QUY ĐỊNH THỰC HIỆN ĐỒ ÁN

MÔN BIG DATA

1. Mô tả Yêu Cầu

Sinh viên sẽ phát triển một hệ thống xử lý và phân tích dữ liệu lớn (Big Data) sử dụng các công nghệ phân tán như Apache Hadoop, Apache Spark hoặc các hệ thống phân tán khác. Dữ liệu đầu vào có thể đến từ các nguồn như dữ liệu mạng xã hội, giao dịch tài chính, hoặc dữ liệu từ các hệ thống cảm biến. Mục tiêu là xây dựng hệ thống có khả năng xử lý, phân tích và trích xuất thông tin giá trị từ tập dữ liệu lớn.

2. Chi Tiết Yêu Cầu

- Lưu trữ phân tán: Sử dụng hệ thống tệp phân tán (HDFS, Ceph, GlusterFS, v.v.) để lưu trữ và quản lý dữ liệu.
- Xử lý dữ liệu: Sử dụng các công cụ như MapReduce, Apache Spark hoặc Flink để xử lý và phân tích dữ liệu lớn.
- Phân tích dữ liệu: Áp dụng các thuật toán phân tích dữ liệu lớn như clustering, regression, machine learning để rút ra thông tin hữu ích.
- Trực quan hóa kết quả: Sử dụng các công cụ trực quan hóa (Matplotlib, Tableau, v.v.) để trình bày kết quả phân tích dưới dạng biểu đồ, giao diện đồ họa hoặc báo cáo.

3. Thang Điểm - Phần app

- Hệ thống lưu trữ phân tán: 2 điểm
- Xử lý và phân tích dữ liệu: 4 điểm
- Trực quan hóa kết quả: 2 điểm
- Tính sáng tạo và hiệu quả của hệ thống: 2 điểm

4. Hình thức báo cáo

- Thời gian nộp bài: Trước Buổi 9 (Theo dõi trên Classroom)
- Báo cáo: Bằng ppt và Demo. Báo cáo vào buổi 10.
- Sinh viên nộp bài trên Classroom bao gồm: Source Code, File word báo cáo và File ppt báo cáo.

Lưu Ý:

- Bài tập lớn cần được nộp đúng thời hạn và tuân thủ các yêu cầu của giảng viên hướng dẫn.

MỘT SỐ ĐỀ TÀI GỢI Ý

1. Phân tích dữ liệu mua sắm trực tuyến

Phân tích hành vi người dùng từ các trang thương mại điện tử để phát hiện xu hướng mua sắm và đề xuất sản phẩm.

2. Phân tích dữ liệu giao thông đô thị

Sử dụng dữ liệu từ hệ thống cảm biến giao thông để dự đoán tắc nghẽn và tối ưu hóa hệ thống đèn giao thông.

3. Phân tích dữ liệu cảm biến sức khỏe

Phân tích dữ liệu từ thiết bị đeo tay để theo dõi và dự đoán tình trạng sức khỏe.

4. Hệ thống phát hiện gian lận tài chính

Xây dựng hệ thống phát hiện các giao dịch đáng ngờ trong lĩnh vực tài chính dựa trên phân tích dữ liệu lớn.

5. Phân tích dữ liệu truyền thông xã hội

Thu thập và phân tích dữ liệu từ mạng xã hội để xác định xu hướng và dự đoán sự kiện quan trọng.

6. Tối ưu hóa hệ thống quản lý kho hàng

Sử dụng dữ liệu lớn để tối ưu hóa quản lý kho và phân phối sản phẩm.

7. Phân tích dữ liệu giao thông hàng không

Sử dụng dữ liệu giao thông hàng không DataExpo để dự đoán phân tích nguyên nhân delay các chuyến bay và dự đoán sự kiện quan trọng.