XỬ LÝ DỮ LIỆU LỚN DÙNG Hadoop và Map Reduce

Aditya B. Patel, Manashvi Birla, Ushma Nair

Tổng quan: Ngày nay, quy mô của cơ sở dữ liệu được sử dụng trong các doanh nghiệp đã tăng theo cấp số nhân. Đồng thời, nhu cầu xử lý và phân tích một khối lượng dữ liệu lớn giúp doanh nghiệp đưa ra quyết định cũng tăng lên. Trong một số ứng dụng thuộc lĩnh vực kinh doanh và nghiêu cứu khoa học, cần phần xử lý hàng terabyte dữ liệu một cách hiệu quả. Điều nầy đã nảy sinh vấn đề về dữ liệu lớn mà ngành công nghiệp phải đối mặt do hệ thống cơ sở dữ liệu thông thường không có khả năng đáp ứng cũng như các công cụ phần mềm để quản lý hoặc xử lý các tập dữ liệu lớn trong giới hạn cho phép. Việc xử lý dữ liệu bao gồm các hoạt động khác nhau là tùy vào cách thức sử dụng như : Loại bỏ, gắn thẻ, làm nổi bật, lập chỉ mục, tìm kiếm, nhận diện khuôn mặt, v..v… Từ đó không thể dùng một hoặc một vài máy tính để lưu trữ hoặc xử lý khối dữ liệu khổng lồ nầy trong một khoảng thời gian qui định. Bài viết này công bố công việc thử nghiệm về dữ liệu lớn và giải pháp tối ưu của nó bằng cách sử dụng Hadoop cluster, hệ thống tập tin phân tán Hadoop(HDFS) dùng để lưu trữ và xử lý song song để từ đó xử lý tập dữ liệu lớn dùng trên nền tảng với cớ chế Map Reduce. Việc triển khai trên mẫu Hadoop cluster, HDFS lưu trữ và xử lý tập dữ liệu lớn bằng cách xem xét dựa trên mẫu kịch bản ứng dụng dữ liệu lớn. Các kết quả thu được từ những thử nghiệm cho ra những kết quả khả quan nhờ tiếp cận Hadoop cluster để giải quyết vấn đề dữ liệu lớn.

I. GIỚI THIỆU

Trong thời đại tin học điện tử ngày nay, ngày càng nhiều tổ chức đang phải đối mặt với vấn đề bùng nổ dữ liệu và quy mô của cơ sở dữ liệu được sử dụng trong các doanh nghiệp ngày nay đã tăng lên theo cấp số nhân. Dữ liệu được tạo thông qua nhiều nguồn như quá trình kinh doanh, giao dịch, trang web mạng xã hội, máy chủ web, v.v. và nó tồn tại ở dạng cấu trúc cũng như không có cấu trúc [1]. các ứng dụng kinh doanh ngày nay đang có các tính năng doanh nghiệp như quy mô lớn, sử dụng nhiều dữ liệu, định hướng web và được truy cập từ các thiết bị khác nhau bao gồm cả thiết bị di động. Xử lý hoặc phân tích lượng dữ liệu khổng lồ hoặc trích xuất thông tin có ý nghĩa là một nhiệm vụ đầy thách thức.

II. BIG DATA

Thuật ngữ "Big Data" được sử dụng cho các tập dữ liệu lớn có kích thước vượt khả năng của những công cụ phần mềm thường được sử dụng để capture, manage và xử lý dữ liệu trong khoảng thời gian khá nhanh. Kích thước dữ liệu lớn là mục tiêu thay đổi liên tục và hiện đang dao động từ vài chục terabyte đến nhiều petabyte dữ liệu trong một tập dữ liệu. Những thách thức gặp phải gồm capture, lưu trữ, tìm kiếm, chia sẻ, phân tích và trực quan hóa. Một số ví dụ điển hình của Big data hiện nay được thu thập từ các nguồn web logs, dữ liệu được tạo bằng RFID, mạng cảm biến, dữ liệu vệ tinh và không gian địa lý, dữ liệu từ mạng xã hội, văn bản và tài liệu từ Internet, các keyword từ Internet, ghi nhận chi tiết cuộc gọi, ngành thiên văn học, khoa học khí quyển , nghiên cứu về cấu trúc, chức năng và sự phát triển của bộ gen, lĩnh vực hóa sinh, sinh học, và nghiên cứu khoa học phức tạp có liên hệ vơi một số lĩnh vực khác như giám sát quân sự, hồ sơ y tế, nhiếp ảnh, lưu trữ video và thương mại điện tử với quy mô lớn. Các tác động của Big data mà cụ thể là Walmart phải xử lý hơn 1 triệu giao dịch khách hàng mỗi giờ, hàng triệu khách hàng đó được imported vào cơ sở dữ liệu ước tính hơn 2,5 petabyte dữ liệu - tương đương với 167 lần thông tin chứa tất cả các cuốn sách trong Thư viện Quốc hội Hoa Kỳ và bản thân Facebook cũng xử lý khoản 40 tỷ bức ảnh từ người dùng của họ.

