ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG NGHIÊN CỨU

MÔN CHUYÊN ĐỀ NGHIÊN CỨU 3

TÊN ĐỀ TÀI: ĐÁNH GIÁ HIỆU NĂNG VÀ TÍNH MỞ RỘNG CỦA HỆ THỐNG HADOOP DISTRIBUTED FILE SYSTEM

TÊN ĐỀ TÀI TIẾNG ANH: EVALUATING THE PERFORMANCE AND SCALABILITY OF THE HADOOP DISTRIBUTED FILE SYSTEM

GIẢNG VIÊN HƯỚNG DẪN
TS. LÊ DUY TÂN

HỌC VIÊN THỰC HIỆN
NGUYỄN HỒNG SƠN

1 Tóm tắt

Trong giai đoạn hiện nay, nhu cầu dữ liệu cho học máy cũng như các lĩnh vực khác liên tục tăng lên, do đó hệ thống lưu trữ chúng phải thích ứng với nhu cầu ngày càng tăng về hiệu suất, độ tin cậy và khả năng chịu lỗi. Điều này làm tăng độ phức tạp quản trị và chi phí. Cải thiện hiệu suất và tăng khả năng mở rộng của hệ thống trong khi vẫn duy trì chi phí thấp là rất quan trọng. Các giải pháp phần mềm lưu trữ được cho là hướng đi mới do giải pháp phần cứng vẫn quá đắt đỏ.

Hadoop Distributes File System (HDFS) do Apache Software Foundation phát triển đến nay đã có phiên bản thứ 3, là một giải pháp đáng tin cậy để lưu trữ và phân phối dữ liệu một cách đáng tin cậy trên nhiều node. Nghiên cứu này xem xét cách HDFS hoạt động trong thiết lập với các phần mềm hỗ trợ phân quyền và bảo mật khác như Kerberos, Openldap, Apache Ranger, YARN. Nghiên cứu thay đổi số lượng các node lưu trữ và tính toán để kiểm tra khả năng mở rộng của hệ thống. Nghiên cứu hướng đến việc thử nghiệm và kiểm tra tính chính xác các cam kết của HDFS, đồng thời phát hiện các điểm có tiềm năng cải tiến. Nghiên cứu này sẽ cải thiện hiểu biết về vệ thống HDFS cho cộng đồng, đồng thời thông qua các kịch bản được trình bày để giải quyết các hạn chế về hiệu suất của HDFS.

2 Giới thiệu

Sự ra đời của Big Data và Machine Learning đã thúc đẩy nhiều nhà khoa học và các tập đoàn tập trung sang phát triển lĩnh vực này, khiến nhu cầu về lưu trữ dữ liệu tăng nhanh. Do đó, các giải pháp quản lý và lưu trữ dữ liệu mới được phát triển. Trong nghiên cứu này, tác giả giới thiệu và thực sự một số đánh giá hiệu năng của HDFS[1], một phần mềm cung cấp hệ thống lưu trữ phân tán được thiết kế với khả năng dễ mở rộng, có hiệu quả cao, khả năng chịu lỗi tốt, đồng thời có chi phí triển khai thấp.

Sự phức tạp của hệ phân tán nói chung và HDFS nói riêng đòi hỏi các cấu hình và điều chỉnh cẩn thận để đạt hiệu suất tốt, các thay đổi được thực hiện ảnh hưởng tới toàn hệ thống và có khả năng làm mất mát dữ liệu nếu không được thao tác đúng. Do đó, tìm hiểu và đánh giá hiệu suất toàn diện của hệ thống là cần thiết. Trong nghiên cứu này, tác giả xây dựng và trả lời các câu hỏi sẽ giúp đáp ứng một số nhu cầu lưu trữ dữ liệu phân tán lớn, yêu cầu khả năng chịu lỗi cao, thuận tiện trong phân quyền và đồng thời tốc độ truy cập nhanh. Khả năng tích hợp với các ứng dụng khác cũng được khảo sát và đề xuất. Trong trường hợp nào thì hệ thống hoạt động tốt nhất? Nếu như vậy thì cái giá phải đánh đổi là gì? Hệ thống hoạt động thế nào khi dữ liệu tăng lên cao hoặc nhiều người dùng cùng truy cập? Khi xảy ra hư hỏng phần cứng thì ứng dụng sẽ điều hướng lưu trữ như thế nào? Các giải pháp theo dõi hiệu suất hệ thống khả dụng là gì? Khi nào thì hiệu suất của hệ thống giảm?

Phần còn lại của nghiên cứu được trình bày như sau: Phần 3 trình bày các thành phần cơ bản của HDFS và các nghiên cứu liên quan. Phần 4 trình bày thiết kế hệ thống và khác thực nghiệm. Phần 5 trình bày chi tiết các kết quả. Cuối cùng, kết luận và đề xuất được trình bày ở phần 6.

3 Cơ sở lý thuyết và các nghiên cứu liên quan

4 Phân tích thiết kế hệ thống

5 Thực nghiệm và kết quả

6 Kết luận và phương hướng

Mục lục

Tài liệu tham khảo

[1] Dhruba Borthakur. “The hadoop distributed file system: Architecture and design”. In: *Hadoop Project Website* 11.2007 (2007), p. 21.