ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG
KHOA CÔNG NGHỆ THÔNG TIN

NGUYỄN THỊ THÚY LINH

NGHIÊN CỨU KIẾN TRÚC VÀ THUẬT TOÁN LÀM VIỆC CỦA HỆ THỐNG LƯU TRỮ VÀ PHÂN TÍCH DỮ LIỆU LỚN

LUẬN VĂN THẠC SỸ KHOA HỌC MÁY TÍNH

Thái Nguyên, năm 2023

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG
KHOA CÔNG NGHỆ THÔNG TIN

NGUYỄN THỊ THÚY LINH

NGHIÊN CỨU KIẾN TRÚC VÀ THUẬT TOÁN LÀM VIỆC CỦA HỆ THỐNG LƯU TRỮ VÀ PHÂN TÍCH DỮ LIỆU LỚN

LUẬN VĂN THẠC SỸ KHOA HỌC MÁY TÍNH

Chuyên ngành: Khoa học máy tính

Mã số: 8480101

Người hướng dẫn: TS. Trần Quang Quý

Thái Nguyên, tháng 07 năm 2023

LỜI CẢM ƠN

Em xin gửi lời cảm ơn sâu sắc tới thầy giáo hướng dẫn TS.Trần Quang Quý. Thầy đã giao đề tài và tận tình hướng dẫn em trong quá trình hoàn thành đề tài này.

Em xin gửi lời cảm ơn tới các thầy cô giáo trong Trường Đại học Công nghệ thông tin và truyền thông – Đại học Thái Nguyên, các thầy cô đã giảng dạy giúp đỡ em trong quá trình học tập tại trường.

Tôi xin chân thành cảm ơn các đồng nghiệp ở cơ quan, Trường Cao đẳng Lào Cai. Cảm ơn gia đình, bạn bè đã giúp tôi hoàn thành luận văn này.

Thái Nguyên 16 tháng 7 năm 2003

Học Viên

Nguyễn Thị Thúy Linh

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này do chính tôi thực hiện, dưới sự hướng dẫn của TS.Trần Quang Quý. Các kết quả lý thuyết được trình bày trong luận văn là sự tổng hợp từ các kết quả đã được công bố, kết quả của chương trình thực nghiệm trong luận văn này được tác giả thực hiện là hoàn toàn trung thực, nếu sai tôi xin chịu hoàn toàn trách nhiệm.

Thái Nguyên 16 tháng 7 năm 2003

Học Viên

Nguyễn Thị Thúy Linh

MỤC LỤC

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

LỜI NÓI ĐẦU

Dữ liệu lớn Big Data là một thuật ngữ cho việc xử lý một tập hợp dữ liệu rất lớn và phức tạp mà các ứng dụng xử lý dữ liệu truyền thống không xử lý được. Dữ liệu lớn bao gồm các thách thức như phân tích, thu thập, giám sát dữ liệu, tìm kiếm, chia sẻ, lưu trữ, truyền nhận, trực quan, truy vấn và tính riêng tư.

Có thể nói tuy thuật ngữ Big Data còn tương đối mới nhưng việc thu thập và lưu trữ một lượng lớn thông tin để phân tích đã được thực hiện từ rất lâu. Lượng dữ liệu đang được tạo ra và lưu trữ ở toàn cầu gần như không thể tưởng tượng được và lượng dữ liệu ấy không dừng lại mà ngày một phát triển tăng lên nhanh chóng. Điều đó nghĩa là có nhiều tiềm năng để thu thập thông tin chi tiết quan trọng nhưng chỉ một phần nhỏ dữ liệu thực sự được phân tích. Vậy làm thế nào để sử dụng tốt hơn những thông tin thô mà các tổ chức có thể thu thập mỗi ngày.

Tuy nhiên, hiện tại thì việc khai phá Big Data đang gặp một số hạn chế:

• Các tổ chức thiếu người tài để tận dụng sức mạnh của Big Data

• Thiếu kiến thức về thống kê, học máy, khai phá dữ liệu (một phần vì đây là vấn đề vẫn mang nặng tính nghiên cứu khoa học, thế nên nhân lực chủ yếu vẫn là giáo sư, tiến sĩ ở các trường công nghệ thông tin)

Những hạn chế này phản ảnh thực tế rằng khoa học cơ bản khó hiểu và khó sử dụng. Cũng như các công nghệ mới khác, công nghệ phân tích Big Data cần thời gian xây dựng và phát triển nhiều hơn.

Nội dung luận văn nhằm trả lời cho các câu hỏi hiện nay như hệ thống lưu trữ dữ liệu lớn làm việc như thế nào, cách lưu trữ, tính toán dữ liệu phân tán và thuật toán làm việc là gì? Big Data là một lĩnh vực đa ngành và đang ngày càng phát triển mạnh mẽ, đóng góp tích cực vào sự tiến bộ và thay đổi trong nhiều lĩnh vực khác nhau của xã hội và kinh tế.

CHƯƠNG 1. DỮ LIỆU VÀ HỆ THỐNG PHÂN TÁN

1.1. Hệ thống phân tán

Điện toán phân tán là một nhánh của khoa học máy tính nghiên cứu các khía cạnh của việc sắp xếp hệ thống phân tán. Nó bao gồm việc kết nối nhiều máy tính thành một mạng, trong đó giao tiếp giữa các máy tính là thông qua giao diện truyền thông phức tạp. Mục tiêu là xử lý dữ liệu và yêu cầu trên hàng trăm máy tính cùng hoạt động để đạt được một mục tiêu chung. Điện toán phân tán đặt ra nhiều thách thức và đã trở thành một lĩnh vực nghiên cứu quan trọng.

Hệ phân tán là tập hợp các máy tính được kết nối với nhau bởi một mạng máy tính và được cài đặt phần mềm hệ phân tán, đây là một hệ thống có chức năng và dữ liệu phân tán trên các trạm (máy tính) được kết nối với nhau bởi một mạng máy tính, ngoài ra nó còn là một tập hợp các máy tính độc lập giao tiếp với người dùng như một hệ thống nhất toàn vẹn.

Mặc dù việc xác định một hệ thống phân tán đã có từ khá lâu, tuy nhiên nó chưa bao giờ là hoàn chỉnh. Để mô tả một hệ thống phân tán, không chỉ dựa trên việc phân phối các thành phần vật lý, mà còn sử dụng các khả năng logic và chức năng của nó. Các chức năng này bao gồm:

- Nhiều quy trình: Hệ thống phân tán bao gồm nhiều hơn một quy trình tuần tự, mỗi quy trình có một luồng kiểm soát độc lập.
- Giao tiếp giữa các quy trình: Kênh giao tiếp giữa các quy trình là rất quan trọng. Độ tin cậy và thời gian trễ của các kênh phụ thuộc vào đặc điểm vật lý của các liên kết trên một nút hoặc mạng.
- Không gian địa chỉ rời rạc: Các quy trình có thể có không gian địa chỉ rời rạc, không chỉ dựa trên bộ nhớ dùng chung.