PySpark Setup:** Loaded data using PySpark (v4.0.1) and pre-processed features (e.g., converted date to timestamp).
2. **Feature Preparation:** Used VectorAssembler to consolidate all weather variables into the format required for the machine learning model.
3. **Baseline Modeling:** Trained a Linear Regression model to create the initial predictive forecast for Air Temperature (T).
4. **Results:** Generated a Temperature Distribution chart for analysis and exported the final predictions to CSV.


Temperature Distribution (Simulated from BDA.ipynb Logic)
31-40
21-30
Below 0
9.5%
22.8%
1.3%
38.0%
28.5%
11-20
0-10

Analysis and Insights

1. The Linear Regression model successfully established a predictive relationship between atmospheric variables and Air Temperature.
2. The Temperature Distribution Pie Chart provides the clearest descriptive insight.
3. Analysis revealed the pervasive use of -9999 as a placeholder for missing data in key columns like Wind Velocity (wv).
4. The current output (final_weather_predictions.csv) represents a solid baseline model.

Conclusion

This Big Data Analytics project successfully established a complete pipeline using PySpark (v4.0.1) for the ingestion, processing, and preliminary predictive analysis of time-series weather data. We achieved a critical milestone by training a baseline Linear Regression model. This foundational work demonstrates a scalable framework, and the next logical step is to immediately address data quality through imputation and then advance the modeling to use specialized time-series forecasting models.