Group: 1. Han Chandeth
2. Li Nita
3. Nuon Roatny

Project Proposal:

Big Data Processing and Analysis Framework for Stock Price Prediction

1. Introduction

This project aims to create a specialized Big Data Processing and Analysis Framework tailored for Stock Price Prediction. By leveraging specific tools and technologies, the framework will encompass data ingestion, storage, and advanced analytics, ultimately predicting stock prices for the next day.

2. Project Scope

Part 1: Big Data Pre-Processing

Objectives:

- Establish a streamlined pipeline for ingesting, processing, and storing stock market data.
- Ensure data quality, consistency, and readiness for predictive analysis.

Tools and Technologies:

- **Python:**
  - Utilized for web scraping from CSX web to gather real-time stock data.
- **Kafka:**
  - Distributed streaming platform for real-time data ingestion.
- **Hadoop:**
  - Hadoop Distributed File System (HDFS) for efficient storage of historical stock data.

Tasks:

- Implement Python scripts for web scraping.
- Set up Kafka for continuous real-time data ingestion.
- Utilize HDFS for storage of historical stock data.

Part 2: Big Data Analysis

Objectives:

- Apply machine learning techniques, specifically SVM algorithm, for stock price prediction.
- Utilize Spark and MLlib for efficient data analytics.

Tools and Technologies:

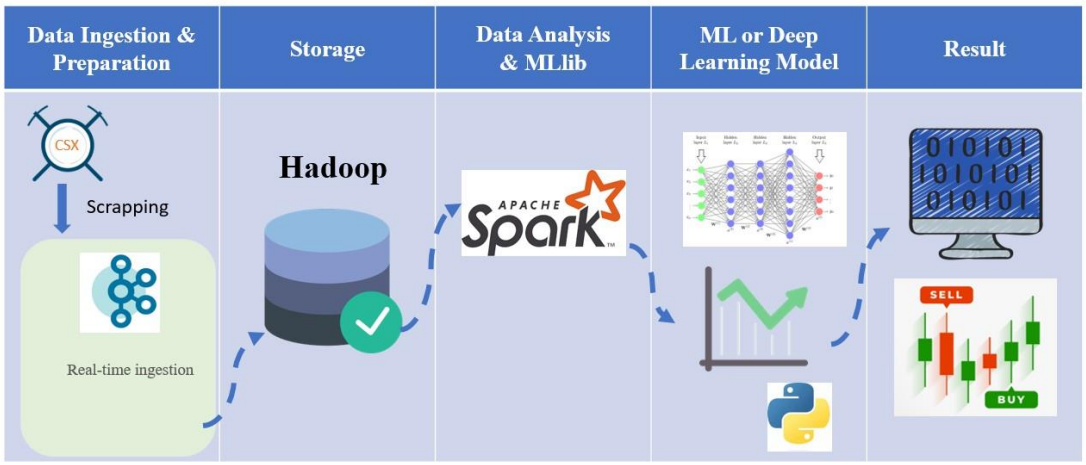
- **Spark:**
  - Core framework for distributed in-memory data processing.
- **MLlib (Machine Learning Library):**
  - Implementation of SVM algorithm for stock price prediction.

Tasks:

- Develop Spark-based analytics pipelines for data processing.
- Apply SVM algorithm from MLlib for stock price prediction.

3. Framework Development

Architecture: Stock Price Prediction with Big Data



Components:

Data Collection:

- Develop Python scripts to scrape stock data from CSX website.
- Implement Kafka for real-time data streaming and ingestion.

Data Storage:

- Utilize Hadoop Distributed File System (HDFS) for efficient storage of historical stock data.

Data Processing and Analytics:

- Leverage Spark for distributed in-memory data processing.
- Design and implement data analytics pipelines for feature extraction and exploration.

Machine Learning:

- Apply SVM algorithm from MLlib for stock price prediction.
- Train the model using historical stock data.

4. Expected Outcomes

- Accurate predictions of stock prices for the next day using the SVM algorithm.
- Efficient handling of real-time stock data through Kafka and seamless processing using Spark.

- Reliable storage and retrieval of historical stock data in Hadoop.

5. Conclusion

This project represents a focused effort to build a specialized Big Data Processing and Analysis Framework tailored for Stock Price Prediction. By combining specific tools and technologies, the framework aims to provide accurate insights for investors and financial analysts, enhancing decision-making processes in the stock market.