

ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

Xây dựng hệ thống hồ dữ liệu phân tích dữ liệu
chuyển 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

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI.....	1
1.1 Đặt vấn đề.....	1
1.2 Mục tiêu và phạm vi đề tài.....	1
1.3 Định hướng giải pháp.....	2
1.4 Bố cục đồ án	3
CHƯƠNG 2. KHẢO SÁT VÀ PHÂN TÍCH YÊU CẦU.....	4
2.1 Khảo sát hiện trạng	4
2.2 Tổng quan chức năng	6
2.2.1 Yêu cầu chức năng	6
2.2.2 Yêu cầu phi chức năng.....	8
CHƯƠNG 3. CÔNG NGHỆ SỬ DỤNG.....	9
3.1 Apache Hadoop.....	9
3.1.1 HDFS	9
3.1.2 YARN.....	9
3.2 Apache Spark.....	10
3.2.1 Spark SQL	10
3.2.2 Spark Streaming	10
3.2.3 Spark YARN	11
3.3 Apache Airflow.....	11
3.4 Apache Kafka	12
3.5 Trino	13
3.6 Apache Hive	14
3.7 Apache Superset.....	14
3.8 Kubernetes.....	15

CHƯƠNG 4. THIẾT KẾ VÀ TRIỂN KHAI HỆ THỐNG	17
4.1 Thiết kế hệ thống.....	17
4.1.1 Kiến trúc tổng quan hệ thống	17
4.1.2 Mô-đun thu thập dữ liệu	18
4.1.3 Mô-đun lập lịch tác vụ.....	21
4.1.4 Mô-đun lưu trữ dữ liệu	23
4.1.5 Mô-đun xử lý dữ liệu.....	25
4.1.6 Mô-đun truy vấn dữ liệu	34
4.1.7 Mô-đun trực quan hóa dữ liệu	37
4.2 Triển khai hệ thống.....	38
4.2.1 Triển khai cụm Hadoop	39
4.2.2 Triển khai cụm Airflow	41
4.2.3 Triển khai nguồn phát dữ liệu.....	42
4.2.4 Triển khai cụm Kafka.....	43
4.2.5 Triển khai cụm Hive.....	44
4.2.6 Triển khai cụm Trino	45
4.2.7 Triển khai cụm Superset	46
4.2.8 Kết quả triển khai.....	46
CHƯƠNG 5. KẾT QUẢ THỰC NGHIỆM.....	49
5.1 Kết quả lập lịch và tự động hóa	49
5.2 Tính chịu lỗi của hệ thống	50
5.3 Tính khả mở của hệ thống	51
5.4 Kết quả lưu trữ dữ liệu	52
5.5 Hiệu suất truy vấn dữ liệu.....	55
5.6 Kết quả trực quan hóa dữ liệu	57

CHƯƠNG 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	62
6.1 Kết luận.....	62
6.2 Hướng phát triển.....	62
TÀI LIỆU THAM KHẢO.....	64

DANH MỤC HÌNH VẼ

Hình 2.1	Kiến trúc kho dữ liệu, hồ dữ liệu, hồ kho dữ liệu. Nguồn [2]	4
Hình 2.2	Sơ đồ phân rã chức năng hệ thống	7
Hình 4.1	Kiến trúc tổng quan hệ thống	17
Hình 4.2	Đồ thị luồng xử lý dữ liệu trên Airflow	21
Hình 4.3	Dữ liệu nguồn tại máy chủ Flask	24
Hình 4.4	Triển khai cụm HDFS	39
Hình 4.5	Triển khai cụm YARN	40
Hình 4.6	Triển khai cụm Airflow	41
Hình 4.7	Triển khai nguồn phát dữ liệu	42
Hình 4.8	Triển khai cụm Kafka	43
Hình 4.9	Triển khai cụm Hive	44
Hình 4.10	Triển khai cụm Trino	45
Hình 4.11	Triển khai cụm Superset	46
Hình 4.12	Cụm Kubernetes	46
Hình 4.13	Tổng quan thành phần của cụm Kubernetes	47
Hình 4.14	Các Deployment của hệ thống	47
Hình 4.15	Các Statefulset của hệ thống	47
Hình 4.16	Các Service của hệ thống	48
Hình 4.17	Các Persistent Volume Claim của hệ thống	48
Hình 5.1	Kết quả lập lịch tác vụ trên Airflow (1)	49
Hình 5.2	Kết quả lập lịch tác vụ trên Airflow (2)	49
Hình 5.3	Tính chịu lỗi trên cụm Kafka (1)	50
Hình 5.4	Tính chịu lỗi trên cụm Kafka (2)	50
Hình 5.5	Tính chịu lỗi khi sử dụng Kubernetes (1)	50
Hình 5.6	Tính chịu lỗi khi sử dụng Kubernetes (2)	51
Hình 5.7	Tính khả mở khi sử dụng Kubernetes (1)	51
Hình 5.8	Tính khả mở khi sử dụng Kubernetes (2)	51
Hình 5.9	Tính khả mở khi sử dụng Kubernetes (3)	52
Hình 5.10	Kết quả lưu trữ trên cụm Kafka (1)	52
Hình 5.11	Kết quả lưu trữ trên cụm Kafka (2)	52
Hình 5.12	Kết quả lưu trữ trên cụm Kafka (3)	53
Hình 5.13	Kết quả lưu trữ dữ liệu trên HDFS (1)	53
Hình 5.14	Kết quả lưu trữ dữ liệu trên HDFS (2)	54
Hình 5.15	Kết quả lưu trữ dữ liệu trên HDFS (3)	54

Hình 5.16	Hiệu suất truy vấn dữ liệu trên Trino (1)	55
Hình 5.17	Hiệu suất truy vấn dữ liệu trên Trino (2)	55
Hình 5.18	Hiệu suất truy vấn dữ liệu trên Trino (3)	56
Hình 5.19	Hiệu suất truy vấn dữ liệu trên Trino (4)	57
Hình 5.20	Độ trễ trung bình theo hãng hàng không năm 2018	58
Hình 5.21	Tỉ trọng chuyến bay và tỉ trọng chuyến bay bị hủy theo ngày trong tháng 1 năm 2019	58
Hình 5.22	Phân phối mạng lưới tiếp thị hàng không quý 1 năm 2020	59
Hình 5.23	Tỉ trọng chuyến bay và chuyến bay bị hủy theo từng hãng hàng không trong quý 4 năm 2021	60
Hình 5.24	Số chuyến bay theo tháng qua các năm	60

DANH MỤC BẢNG BIỂU

Bảng 2.1	So sánh kho dữ liệu, hồ dữ liệu và hồ kho dữ liệu	5
Bảng 4.1	Dữ liệu thời gian của chuyến bay	29
Bảng 4.2	Hãng vận chuyển tiếp thị và hãng vận chuyển điều hành của chuyến bay	30
Bảng 4.3	Địa điểm xuất phát và địa điểm đến của chuyến bay	31
Bảng 4.4	Thời gian khởi hành của chuyến bay	32
Bảng 4.5	Thời gian vận hành của chuyến bay	32
Bảng 4.6	Thời gian di chuyển của chuyến bay	33
Bảng 4.7	Thời gian trễ của chuyến bay	33
Bảng 4.8	Dữ liệu khác của chuyến bay	34
Bảng 4.9	Cấu hình của cụm Hadoop	40
Bảng 5.1	Hiệu suất câu truy vấn trên Trino	56