Hemanth Kumar K(A20516013)

Project Draft: PySpark Analytical Framework: Integrating Model Deployment for Real-Time Big Data Insights"

Abstract:

This project utilizes PySpark, a powerful big data processing framework, to analyze historical stock data, develop a predictive model using linear regression, and deploy the model for real-time stock price predictions. The project aims to showcase the capabilities of PySpark in handling large datasets, performing machine learning tasks, and deploying models for practical use.

Introduction:

The project introduces the use of PySpark in analyzing historical stock data. The primary goals are to develop a robust predictive model using linear regression and deploy the model for real-time stock price predictions. By utilizing PySpark's capabilities, the project demonstrates efficient handling of large datasets, machine learning tasks, and practical deployment scenarios.

Setup and Installation:

- Installation of PySpark
- Configuration of SparkSession
- Loading and exploring the dataset

Data Analysis and Feature Engineering:

- Data cleaning and preprocessing
- Feature development and selection
- Exploratory data analysis (EDA)

Model Development:

- Linear regression model
- Training and testing the model
- Model evaluation using RegressionEvaluator

Big Data Implementation:

- Leveraging PySpark for scalable data processing
- Handling large volumes of stock data efficiently
- Demonstrating the advantages of a big data framework

Real-Time Deployment:

- Converting the model for real-time predictions
- Incorporating the model into a PySpark pipeline
- Displaying real-time predictions

Model Deployment and Real-Time Usage:

- Steps for deploying the model
- Real-time usage scenarios and benefits

Conclusion:

- Summary of achievements
- Reflection on the project's success and challenges
- Future enhancements and possibilities