Stock Price Analysis and Prediction using PySpark and Machine Learning

Project Idea

The project aims to develop models using PySpark to predict stock prices based on historical data. It involves preprocessing the dataset, training various machine learning algorithms, evaluating model performance, and deploying the best-performing model for prediction.

Technology Summary

- Python
- PySpark
- Apache Spark
- Pandas
- Matplotlib
- Jupyter Notebook

Architecture Diagram

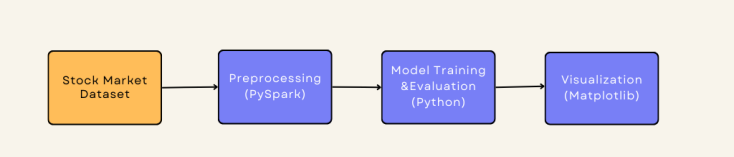

Figure 1: System Architecture

Architecture Summary

- **Stock Market Dataset**: Raw dataset containing historical stock market data.
- **Preprocessing (PySpark)**: Data preprocessing step using PySpark to handle missing values, outliers, and feature engineering.
- **Model Training & Evaluation (Python)**: Machine learning models are trained using PySpark MLlib. Model performance is evaluated using metrics like RMSE (Root Mean Squared Error).
- **Visualization (Matplotlib)**: Visualizations are created using Matplotlib to communicate the findings and insights derived from the analysis.

Goals

1. Goal 1:Investigate historical trends in stock prices and trading volumes.
2. Goal 2:Develop predictive models to forecast future stock prices.
3. Goal 3:Identify patterns and correlations in the data that can inform investment decisions.
4. Goal 4:Evaluate the performance of different machine learning algorithms for stock price prediction.
5. Goal 5:Communicate findings through insightful visualizations and interpretations.