A. Vấn đề Big data là gi ?

Dữ liệu lớn đã hiện hữu bởi vì chúng ta đang sống trong một xã hội mà con người tạo nên sự gia tăng do sử dụng nhiều công nghệ có liên quan đến dữ liệu. Một tính năng hiện tại của dữ liệu lớn là gặp khó khăn khi làm việc với cơ sở dữ liệu quan hệ và các gói thống kê / trực quan hóa từ máy tính để bàn, thay vì đòi hỏi "phần mềm xử lý song song trên hàng chục, hàng trăm hoặc thậm chí hàng nghìn máy chủ". Những thách thức khác nhau phải đối mặt trong quản lý dữ liệu lớn cũng bao gồm - khả năng mở rộng, dữ liệu phi cấu trúc, khả năng truy cập, phân tích thời gian thực, khả năng chịu lỗi và nhiều hơn thế nữa. Ngoài các biến thể về lượng dữ liệu được lưu trữ trong các sector khác nhau như các dạng dữ liệu video, hình ảnh, âm thanh hoặc thông tin văn bản / dạng số có nên mã hóa hay không cũng có sự khác biệt rõ rệt từ các lĩnh vực với nhau.

B. Kỹ thuật và công nghệ trong Big data

Big data cần những công nghệ đặc thù để xử lý lượng lớn dữ liệu môt cách hiệu quả trong thời gian chấp nhận được. Những công nghệ đang được áp dụng cho Bisg data gồm cơ sở dữ liệu xử lý song song (MPP), khai thác dữ liệu qua gids, hệ thống tập tin phân tán, cơ sở dữ liệu phân tán, nền tảng điện toán đám mây, Internet và những hệ thống lưu trữ mở rộng. Việc điều phối thông tin theo thời gian thực là một trong những đặc tính của phân tích dữ liệu lớn. Độ trễ do đó có thể được tránh bất cứ khi nào . Một loạt các kỹ thuật lẫn công nghệ phát triển và điều chỉnh nhằm tổng hợp, thao tác, phân tích và trực quan hóa dữ liệu lớn [2]. Những kỹ thuật lẫn công nghệ này có được từ một số lĩnh vực như thống kê, khoa học máy tính, toán học ứng dụng và kinh tế. Điều này có nghĩa khi náo một tổ chức có dư định lấy dữ liệu từ Big data phải áp dụng cách tiếp cận một cách linh hoạt, đa lĩnh vực.

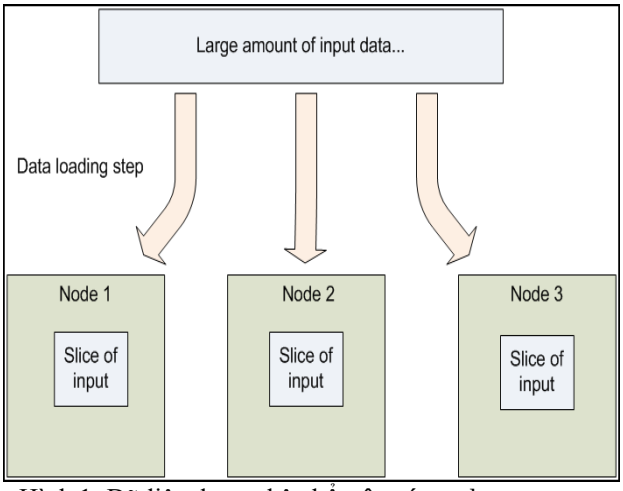
C. Hadoop

Dự án Apache Hadoop phát triển dựa trên phần mềm mã nguồn mở cho hệ thống tính toán phân tán, có độ tin cậy và có thể mở rộng. Thư viện phần mềm Apache Hadoop

là một framework cho phép xử lý phân tán các tập dữ liệu lớn trên các cụm máy tính bằng một mô hình lập trình đơn giản. Nó cho phép các ứng dụng hoạt động với hàng ngàn máy tính độc lập tính toán với hàng petabyte dữ liệu. Hadoop được phát triển từ MapReduce và GFS của Google.

