ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

Xây dựng hệ thống hồ dữ liệu phân tích dữ liệu chuyến bay sử dụng các thành phần của hệ sinh thái Hadoop

TRẦN NGỌC BẢO
bao.tn215529@sis.hust.edu.vn

Ngành Kỹ thuật máy tính

Giảng viên hướng dẫn: TS. Trần Việt Trung _____

Chữ kí GVHD

Khoa: Khoa học máy tính

Trường: Công nghệ Thông tin và Truyền thông

HÀ NỘI, 12/2024

LỜI CẢM ƠN

Trong những năm tháng thanh xuân rực rỡ, em thật may mắn và tự hào khi được trở thành sinh viên của Đại học Bách khoa Hà Nội. Học tập và rèn luyện trong môi trường chuyên nghiệp, năng động tại đây là một trải nghiệm quý giá, đáng nhớ trong cuộc đời. Bốn năm gắn bó với mái trường Bách khoa là hành trình đầy ắp những thử thách, vấp ngã nhưng cũng tràn đầy niềm tự hào và vinh quang. Trước hết, em xin gửi lời tri ân sâu sắc đến gia đình – chỗ dựa vững chắc và nguồn động lực to lớn giúp em vượt qua mọi khó khăn trên con đường học vấn. Em cũng chân thành cảm ơn TS. Trần Việt Trung, người đã tận tình hướng dẫn, định hướng và đồng hành cùng em trong quá trình thực hiện đồ án này. Đồng thời, em bày tỏ lòng biết ơn đến các thầy cô giảng viên của Đại học Bách khoa Hà Nội, đặc biệt là các thầy cô thuộc trường Công nghệ Thông tin và Truyền thông, những người đã truyền đạt cho em những kiến thức quý báu – hành trang vững chắc cho tương lai. Cuối cùng, em không thể không nhắc đến những người bạn đồng hành tuyệt vời – những người luôn sẵn sàng sẻ chia, hỗ trợ em trong suốt thời gian học tập. Những năm tháng tươi đẹp ấy sẽ mãi là ký ức đáng trân trọng và là nguồn động lực to lớn để em không ngừng phấn đấu trên hành trình phía trước.

TÓM TẮT NỘI DUNG ĐỒ ÁN

Dữ liệu lớn ngày càng được xem như một "tài nguyên" cực kỳ giá trị trong hiện tại và tương lai. Việc lưu trữ và xử lý dữ liệu lớn đóng vai trò thiết yếu đối với các doanh nghiệp, đặc biệt là trong việc quyết định mô hình hệ thống và công nghệ phù hợp để triển khai. Hiện nay, các hệ thống dữ liệu lớn có thể được triển khai theo các mô hình như kho dữ liệu, hồ dữ liệu và hồ kho dữ liệu, cùng với nhiều hệ sinh thái công nghệ hỗ trợ như Hadoop, Amazon Web Services, Google Cloud Platform. Việc lựa chọn mô hình và hệ sinh thái phù hợp để xây dựng hệ thống đòi hỏi phải giải quyết các thách thức về khối lượng lưu trữ và xử lý, chi phí và bảo mật.

Đồ án này được em phát triển và thực hiện với mục tiêu xây dựng một hệ thống hồ dữ liệu phân tích chuyến bay dựa trên các công nghệ của hệ sinh thái Hadoop. Em đã lấy một nguồn dữ liệu về chuyến bay tại Hoa Kỳ trên Kaggle để đáp ứng được yêu cầu dữ liệu lớn áp dụng cho hệ thống. Sau đó em dựng lên một hệ thống trên nền tảng ảo hóa Kubernetes để giả lập quá trình lấy dữ liệu theo chu kỳ, lấy và xử lý dữ liệu thời gian thực. Tiếp tiến hành lưu trữ dữ liệu và xử lý dữ liệu theo lô. Cuối cùng là thực hiện phân tích, truy vấn và trực quan hóa dữ liệu để cho ra kết quả cuối cùng. Luồng xử lý dữ liệu và các bước trong luồng xử lý đều được lập lịch và giám sát thực hiện.

Hệ thống được em xây dựng trong đồ án tốt nghiệp này đã thực hiện được một luồng xử lý dữ liệu lớn hoàn chỉnh, là một hệ thống phân tán, có các mô-đun chức năng riêng biệt, có khả năng mở rộng hệ thống, dễ dàng sử dụng và quản lý. Đặc biệt, hệ thống của em có thể đáp ứng được việc triển khai một sản phẩm trên môi trường thực tế trong việc lưu trữ và xử lý dữ liệu lớn cho doanh nghiệp giúp tiết kiệm chi phí.

Sinh viên thực hiện

(Ký và ghi rõ họ tên)

ABSTRACT

Big Data is increasingly regarded as an invaluable "resource" in both the present and the future. Storing and processing Big Data play a crucial role for businesses, particularly in determining the appropriate system model and technology for implementation. Currently, Big Data systems can be deployed using models such as data warehouse, data lake, and lakehouse, supported by various technological ecosystems like Hadoop, Amazon Web Services, and Google Cloud Platform. Selecting the right model and ecosystem to build these systems involves addressing challenges related to storage and processing capacity, costs, and security.

This project is developed with the goal of building a flight data lake system based on technologies from the Hadoop ecosystem. I used a dataset of U.S. flight information available on Kaggle to meet the Big Data requirements of the system. The system was deployed on a Kubernetes-based virtualization platform to simulate cyclical data collection and real-time data ingestion and processing. The project also included storing data, batch processing, analyzing, querying, and visualizing data to produce final results. The data processing flow and all steps within it were scheduled and monitored throughout execution.

The system developed in this graduation project successfully implements a complete Big Data processing flow. It is a distributed system with distinct functional modules, scalability, and ease of use and management. Notably, the system is capable of being deployed in real-world environments to support enterprises in storing and processing Big Data, thereby helping to reduce costs.

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU