Assignment: Triển khai hệ thống phân tích và xử lý dữ liệu lớn

Case study: Stock Price Bigdata

Mục tiêu:

- Tìm hiểu một số chủ đề về phân tích dữ liệu lớn.
- Vận dụng kỹ thuật phân tích và xử lý dữ liệu.
- Xây dựng, triển khai hệ thống dữ liệu lớn với bài toán cụ thể: Stock Price Bigdata
- Kiểm tra kết quả thực hiện trên màn hình giao diện ứng dụng.
- Submit kết quả thực hiện Project lên hệ thống (LMS Canvas, Github, jupyter notebook…)

Nội dung:

1. **Kiểm tra thực hiện việc tìm hiểu bài toán trong Lab trước:**

- **Bài toán:**

  Kịch bản của thị trường chứng khoán không ngừng chuyển động khi hàng ngày có đến hàng triệu lượt giao dịch. Việc áp dụng xử lý dữ liệu lớn vào môi trường chứng khoán trở thành một công cụ hữu hiệu cho các nhà đầu tư. Các phân tích dữ liệu lớn sẽ giúp các nhà đầu tư đưa ra những quyết định thông minh, hạn chế rủi ro trên môi trường đầy biến động này.

  Project hướng đến việc mô phỏng một hệ thống big data sử dụng hadoop và spark cho việc lưu trữ và xử lý dữ liệu chứng khoán.

  # Mô hình hệ thống:

  Dữ liệu được lưu trữ trên một cụm HDFS bao gồm:
  - 1 Namenode để quản lý các datanode
  - 4 Datanode để lưu trữ dữ liệu.

  Để lấy dữ liệu ra và xử lý, nhóm demo sử dụng một cụm spark gồm 1 Spark Master và 4 Spark Worker.

  # Công nghệ sử dụng:
  Hadoop và Spark.
  Dùng nền tảng docker để xây dựng hệ thống mô phỏng. Sử dụng các image của Big Data Europe (bde2020) để tạo ra namenode, các datanode, dịch vụ yarn, sparkmaster và các sparkworker.
  Sử dụng image pyspark-notebook của jupyter để demo việc xử lý dữ liệu của cụm spark (xem file docker-compose.yml)
  Tài liệu đính kèm: Stock Price Bigdata

2. **Triển khai xây dựng hệ thống theo các bước**
3. **Viết report báo cáo kết quả thực hiện**
4. **Submit report và kết quả thực hiện trên hệ thống**