D- HDFS (Hadoop Distributed File System) [5]

Hadoop Distributed File System (HDFS) là một hệ thống tệp phân tán cung cấp khả năng chịu lỗi và được thiết kế để chạy trên phần cứng cơ bản. HDFS cung cấp quyền truy cập thông lượng cao vào dữ liệu ứng dụng và phù hợp với các ứng dụng có tập dữ liệu lớn. Hadoop cung cấp một hệ thống tệp phân tán (HDFS) có thể lưu trữ dữ liệu trên hàng ngàn máy với ý nghĩa lá nó thực hiện công việc Map/Reduce trên những node đó. HDFS có kiến trúc master/sliver. Dữ liệu lớn được tự động cắt ra thành các khối và được quản lý bởi các node khác nhau trong hadoop cluster.



Hình 1: Dữ liệu được phân bổ trên các node

A. Nền tảng chương trình MapReduce [6]

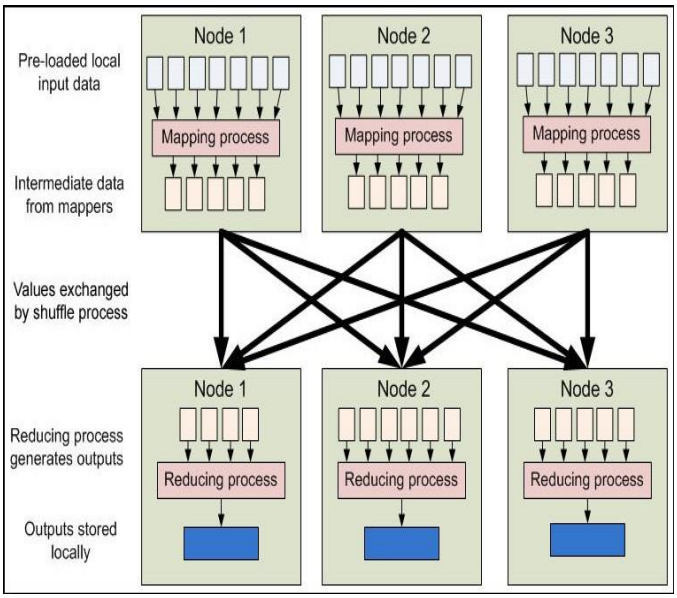
MapReduce là nền tảng phần mềm được Google giới thiệu vào năm 2004 nhằm hỗ trợ tính toán phân tán trên các tập dữ liệu lớn trên các cụm máy tính có nối mạng. MapReduce là một mô hình lập trình để xử lý và tạo các tập dữ liệu lớn. Người dùng chỉ định một hàm map xử lý cặp key/value để tạo cặp key/ value trung gian và hàm reduce gôm nhóm tất cả các giá trị trung gian có cùng key.

Bước "Map": Node chính nhận dữ liệu đầu vào, chia nhỏ và điều phối đến các node worker. Node worker thực hiện lại lần lượt việc điều phối nầy nhằm hình thành cấu trúc cây phân cấp. Node worker xử lý phần dữ liệu được chỉ định sau đó phản hồi lại với Master node. Phần việc của Map là lấy cặp key/value ở dữ liệu đầu vào, xử lý trả về kết quả là một list ở đầu ra :

Map (k1, v1) → list (K2, v2)

Bước "Reduce": Trong bước nầy, Master node ghi nhận tất cả các phản hồi và tổng hợp để làm kết quả đầu ra, kế tiếp hàm reduce lần lượt xử lý song song trên mỗi nhóm để gôm các giá trị trong cùng nhóm.

Reduce (K2, list (v2)) → list (v3)



Hình 2: Distributed Map and Reduce processes

III-KIẾN TRÚC HỆ THỐNG

Kiến trúc hệ thống cũng đồng nghĩa với kiến trúc hadoop vì nó được triển khai trong một cluster có nhiều node và công việc cài đặt HDFS lẫn triển khai MapReduce nhằm giải quyết vấn đề phức tạp của dữ liệu.

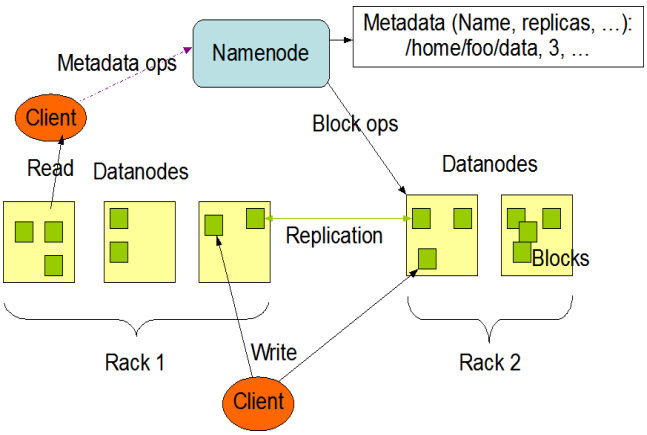
A. Kiến trúc HDFS [6]

Như miêu tả trong Hình 3 , một cụm HDFS gồm một NameNode duy nhất, một master server quản lý một file namespace và quy định cách mà client truy cập tập tin. Ngoài ra, trong số DataNodes, thường thì mỗi node tự quản lý việc lưu trữ. Hệ thống HDFS đưa ra một file namespace và cho phép dữ liệu người dùng được lưu trữ thành các tệp. Trong đó, một tệp được chia thành một hoặc nhiều block và các block được lưu trữ trong một dãy DataNode. NameNode đảm nhiệm map các block tới các DataNodes. HDFS cũng được thiết kế để lưu trữ với số lượng rất lớn các tập tin đáng tin cậy trên một cluster lớn gồm nhiều máy. Nó lưu trữ tập tin thành một dãy các khối.

B-Kiến trúc phân cấp Hadoop cluster

Hadoop cluster gồm một master và nhiều slaver hay còn gọi là các "worker nodes". Trong khi JobTracker là dịch vụ trong Hadoop cung cấp các tác vụ MapReduce cho

các nút được chỉ định trong cluster, lý tưởng là các node có chứa dữ liệu hoặc ít nhất thì có cùng rack.



Hình 3: Kiến trúc HDFS

Tasktracker là một node trong cluster nhận những nhiệm vụ và thực thi công việc Map, Reduce và Shuffle từ Jobtracker. Trong khi master node gồm JobTracker, TaskTracker, NameNode và DataNode. Một slave hay worker hoạt động với vai trò vừa là DataNode vừa là TaskTracker. Trong một cluster lớn như vậy, hệ thống HDFS được quản lý thông qua NamdeNode server bằng các hệ thống tập tin chỉ mục, nhằm ngăn chặn việc xảy ra sự cố và mất dữ liệu thì NamdeNode secondary sẽ chụp lại một bản dự phòng cho NameNode master.

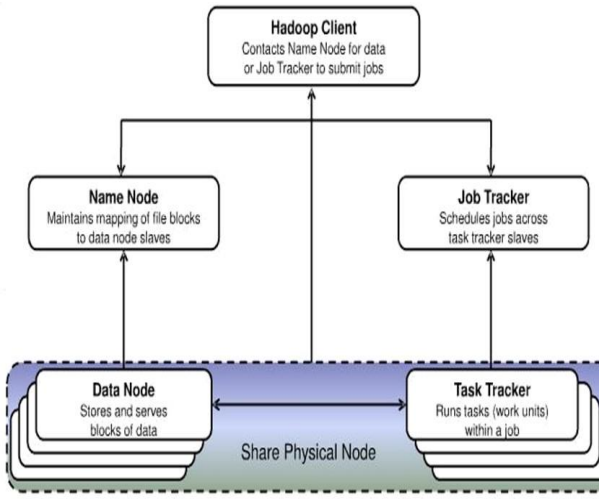
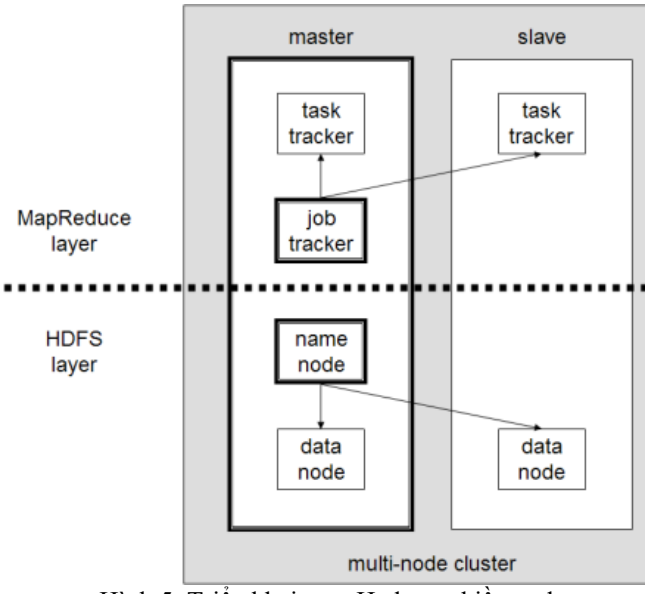


Figure 4: Kiến trúc phân cấp Hadoop+

IV.THỰC NGHIỆM CÀI ĐẶT

Để thực nghiệm với Big data, việc triển khai Hadoop được sử dụng 4 nodes và hệ thống lưu trữ tập tin (HDFS). Để hướng đến một cluster nhiều node, thì trước tiên một node cluster phải được cấu hình và kiểm tra. Hadoop có rất nhiều tham số cấu hình được mô tả ở đây, nhưng mục đích quan trọng nhất đó là những tác vụ Map và Reduce được phép chạy trên mỗi node. Chúng tôi đã cấu hình trong cluster từ 4 đến 8 tác vụ trên mỗi server. Mỗi khi chương trình Map/Reduce được thực thi thì nó được chia thành những tác vụ map ký hiệu là M, còn những tác vụ Reduce gọi là R. Dữ liệu vào/ra cho mỗi lần thực thi Map/Reduce được lưu trên HDFS, trong khi dữ liệu vào/ra song song dựa trên cơ chế stack được lưu trực tiếp trên những đĩa lưu trữ cục bộ. Một node được cấu hình giữ vai trò Master node trong khi những node còn lại được cấu hình giữ vai trò là những slave node. Master node chạy ngầm một "master" daemons: DataNode ở tầng HDFS trong khi TaskTracker thuộc tầng xử lý MapReduce. Master node cũng được dùng như một slave node để tăng quá trình xử lý các node. Phần mềm để sử dụng làm master và slave node là Sun Java phiên bản 1.6, Ubuntu Linux 10.04 LTS and hadoop phiên bản 1.0.3.



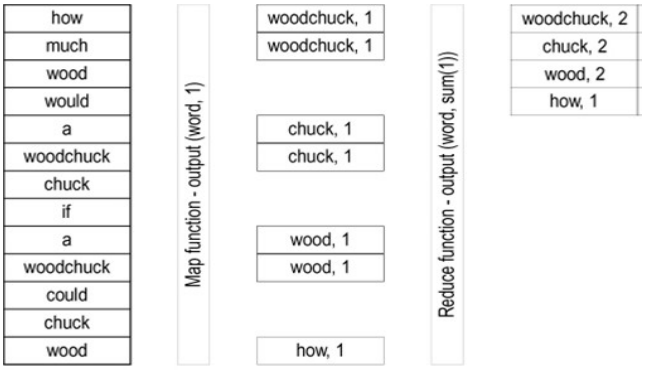
Hình 5: Triển khai cụm Hadoop nhiều node

V-THỰC NGHIỆM VÀ CÁC KẾT QUẢ ĐẠT ĐƯỢC

Thực nghiệm đầu tiên là xử lý từ dạng text bằng cách đếm số lượng từ xuất hiện trong những tài liệu có kích thước lớn. Hàm map sẽ cắt mỗi tập tin tài liệu thành nhiều từ và cho ra cùng với nhau ứng với mỗi từ có giá trị "1". Vì thế các record đầu ra có dạng (word,1). MapReduce gôm nhóm các record có cùng giá trị (i.e,

word) và chuyển qua hàm reduce. Hàm reduce tính tổng các giá trị đầu vào và xuất ra từ và tổng số từ có trong các tài liệu.

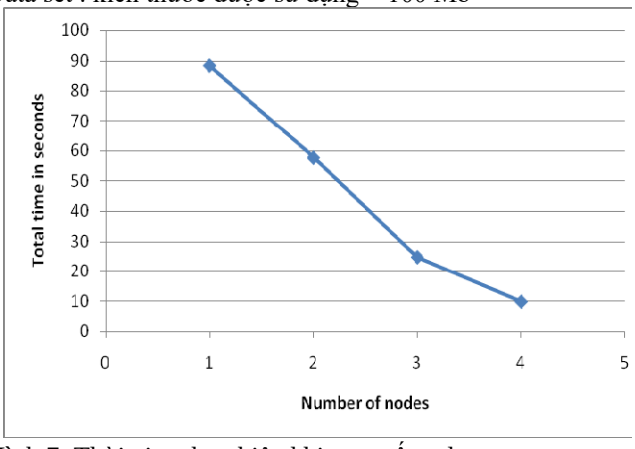
A- Ứng dụng xử lý dạng Text



Hình 6: Map Reduce đếm số từ

1) Thực nghiệm với việc tang số node

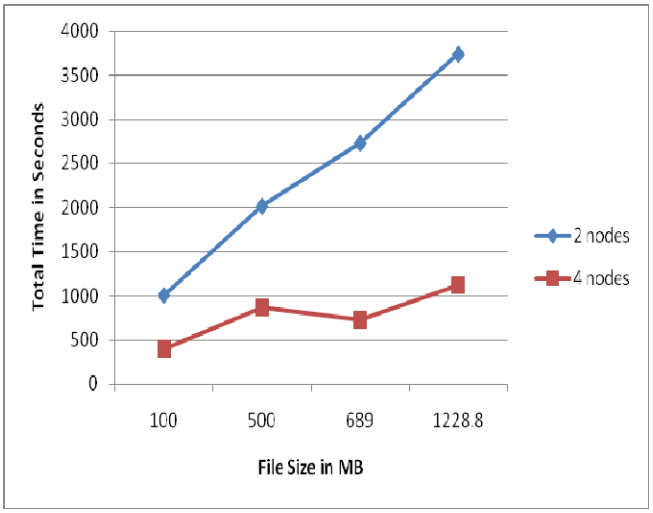
Data set : kích thước được sử dụng = 100 Mb



Hình 7: Thời gian thực hiện khi tang số node

Kết quả trên cho thấy thời gian thực thi giảm khi tăng số node trong Hadoop cluster.

2-Thực hiện với việc tăng tập dữ liệu và số lượng node



Hình 8: Thực gian thực thi với việc tăng tập dữ liệu và node

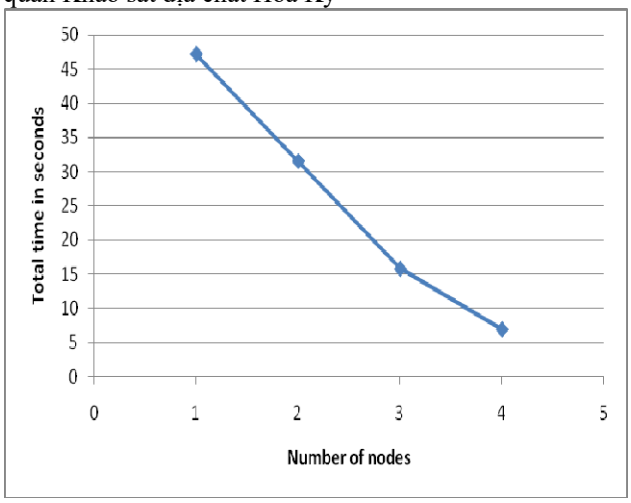
Kết quả trên cho thấy thời gian thực thi giảm khi tăng số node trong Hadoop cluster.

B-Phân tích dữ liệu của trận động đất

Trong thí nghiệm thứ hai, chúng tôi đã phân tích dữ liệu trận động đất đã công bố bởi Cơ quan Khảo sát Địa chất Hoa Kỳ (USGS). Dữ liệu động đất có sẵn ở dạng tệp định dạng CSV (các giá trị ghi nhận phân cách bằng dấu phẩy) được xuất bản theo định kỳ. Việc phân tích dữ liệu trận động đất cung cấp câu trả lời cho vị trí, nơi xảy ra trận động đất, số lượng trận động đất vào ngày cụ thể.

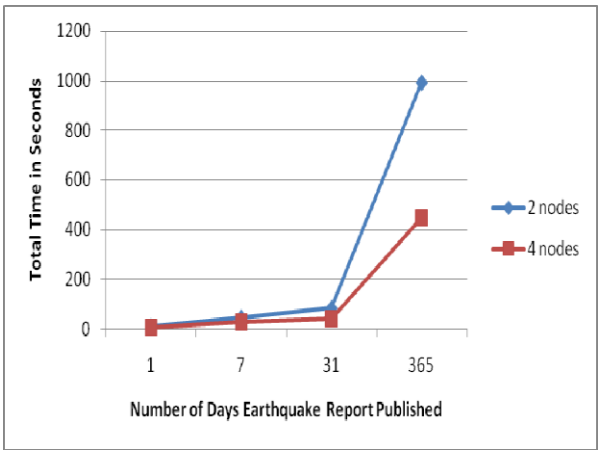
1)Thực nghiệm với việc tăng số lượng node

Tập dữ liệu – 7: Ngảy báo cáo trận động đất của cơ quan Khảo sát địa chất Hoa Kỳ



Hình 9: Phân tích trận động đất – Số lượng node và thời gian thực hiện

2)Thực hiện với việc tăng tập dữ liệu và số lượng node



Hình 10: Phân tích động đất – Số ngày báo cáo vs thời gian thực hiện

Kết quả thử nghiệm ở trên chỉ ra rằng Hadoop cluster có

thể mở rộng để hỗ trợ tập dữ liệu tăng lên và mất ít thời gian thực hiện hơn với tăng số lượng node.

VI-KẾT LUẬN VÀ THỰC NGHIỆM TRONG TƯƠNG LAI

Trong thí nghiệm này, chúng tôi đã khám phá giải pháp cho vấn đề dữ liệu lớn bằng cách sử dụng dữ liệu Hadoop cluster, cùng với các tình huống ứng dụng mẫu dữ liệu lớn vào HDFS và nền tảng lập tình Map Reduce. Các kết quả thu được từ những lần thí nghiệm khác nhau cho kết quả khả quan từ phương pháp đề cập ở trên nhằm giải quyết vấn đề dữ liệu lớn. Trong tương lại, thí nghiệm sẽ tập trung vào đánh giá hiệu suất, mô hình hóa những ứng dụng dựa trên nền tảng đám mây như Amazon Elastic Compute Cloud (EC2).

VII-NGUỒN TRÍCH DẪN

[1] Impetus white paper, March, 2011, "Planning Hadoop/NoSQL Projects for 2011" by Technologies, Available: http://www.techrepublic.com/whitepapers/planning- hadoopnosql-projects-for-2011/2923717, March, 2011.

[2] McKinsey Global Institute, 2011, Big Data: The next frontier for innovation, competition, and productivity, Available: www.mckinsey.com/~/media/McKinsey/dotcom/Insights %20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ash x, Aug, 2012.

[3] Thomas Herzog, Associate Commissioner, New York State, Thomas Kooy, IJIS Institute Big Data and the Cloud, IJIS Institute Emerging Technologies, Available: http://www.correctionstech.org/meeting/2012/Presentatio ns/Red_01.pdf, Aug, 2012.

[4] Jacobs, A., The Pathologies of Big Data, ACM Queue, Available: http://queue.acm.org/detail.cfm?id=1563874, 6th July 2009.

[5] Apache Software Foundation. Official apache hadoop website, http://hadoop.apache.org/, Aug, 2012.

[6] The Hadoop Architecture and Design, Available: http://hadoop.apache.org/common/docs/r0.16.4/hdfs_des ign.html, Aug, 2012

[7] Hung-Chih Yang, Ali Dasdan, Ruey-Lung Hsiao, and D. Stott Parker from Yahoo and UCLA, "Map-Reduce- Merge: Simplified Data Processing on Large Clusters", paper published in Proc. of ACM SIGMOD, pp. 1029– 1040, 2007.

[8] White, Tom. Hadoop The Definitive Guide 2nd Edition. United States : O'Reilly Media, Inc., 2010.