Hemanth Kumar K(A20516013)

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

Introduction:

In the era of big data, the need for robust and scalable data processing tools is paramount. This project aims to elevate the analytical capabilities of PySpark, a global data processing framework. Leveraging PySpark's strengths, we will focus on feature development, model deployment, and real-time usage to create a comprehensive solution for large-scale data analytics.

Objectives

Feature Development:

- Enhance documentation, performance, usability, and functionality of PySpark.
- Implement new features to facilitate data cleansing, loading, and scalable data manipulation.

Model Deployment:

- Develop a mechanism for deploying machine learning models using PySpark.
- Explore integration with external tools or platforms for efficient model deployment.

Real-Time Usage:

- Implement real-time data processing capabilities to enhance PySpark's versatility.

- Demonstrate the effectiveness of PySpark in handling dynamic data streams.

Methodology:

The project will follow a systematic approach:

- Conduct a thorough analysis of PySpark's existing features and limitations.
- Implement new features and enhancements based on identified areas for improvement.
- Explore tools and technologies for model deployment, ensuring seamless integration.
- Develop and test real-time data processing capabilities to align PySpark with evolving data needs.

Expected Outcomes

- Improved PySpark Functionality:
- Enhanced documentation, performance, and usability.
- Additional features for data manipulation and analytics.

Model Deployment Mechanism**:**

- A robust system for deploying machine learning models using PySpark.

Real-Time Data Processing:

- Implementation of real-time data processing capabilities.

References:

- Azhari, M., Abarda, A., Ettaki, B., Zerouaoui, J. and Dakkon, M., 2020. Higgs boson discovery using machine learning methods with pyspark. *Procedia Computer Science*, *170*, pp.1141-1146.
- Shaikh, E., Mohiuddin, I., Alufaisan, Y. and Nahvi, I., 2019, November. Apache spark: A big data processing engine. In *2019 2nd IEEE Middle East*

and North Africa COMMunications Conference (MENACOMM) (pp. 1-6). IEEE.

- Pradhan, R., Mannepallli, P.K. and Rajpoot, V., 2021, March. Analysing uber trips using pyspark. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1119, No. 1, p. 012013). IOP Publishing.

Conclusion

This project endeavors to elevate PySpark's capabilities to meet the demands of modern big data analytics. By focusing on feature development, model deployment, and real-time usage, we aim to provide users with an enhanced, versatile tool for efficiently processing and analyzing vast volumes of data. This project aligns with the growing needs of the data analytics community and contributes to the continuous evolution of PySpark as a reliable solution for large-scale data processing.