

**ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH**



LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC

**NGHIÊN CỨU VÀ XÂY DỰNG TẬP DỮ LIỆU HỮU
ÍCH CHO BÀI TOÁN CDNS (CONTENT
DELIVERY/DISTRIBUTION NETWORKS) DỰA TRÊN
PHÂN TÍCH LOG FILES CỦA HỆ THỐNG SERVERS**

HỘI ĐỒNG: HỆ THỐNG VÀ MẠNG

GVHD : PGS. TS. THOẠI NAM

GVPB : ThS. NGUYỄN CAO ĐẠT

SVTH : NGUYỄN TUẤN MINH (1412305)

TP. HỒ CHÍ MINH, 06/2018

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của tôi và được sự hướng dẫn khoa học của PGS. TS. Thoại Nam. Các nội dung trong luận văn này, kết quả trong đề tài này là trung thực và chưa được công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được tôi thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo. Ngoài ra, trong luận văn còn có sử dụng một số hình ảnh, nhận xét, đánh giá cũng như số liệu của các tác giả, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc. Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về mặt nội dung luận văn tốt nghiệp của mình. Trường đại học Bách Khoa – Đại Học Quốc Gia Thành Phố Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày 11 tháng 06 năm 2018

LỜI CẢM ƠN

Trên thực tế không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp của người khác. Trong suốt thời gian từ khi bắt đầu học tập ở trường đại học Bách Khoa – Đại Học Quốc Gia Thành Phố Hồ Chí Minh cho đến nay, tôi đã nhận được rất nhiều sự quan tâm, giúp đỡ của quý thầy, cô, gia đình và bạn bè.

Với lòng biết ơn chân thành và sâu sắc nhất, tôi xin gửi lời cảm ơn đến PGS. TS. Thoại Nam. Thầy đã tạo điều kiện cho tôi có cơ hội được tham gia và thực hiện đề tài với chủ đề **“Nghiên cứu và xây dựng tập dữ liệu hữu ích cho bài toán CDNs (content delivery/distribution networks) dựa trên phân tích log files của hệ thống Servers”**. Trong quá trình thực hiện, thầy thường xuyên trao đổi, hỗ trợ tài liệu và giải đáp mọi thắc mắc liên quan đến nội dung cần tìm hiểu. Qua đó, tôi có thể trau dồi kiến thức và kỹ năng giải quyết vấn đề cũng như định hướng được các công việc trong quá trình thực hiện luận văn tốt nghiệp.

Bên cạnh đó, tôi cũng gửi lời cảm ơn đến các thành viên của phòng thí nghiệm Tính Toán Hiệu Năng Cao đã hết lòng hỗ trợ tôi về mặt máy móc, kỹ thuật trong suốt quá trình làm luận văn tốt nghiệp.

Ngoài ra, tôi cũng xin gửi lời cảm ơn chân thành đến các thầy, cô của trường đại học Bách Khoa – Đại Học Quốc Gia Thành Phố Hồ Chí Minh nói chung cũng như các thầy, cô của khoa Khoa Học và Kỹ Thuật Máy Tính nói riêng, đã cung cấp những kiến thức nền tảng vô cùng quý giá để chúng tôi có thể áp dụng nó vào đề tài luận văn tốt nghiệp.

Bước đầu đi vào lĩnh vực sáng tạo trong nghiên cứu khoa học, kiến thức của tôi vẫn còn nhiều hạn chế. Do đó, không tránh khỏi những thiếu sót là điều chắc chắn, tôi rất mong nhận được những ý kiến đóng góp quý báu của quý thầy, cô và các bạn sinh viên để tôi ngày càng hoàn thiện hơn.

Sau cùng, tôi xin kính chúc các thầy, cô trong Khoa Khoa Học và Kỹ Thuật Máy Tính nói chung, PGS. TS. Thoại Nam nói riêng thật dồi dào sức khỏe, niềm tin để tiếp tục thực hiện sứ mệnh truyền đạt kiến thức cao đẹp của mình cho thế hệ sinh viên mai sau.

Xin chân thành cảm ơn!

TÓM TẮT LUẬN VĂN

Các nội dung mới ngày càng nhiều, khi các người dùng truy cập đến các nội dung nhiều lên một cách đáng kể sẽ gây ra tắc nghẽn cho các hoạt động trên mạng. Nếu cứ tắc nghẽn nhiều sẽ gây ra kết quả không tốt như thời gian phản hồi quá lâu, có thể mất dữ liệu, tốn băng thông,... Giải pháp mạng phân phối nội dung (CDN) nhằm giải quyết tình trạng đó và cải thiện các hoạt động và tăng tính hiệu quả khi truy cập trên mạng. Trong đề tài luận văn này, tôi xây dựng các tập dữ liệu về các yêu cầu người dùng để phục vụ cho bài toán CDN.

Tôi sử dụng log file của trường làm dữ liệu để thực hiện tìm hiểu và phân tích. Tôi sử dụng các công cụ, phần mềm hỗ trợ cho việc phân tích như Apache Spark. Sử dụng phần mềm Tableau để trực quan hóa các tập dữ liệu mới.

Spark phân tích các log file lớn để tạo ra tập dữ liệu có ý nghĩa. Thực hiện loại bỏ bớt các dòng lỗi có trong log bằng các phép so sánh và biểu thức chính quy. Sau đó thống kê lại các nội dung được truy cập bao nhiêu lần để sinh ra một tập dữ liệu mới. Qua các IP có trong log, chúng ta có thể lấy được thêm các thông tin về quốc gia, thành phố và tổ chức của các ISP. Tổng hợp với các nội dung có trong log sẽ thành một tập dữ liệu về vị trí của các người dùng truy cập đến.

Sau khi có được các tập dữ liệu, tôi cố gắng trực quan hóa các dữ liệu đó dưới dạng hình ảnh nhờ vào Tableau. Tableau hỗ trợ ta tạo ra rất nhiều biểu đồ. Chúng ta sẽ tạo biểu đồ về lượng truy cập của các quốc gia trên thế giới đến Server của trường. Kế tiếp đó tôi sẽ tạo ra thêm biểu đồ thống kê lượng truy cập của các thành phố ở Việt Nam. Cuối cùng có biểu đồ về phân bố nội dung truy cập theo các tháng trong năm.

Kết quả đạt được các tập dữ liệu về nội dung truy cập của người dùng để áp dụng cho các Cache Server của mạng CDN được hoạt động.

Trong tương lai, chúng ta còn có thể phát triển thêm về các nội dung khác trong log như thời gian, liên quan giữa các nội dung truy cập với nhau,... để có thể sinh ra các tập dữ liệu có ý nghĩa khác.

ABSTRACT

These new contents are more and more, when users access to the content of a considerable number of ways will lead to a blockage of the active network. A blockage can cause a lot of bad results, such as the long reaction time; the data may be lost; the bandwidth is wasted,... Content delivery network (CDN) resolve the situation, improve the operation and increase the efficiency when we accessed on the network. I will build the data set are useful for CDN.

I use the log file of the university to do data to make understanding and analysis. I use the tools, the support software for analyzing such as Apache Spark. Using Tableau software to visualize the data set new.

Spark analyzes large log files to meaningful data collection. Make the type of deregulation error line in the log by the comparisons and regular expressions. Then the statistics content to be accessed many times to produce a new data file. Through the IPS are in the log, we can get more information about countries, cities and organizations of the ISP. With the content contained in the log file data to be a position of access to users. After you get the data set, I tried to indicate that data into the chart by the Tableau.

Tableau supports have created a lot of diagrams. We will create a graph of the traffic of the countries in the world to the university Server. Next I will create more traffic statistics graph of the cities in Vietnam. Finally there is the chart of the distribution of content accessed by the months of the year.

Results achieved datasets about the user's access to content to apply for the Cache Server of the CDN network action.

In the future, we can develop more of the other items in the log as the time, the relationship between the content accessible to each other, ... to be able to meaningful data sets other.

MỤC LỤC

LỜI CAM ĐOAN.....	i
LỜI CẢM ƠN	ii
TÓM TẮT LUẬN VĂN	iii
ABSTRACT	iv
MỤC LỤC	v
DANH MỤC TỪ VIẾT TẮT	vii
DANH MỤC HÌNH ẢNH	viii
CHƯƠNG 1 : TỔNG QUAN	1
1.1 Giới thiệu đề tài	1
1.2 Lý do lựa chọn đề tài	1
1.3 Mục tiêu đề tài	1
1.4 Đối tượng và phạm vi nghiên cứu	2
1.5 Phương pháp nghiên cứu và hướng tiếp cận	2
1.6 Cấu trúc luận văn.....	2
CHƯƠNG 2 : KIẾN THỨC NỀN TẢNG	4
2.1 Hệ thống mạng phân phối nội dung.....	4
2.2 Cấu trúc log file	7
2.3 Apache Spark.....	8
2.4 Apache Zeppelin.....	13
2.5 Tableau	15
CHƯƠNG 3 : HIỆN THỰC	18
3.1 Phân tích log file sử dụng Spark.....	18
3.1.1 Loại bỏ các lỗi trong tập dữ liệu ban đầu	18
3.1.2 Rút gọn tập dữ liệu	20
3.1.3 Tập dữ liệu truy cập thật sự	20
3.1.4 Xây dựng tập dữ liệu vị trí các IP	21
3.1.5 Xây dựng tập dữ liệu thống kê truy cập	23
3.2 Cách trực quan hóa dữ liệu	23
3.2.1 Biểu đồ trực quan truy cập toàn thế giới	24
3.2.2 Biểu đồ truy cập ở Việt Nam.....	26
3.2.3 Biểu đồ lượng truy cập của các ISP tại Việt Nam.....	27
3.2.4 Biểu đồ độ phổ biến nội dung trong một năm	27
CHƯƠNG 4 : ĐÁNH GIÁ	28
4.1 Môi trường thực nghiệm.....	28
4.2 Phân tích dữ liệu dùng Spark.....	29

4.2.1 Tập dữ liệu về thống kê lượng truy cập.....	29
4.2.2 Tập dữ liệu về vị trí và ISP của IP.....	30
4.3 Trực quan hóa dữ liệu dùng Tableau.....	30
4.3.1 Lượng truy cập của các quốc gia trên thế giới	31
4.3.2 Lượng truy cập tại Việt Nam.....	32
4.3.3 Thống kê theo ISP về lượng truy cập ở Việt Nam	34
4.3.3 Độ phổ biến nội dung	35
CHƯƠNG 5 : TỔNG KẾT VÀ HƯỚNG PHÁT TRIỂN	36
5.1 Tổng kết.....	36
5.2 Hướng phát triển.....	37
TÀI LIỆU THAM KHẢO.....	38

DANH MỤC TỪ VIẾT TẮT

TỪ VIẾT TẮT	NỘI DUNG
API	Application Programming Interface
CDN	Content Delivery/Distribution Network
JDBC	Java Database Connectivity
JVM	Java Virtual Machine
ISP	Internet Service Provider
Mlib	Machine Learning Library (MLlib)
ODBC	Open Database Connectivity
QoS	Quality of Service
RDD	Resilient Distributed Datasets
TCP	Transmission Control Protocol

DANH MỤC HÌNH ẢNH

Hình 2.1 Hình ảnh về mạng phân phối nội dung CDN	5
Hình 2.2 Logo của Apache Spark.....	8
Hình 2.3 Hình ảnh thành phần của Apache Spark.....	9
Hình 2.4 Cấu trúc thành phần của Apache Spark.....	12
Hình 2.5 Logo của Apache Zeppelin.....	14
Hình 2.6 Giao diện trên web của Zeppelin.....	15
Hình 2.7 Logo Tableau Software	16
Hình 2.8 Kết nối dữ liệu đầu vào của Tableau Desktop.....	17
Hình 2.9 Giao diện thực hiện trực quan hóa.....	17
Hình 3.1 Khử lỗi từ tập dữ liệu ban đầu.....	18
Hình 3.2 Các lỗi trên log file	19
Hình 3.3 Lọc lại các thông tin sử dụng.....	20
Hình 3.4 Lọc lại lượng truy cập thật sự.....	20
Hình 3.5 Mô hình lấy vị trí từ IP.....	21
Hình 3.6 Tra cứu thông tin ip trên một trang web.....	22
Hình 3.7 Giao diện sau khi kết nối với dữ liệu.....	24
Hình 3.8 Các quốc gia truy cập đến Server.....	25
Hình 3.9 Các thành phố ở Việt Nam truy cập đến Server.....	26
Hình 4.1 Thống kê lượng truy cập của từng nội dung	29
Hình 4.2 Thông tin về vị trí và ISP của từng IP	30
Hình 4.3 Số lượng truy cập của các quốc gia trên thế giới	31
Hình 4.4 Tỷ lệ phần trăm lượng truy cập các nước.....	31
Hình 4.5 Số lượng truy cập của các thành phố trong nước Việt Nam	32
Hình 4.6 Tỷ lệ phần trăm lượng truy cập của 2 thành phố lớn	33
Hình 4.7 Thống kê lượng truy cập của các ISP ở Hà Nội và Hồ Chí Minh	34
Hình 4.8 Độ phổ biến nội dung trên web trường trong một năm	35

CHƯƠNG 1 : TỔNG QUAN

1.1 Giới thiệu đề tài

Hiện nay chúng ta đang thấy sự bùng nổ mạnh mẽ của Internet và các nội dung trên mạng. Ngày càng tạo ra nhiều nội dung mới và lượng truy cập quá nhiều sẽ gây ra tắc nghẽn cho các hoạt động trên mạng. Để hạn chế tình trạng này, các kỹ thuật giúp cải thiện việc truy cập và tránh tắc nghẽn đã thể hiện vai trò quan trọng của nó. Các kỹ thuật như định cỡ mạng, QoS, v.v... Cũng như các cách trên thì kỹ thuật để giúp mạng hoạt động tốt hơn là giải pháp mạng phân phối nội dung (Content Delivery Networks-CDN) nhằm tránh sự tắc nghẽn của các hoạt động trên mạng. Các Client kết nối với Server nếu đi qua vùng bị tắc nghẽn sẽ bị trì trệ. Thay vào đó các Server sẽ liên kết với nhau để trả về nội dung theo yêu cầu của người dùng. Như vậy sẽ tránh được sự ùn tắc trong mạng và có thể khả năng tốc độ đường truyền sẽ cao hơn. Lưu trữ ở các Cache Server như thế nào cũng là vấn đề được chú ý đến và tính toán kỹ hơn. Chúng ta nên lưu lại những nội dung mà mọi người quan tâm, được truy cập nhiều sẽ giúp các hoạt động trên mạng tốt hơn. Phân bố nội dung trên mạng cho phù hợp với các Server để người dùng truy cập thông tin nhanh nhất mà ít bị tắc nghẽn. Ở các Server sẽ có lưu lại các hoạt động của các người dùng truy cập đến. Qua đó tìm hiểu cách thức phân tích và tìm ra các nội dung nên được lưu lại ở các Cache Server. Từ việc muốn tìm ra nội dung trên, chúng ta nên xây dựng các tập dữ liệu yêu cầu nội dung từ người dùng cho việc tính toán ở các Cache Server. Rõ ràng nếu lưu lại được các nội dung phù hợp ở các Cache Server thì sẽ giảm bớt đi sự tắc nghẽn cho hoạt động trong mạng[1].

1.2 Lý do lựa chọn đề tài

Để hỗ trợ cho việc tính toán những nội dung ở Cache Server. Ở các Cache Server cần biết được nội dung truy cập của người dùng thay đổi như thế nào để cho mạng CDN được hoạt động tốt. Do đó tôi thực hiện đề tài xây dựng ra các tập dữ liệu có ý nghĩa cho các Cache Server.

1.3 Mục tiêu đề tài

- Xây dựng tập dữ liệu thống kê số truy cập của các nội dung trên mạng
- Xây dựng tập dữ liệu về vị trí của các người dùng
- Phân tích và biểu thị ý nghĩa từ các tập dữ liệu

1.4 Đối tượng và phạm vi nghiên cứu

Mục đích nghiên cứu của luận văn là xây dựng các tập dữ liệu hữu ích cho giải pháp mạng CDN. Tôi sử dụng log file ở Server của trường để phân tích và làm dữ liệu cho bài luận văn. Log file của website trường cụ thể là: www.hcmut.edu.vn làm đối tượng của phạm vi nghiên cứu. Với phạm vi nghiên cứu là áp dụng cho mạng CDN.

1.5 Phương pháp nghiên cứu và hướng tiếp cận

Từ dữ liệu ban đầu là log file có kích thước lớn, tôi tiến hành tìm hiểu và phân tích nó. Tôi sử dụng Apache Spark để phân tích log file đó thành các tập dữ liệu có ý nghĩa cho Cache Server. Tôi có trực quan hóa các dữ liệu mới được tạo ra để dễ dàng hiểu ý nghĩa dữ liệu bằng Tableau.

1.6 Cấu trúc luận văn

Luận văn được chia thành 5 chương, nội dung của mỗi chương như sau:

Chương 1: Tổng quan

Chương đầu giới thiệu ngắn gọn về sự phát triển của CDN. Bên cạnh đó, chương này tôi cũng nêu về lý do, mục tiêu, đối tượng và phạm vi nghiên cứu đề tài. Giới thiệu về cách thức nghiên cứu và hướng tiếp cận thực hiện cho đề tài.

Chương 2: Kiến thức nền tảng

Chương này cung cấp các kiến thức nền tảng để thực hiện ở phần sau của luận văn. Các phần của chương này nói về các hệ thống mạng phân phối nội dung(CDN), cấu trúc của log file ở trường cụ thể hơn là access log, cũng có thêm các công cụ hỗ trợ thực hiện như : Apache Spark, Zeppelin, Tableau.

Chương 3: Hiện thực

Chương này tôi sẽ nêu ra cách thức tìm hiểu và phân tích log với việc sử dụng Apache Spark và trực quan hóa bằng Tableau. Phân tích tạo ra các tập dữ liệu mới. Sử dụng các tập dữ liệu để hiển thị lên các dạng biểu đồ với một cách nhìn trực quan hơn về các khía cạnh có trong log.

Chương 4: Đánh giá

Chương này tôi sẽ thực hiện đánh giá lại các kết quả sau khi phân tích ra các tập dữ liệu bằng Spark và hiển thị bằng công cụ Tableau.

Chương 5: Tổng kết và hướng phát triển

Tổng kết các kết quả đạt được qua quá trình phân tích và trực quan hóa. Ứng dụng các tập dữ liệu cho bài toán CDN như thế nào. Các thách thức, khó khăn trong quá trình xử lý từ đầu.

Có thể vận dụng thêm các trường trong log để tìm tiếp các kết quả mới cho các đề tài sau này. Sử dụng trường thời gian và kết hợp tìm hiểu rõ hơn về các mối liên quan với nhau trong phần nội dung. Xây dựng thêm các tập dữ liệu có ích khác phục vụ cho bài toán CDN.

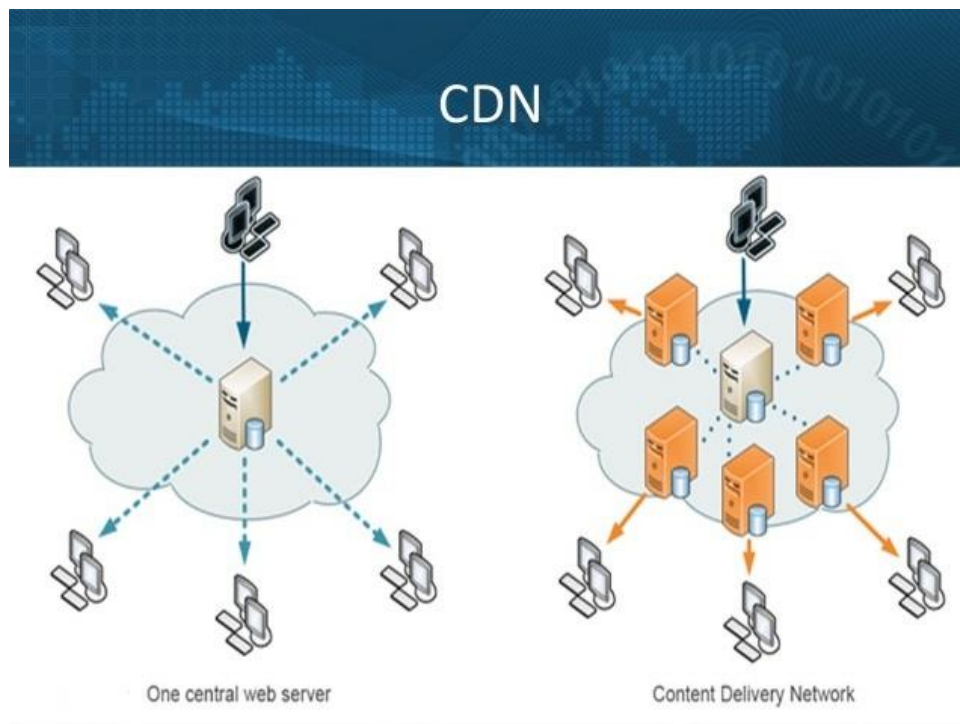
CHƯƠNG 2 : KIẾN THỨC NỀN TẢNG

Chương này giới thiệu sơ bộ về các kiến thức nền tảng phục vụ cho việc thực hiện các bước ở phía sau. Phần đầu tiên giới thiệu về hệ thống mạng phân phối nội dung. Kế tiếp nói về nội dung log file sẽ đem phân tích cụ thể về các thành phần của log. Ở các phần tiếp theo sẽ giới thiệu về các phần mềm hỗ trợ cho phân tích đề tài: Apache Spark, Apache Zeppelin, Tableau.

2.1 Hệ thống mạng phân phối nội dung

Nhằm cải thiện hoạt động của người sử dụng, tránh đi hiện tượng “thất cổ chai” khi các Client kết nối đến các Server nên giải pháp mạng phân phối nội dung đã được triển khai trong hạ tầng hệ thống mạng internet. CDN phục vụ một phần lớn nội dung Internet ngày nay, bao gồm các đối tượng web (văn bản, đồ họa và tập lệnh), các đối tượng tải xuống (media file, phần mềm, tài liệu), ứng dụng, phương tiện phát trực tiếp. Các nhà khai thác CDN phân phối các nội dung của các nhà chủ sở hữu nội dung cho các người dùng.

Các Cache Server thường được triển khai ở nhiều vị trí hình thành nên một mạng lưới. Lợi ích bao gồm giảm chi phí băng thông, cải thiện thời gian tải trang hoặc tăng tính sẵn có của nội dung trên toàn cầu. Số lượng các Cache Server và Server tạo nên các mạng phân phối nội dung khác nhau, tùy thuộc vào kiến trúc, một số mạng đạt đến hàng ngàn Cache Server với hàng chục nghìn Server.



Hình 2.1 Hình ảnh về mạng phân phối nội dung CDN [2]

Các yêu cầu về nội dung thường được định hướng bởi các thuật toán cho các Cache Server theo hướng tối ưu. Tính toán khoảng cách giữa các Cache Server với nhau, giữa các Cache Server với Server gốc sao cho phù hợp. Phân tích các nội dung phù hợp để lưu lại ở các Cache Server một cách hiệu quả nhất. Các mạng lưới CDN có quy mô từ lớn đến nhỏ khác nhau. Thực hiện trên các châu lục, các quốc gia với quốc gia, các thành phố với nhau, v.v...

- Cách hoạt động của mạng CDN:

Thay vì các người dùng truy cập đến nội dung họ mong muốn được lưu trực tiếp trên Server gốc thì người dùng có thể sẽ truy cập đến các Cache Server có chứa nội dung đó ở gần nhất có thể. Nếu như nội dung đó không có ở Cache Server thì nó sẽ thực hiện lấy từ Server gốc về. Khi các nội dung mới được phản hồi về thì các Cache Server sẽ có thể lưu lại các nội dung mới đó để xuất ra thông tin nhanh nhất cho người dùng khác có truy cập đến.

- Các lợi ích của hệ thống mạng phân phối nội dung:

+ Lợi ích cho các nhà thực hiện CDN:

- Giảm tải cho hệ thống máy chủ vận hành chính

- Các file tĩnh của website sẽ được bố trí trên các Cache Server của mạng CDN giúp cho các Server gốc giảm tải trong quá trình vận hành hệ thống.

- Dùng cơ chế xác định vị trí Server gần nhất so với Client giúp cho việc truyền tải dữ liệu nhanh hơn giúp website bạn có tốc độ truy xuất nhanh hơn dù ở bất kỳ nơi đâu

- Tương thích với các mã nguồn thông dụng: wordpress, joomla, v.v...

- Tiết kiệm chi phí đầu tư nâng cấp cho hệ thống Server hiện tại

- Thay vì phải trang bị nhiều Server đặt tại nhiều nơi bạn có thể dùng dịch vụ CDN để tiết kiệm chi phí đầu tư thiết bị và tập trung vào công việc kinh doanh của bạn và mang lại hiệu quả cao hơn

- Giúp tăng thêm đối tượng truy cập ở nhiều nơi trên thế giới

- Cùng với việc mở rộng phạm vi truy cập sẽ giúp bạn tìm kiếm được các khách hàng tiềm năng và mở rộng hoạt động kinh doanh của bạn sang các khu vực và quốc gia khác

+ Lợi ích cho các người dùng:

- Tiết kiệm băng thông đáng kể đối với các dữ liệu tĩnh (hình ảnh, css, javascript)

- Tăng tốc độ truy cập website, load nội dung nhanh, giảm thiểu độ trễ, giật hình khi truy cập và xem các trang website phân phối nội dung như: Movies, Video clip, v.v...

- Cho phép người dùng Internet có thể tương tác nhanh chóng, gia tăng sự hài lòng khi tiếp cận website trong thời gian thực

+ Các đối tượng cần dùng CDN :

- Các website có lượng truy cập lớn, website chứa nhiều nội dung tĩnh (hình ảnh, css, javascript). Sử dụng CDN sẽ tiết kiệm hơn là dùng Server riêng cho các website tầm trung

- Máy chủ gốc đặt ở xa đối tượng người dùng hoặc cần phân phối nội dung với chất lượng tốt nhất trên toàn thế giới

- Các nhà cung cấp dịch vụ Media, các doanh nghiệp, cá nhân sử dụng hạ tầng CDN để phân phối nội dung (Movies, Video clip,...) trên Internet nhằm quảng bá

và kinh doanh các sản phẩm dịch vụ do doanh nghiệp, cá nhân cung cấp tới người dùng cuối

- Đối với các đài truyền hình, đơn vị có thể phát triển kênh truyền hình cung cấp cho người xem thông qua mạng Internet trên trang web của chính đài truyền hình

2.2 Cấu trúc log file

Hoạt động của các công ty, trường học và các Server khác xảy ra liên tục ngày này qua ngày khác. Tại các Server luôn phải ghi lại các hoạt động xảy ra ở Server của mình. Những hành động tác động đến nội dung, truy cập đến trang web, các lỗi xảy ra,... sẽ được Server lưu trữ lại. Các file chứa các nội dung như vậy là gọi là log. Có nhiều loại log ở từng Server khác nhau như: access log, error log, sql log, system log,...

Mỗi loại log sẽ có các cấu trúc khác tùy vào người cấu hình bên Server. Dưới đây là một cấu trúc mẫu của access log mà tôi xem xét trong đề tài này:

```
113.162.189.51      -      -      [08/Nov/2016:12:00:29      +0700]      "GET
/includes/js/jquery/jquery-1.6.4.min.js      HTTP/1.1"      200      91669
"http://www.hcmut.edu.vn/vi" "Mozilla/5.0 (Linux; Android 5.1.1; D2502
Build/19.4.A.0.182; wv) AppleWebKit/537.36 (KHTML, like Gecko)
Version/4.0 Chrome/46.0.2490.76 Mobile Safari/537.36"
```

Gồm các thành phần :

+ Địa chỉ IP (Ip Address): 113.162.189.51 - đây là một địa chỉ IP public của người dùng truy cập đến trang web

+ Tên truy cập (Username) : - - sẽ có liên quan khi truy cập vào các trang web có bảo vệ bằng mật khẩu

+ Thời gian (Timestamp): [08/Nov/2016:12:00:29 +0700] trường này chứa thời gian truy cập và timezone của web Server

+ Yêu cầu truy cập (Access request): "GET /includes/js/jquery/jquery-1.6.4.min.js HTTP/1.1" sử dụng GET request là hiển thị trang web cho file "/includes/js/jquery/jquery-1.6.4.min.js" bằng giao thức "HTTP/1.1"

+ Mã trạng thái kết quả (Result status code): 200 - mã trạng thái kết quả 200 là thành công. Nếu đường dẫn không tồn tại thì mã là 404. Còn nhiều mã lỗi khác nhau.

+ Dung lượng truyền đi (bytes transferred): 91669 là dung lượng byte chuyển đi.

+ Đường dẫn tham khảo (Referrer URL): <http://www.hcmut.edu.vn/vi> - đây là trang web mà khách truy cập đến. Nó có thể là đường dẫn có liên kết từ các trang web khác hoặc nó chỉ là địa chỉ mà họ gõ vào để truy cập. Có thể là dấu "-" vì :

- Nó được nhập trực tiếp vào browser hoặc sử dụng bookmark
- Có thể là đường link dẫn từ bên ngoài ví dụ như đường dẫn trong mail hay điện thoại
- Có thể chính browser của họ không gửi đường dẫn tham khảo

+ User agent : "Mozilla/5.0 (Linux; Android 5.1.1; D2502 Build/19.4.A.0.182; wv) AppleWebKit/537.36 (KHTML, like Gecko) Version/4.0 Chrome/46.0.2490.76 Mobile Safari/537.36" Giới thiệu về các browser dùng để truy cập và một số phiên bản.

2.3 Apache Spark



Hình 2.2 Logo của Apache Spark [3]

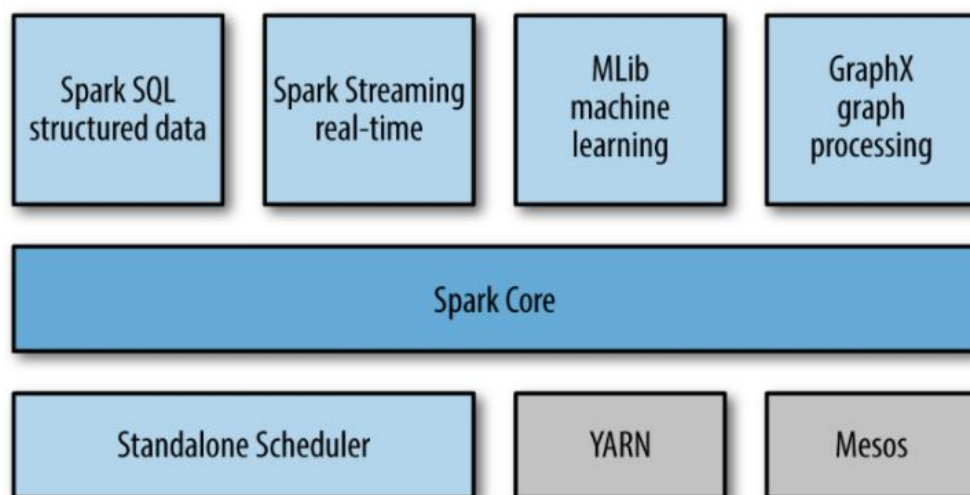
Dữ liệu ngày càng có nhiều nên việc lưu trữ và xử lý dữ liệu cũng là vấn đề được nhiều công ty quan tâm đến. Xử lý các dữ liệu có dung lượng lớn sẽ rất lâu gây tốn nhiều thời gian. Trong đó có những giải pháp đã được đưa ra để phục vụ cho việc khai thác và xử lý dữ liệu như Apache Spark.

Spark được phát triển sơ khởi vào năm 2009 bởi AMPLab tại đại học California. Sau này, Spark đã được chuyển giao cho Apache Software Foundation vào năm 2013 và được phát triển cho đến nay [4].

Apache Spark là một hệ thống mã nguồn mở cho phép thực hiện tính toán trên các hệ thống máy tính hiệu năng cao nhằm tạo ra khả năng phân tích dữ liệu nhanh với 2 tiêu chí: nhanh về cả lúc chạy và nhanh cả lúc ghi dữ liệu. Để chạy chương trình nhanh hơn, Spark cung cấp một mô hình thực thi cho phép tối ưu các tính toán đồ thị một cách tùy ý (optimize arbitrary operator graphs) và hỗ trợ tính toán tại bộ nhớ chính (Ram) giúp việc query dữ liệu nhanh hơn công nghệ tính toán dựa trên bộ nhớ ngoài (disk-based) như là Hadoop [5].

Spark có kết hợp với nhiều ngôn ngữ lập trình để thuận tiện cho người sử dụng như: Scala, Java, Python. Spark ban đầu được phát triển cho hai loại thuật toán: thuật toán lặp (iterative algorithms) thường được sử dụng trong học máy và thuật toán tương tác trong khai phá dữ liệu. Trong các trường hợp trên, Spark có thể chạy nhanh gấp 100 lần so với Hadoop MapReduce [3].

+ Các thành phần của Apache Spark



Hình 2.3 Thành phần của Apache Spark [6]

Mọi ứng dụng Spark đều bao gồm một chương trình điều khiển chạy chức năng chính của người dùng và thực hiện các hoạt động song song khác nhau trên một hệ thống máy tính hiệu năng cao. Spark cung cấp một tập dữ liệu phân phối đàn hồi (resilient distributed dataset- RDD), là tập hợp các phần tử được phân đoạn trên các nút tính toán của hệ thống máy tính hiệu năng cao có thể hoạt động song song. Các thành phần của Spark được nêu lên ở hình 2.3.

+ **Spark SQL**: là mô-đun của Apache Spark để làm việc với dữ liệu có cấu trúc. Spark SQL cho phép truy vấn dữ liệu cấu trúc qua các câu lệnh SQL. Spark SQL có thể thao tác với nhiều nguồn dữ liệu như Hive tables, Parquet, và JSON.

Các tính năng của Spark SQL :

- Tích hợp (Integrated) - Kết hợp liền mạch các truy vấn SQL với các chương trình Spark. Spark SQL cho phép bạn truy vấn dữ liệu có cấu trúc dưới dạng bộ dữ liệu phân tán (RDD) trong Spark, với các API được tích hợp trong Python, Scala và Java. Việc tích hợp chặt chẽ này giúp dễ dàng chạy các truy vấn SQL cùng với các thuật toán phân tích phức tạp.

- Truy cập dữ liệu thống nhất (Unified Data Access) - Tải và truy vấn dữ liệu từ nhiều nguồn khác nhau. Lược đồ-RDD cung cấp một giao diện duy nhất để làm việc hiệu quả với dữ liệu có cấu trúc, bao gồm các bảng Apache Hive và các tệp JSON.

- Khả năng tương thích Hive (Hive Compatibility) - Chạy truy vấn Hive chưa sửa đổi trên kho hiện có. Spark SQL tái sử dụng lối vào Hive và MetaStore, cho bạn khả năng tương thích đầy đủ với dữ liệu Hive, truy vấn và UDF hiện có. Đơn giản chỉ cần cài đặt nó cùng với Hive.

- Kết nối chuẩn - Kết nối thông qua JDBC hoặc ODBC.

+ **Spark Streaming**: là phần mở rộng của API Spark core cho phép xử lý luồng có khả năng mở rộng, thông lượng cao, xử lý luồng có khả năng chịu lỗi của luồng dữ liệu trực tiếp. Dữ liệu có thể được nhập từ nhiều nguồn như Kafka, Flume, Kinesis hoặc TCP socket, và có thể được xử lý bằng cách sử dụng các thuật toán phức tạp được thể hiện với các hàm bậc cao như map, reduce, join và window. Cuối cùng, dữ liệu đã xử lý có thể được đẩy ra ngoài hệ thống thành tệp, cơ sở dữ liệu và trang tổng quan trực tiếp. Trên thực tế, bạn có thể áp dụng các thuật toán xử lý đồ họa và học máy của Spark trên luồng dữ liệu.

+ **Mlib machine learning**:

MLlib chứa nhiều thuật toán và tiện ích. Một số thuật toán Machine Learning bao gồm:

- Phân loại: hồi quy logistic, naive Bayes,...
- Hồi quy: hồi quy tuyến tính tổng quát, hồi quy sống còn(survival regression), ...
- Cây quyết định, cây có độ dốc

Các tập hợp thường xuyên, quy tắc kết hợp và khai thác mẫu tuần tự.

Các tiện ích dòng công việc ML(Machine Learning) [3] bao gồm:

- Chuyển đổi tính năng: chuẩn hóa, băm, ...
- ML Pipeline construction
- Đánh giá mô hình và điều chỉnh tham số
- ML persistence: lưu và tải các mô hình và đường ống

Mlib là thư viện học máy có khả năng mở rộng của Apache Spark.

+ **GraphX graph processing:**

Việc phân tích dữ liệu lớn của Apache Spark ngày càng rộng và trở nên dễ dàng hơn, Spark triển khai các thuật toán hữu ích cho việc khai phá dữ liệu, phân tích dữ liệu, học máy, các thuật toán trên đồ thị. Sử dụng khả năng chạy trên hệ thống máy tính hiệu năng cao, Spark giải quyết các vấn đề như khả năng chịu lỗi và cung cấp API đơn giản để thực hiện tính toán song song.

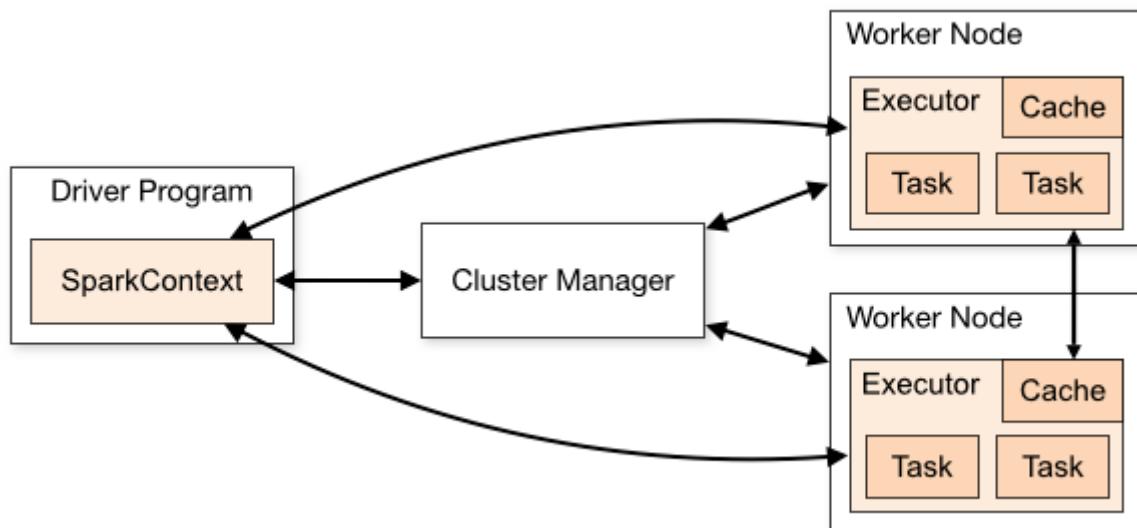
GraphX là API của Apache Spark cho đồ thị và tính toán song song đồ thị. GraphX đơn giản hóa các nhiệm vụ phân tích biểu đồ. Có rất nhiều thuật toán trong lý thuyết đồ thị được sử dụng trong phân tích dữ liệu.

+ **Spark Core:**

Spark Core là nền tảng của toàn bộ dự án. Nó cung cấp các chức năng phân phối nhiệm vụ, lịch trình và các chức năng nhập xuất cơ bản, được hiển thị thông qua giao diện lập trình ứng dụng. Mô hình lập trình chức năng hay bậc cao: các “Driver” gọi các hoạt động song song như map, filter, reduce trên RDD bằng cách chuyển một hàm tới Spark sau đó lên lịch thực hiện hàm song song trên các hệ thống máy tính hiệu năng cao. Các hoạt động kết nối với nhau từ RDD đầu tạo ra các RDD mới là không thay đổi và tính toán lười. RDD có thể thực hiện trên các đối tượng như Python, Scala, Java.

Spark có thể chạy trên nhiều loại công cụ quản lý hệ thống máy tính hiệu năng cao như Hadoop YARN, Apache Mesos hoặc trên chính công cụ hệ thống máy tính hiệu năng cao được cung cấp bởi Spark được gọi là Standalone Scheduler.

+ Cấu trúc spark



Hình 2.4 Cấu trúc thành phần của Apache Spark [8]

Các ứng dụng Spark chạy như các bộ quy trình độc lập trên một hệ thống máy tính hiệu năng cao, được phối hợp bởi đối tượng SparkContext trong chương trình chính của bạn (được gọi là chương trình trình điều khiển – driver program).

Cụ thể, để chạy trên một hệ thống máy tính hiệu năng cao, SparkContext có thể kết nối với một số loại công cụ quản lý hệ thống máy tính hiệu năng cao, phân bổ tài nguyên trên các ứng dụng. Khi đã kết nối, Spark sẽ thu nhận các thực thi trên các nút trong hệ thống máy tính hiệu năng cao, đó là các quá trình chạy các tính toán và lưu trữ dữ liệu cho ứng dụng. Tiếp theo, nó sẽ gửi mã ứng dụng (được xác định bởi các tệp JAR hoặc Python được chuyển đến SparkContext) cho các trình thực thi. Cuối cùng, SparkContext gửi nhiệm vụ đến các trình thực thi để chạy.

Mỗi ứng dụng sẽ có các quy trình thực thi riêng của nó, nó sẽ ở lại trong suốt thời gian của toàn bộ ứng dụng và chạy các tác vụ trong nhiều luồng. Điều này có lợi ích của việc tách biệt các ứng dụng với nhau, trên cả hai mặt lịch trình (mỗi trình điều khiển lên lịch các nhiệm vụ riêng của mình) và bên thi hành (các nhiệm vụ từ các ứng dụng khác nhau chạy trong các JVM khác nhau).

Hệ thống hiện hỗ trợ các công cụ quản lý hệ thống máy tính hiệu năng cao:

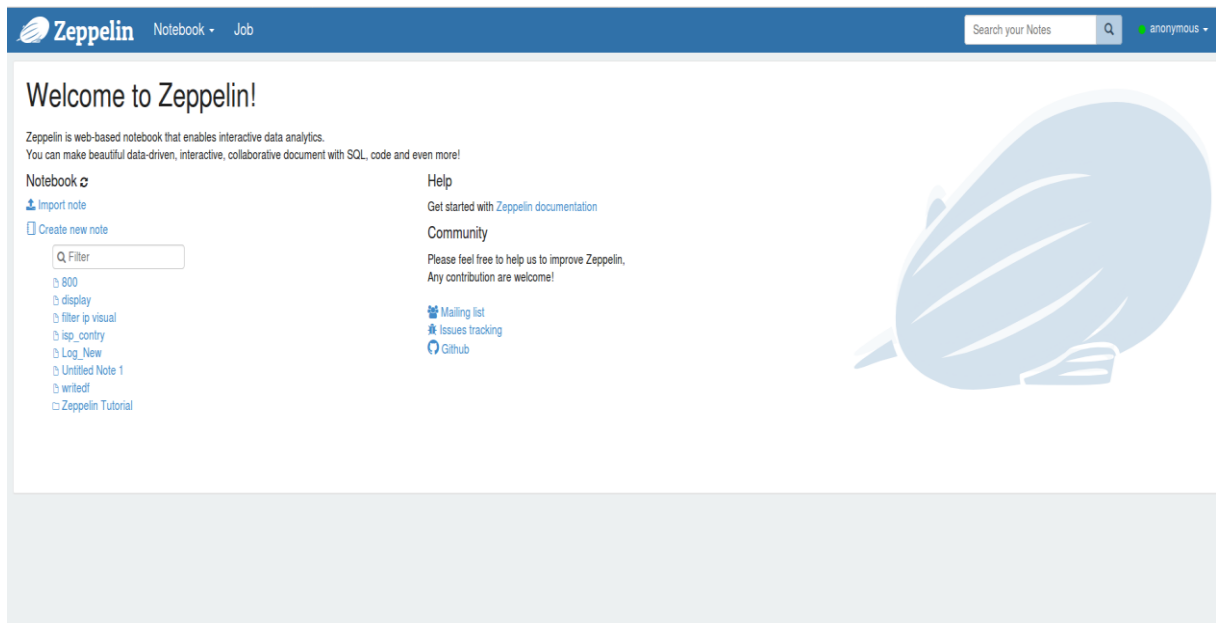
- + Standalone - một trình quản lý cụm đơn giản đi kèm với Spark giúp dễ dàng thiết lập một cụm.
- + Apache Mesos - một trình quản lý cụm chung có thể chạy các ứng dụng dịch vụ và MapReduce của Hadoop.
- + Hadoop YARN - trình quản lý tài nguyên trong Hadoop 2.
- + Kubernetes - một hệ thống mã nguồn mở để tự động hóa việc triển khai, mở rộng quy mô và quản lý các ứng dụng được container.

2.4 Apache Zeppelin



Hình 2.5 Logo của Apache Zeppelin [9]

Apache Zeppelin là web-based notebook dựa trên web cho phép phân tích dữ liệu, tương tác với các hình ảnh được tích hợp sẵn. Nó hỗ trợ nhiều ngôn ngữ với một khung thông dịch. Hiện tại, Zeppelin hỗ trợ các khung thông dịch như Spark, Markdown, Shell, Hive, Phoenix, Tajo, Flink, Ignite, Lens, HBase, Cassandra, Elasticsearch, Geode, PostgreSQL và Hawq. Nó có thể được sử dụng để nhập dữ liệu, khám phá, phân tích dữ liệu và trực quan hóa bằng cách sử dụng notebooks tương tự IPython Notebooks. Đặc biệt, Apache Zeppelin cung cấp tích hợp Apache Spark tích hợp. Bạn không cần phải xây dựng một module, plugin hoặc thư viện riêng cho nó.



Hình 2.6 Giao diện trên web của Zeppelin

Giao diện khi chúng ta cài đặt thành công. Chúng ta có thể tạo các note mới.

Apache Zeppelin có tích hợp Spark cung cấp:

- Tự động SparkContext và SQLContext injection
- Tải xuống phụ thuộc vào thời gian chạy của jar từ hệ thống tệp cục bộ hoặc kho lưu trữ maven.
- Hủy công việc và hiển thị tiến trình của nó

Apache Zeppelin có thể trực quan hóa dữ liệu dưới nhiều dạng đồ thì khác nhau như: biểu đồ cột, biểu đồ đường, biểu đồ tròn,... với các mục đích khác nhau. Các biểu đồ thì không giới hạn truy vấn SparkSQL. Các note trong Zeppelin có thể xuất ra lưu trữ bên ngoài thuận tiện cho việc sử dụng trên các máy tính khác nhau.

2.5 Tableau



Hình ảnh 2.7 Logo Tableau Software [10]

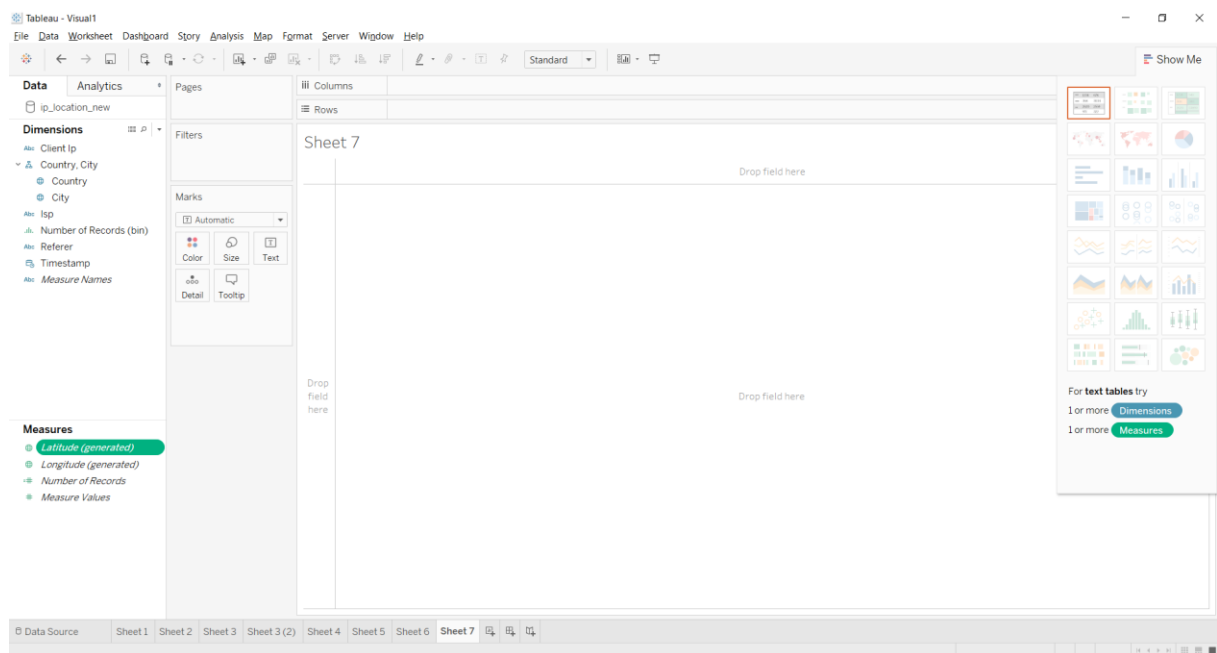
Tableau Software là một công ty phần mềm máy tính của Mỹ có trụ sở tại Seattle, WA, Hoa Kỳ. Nó tạo ra các sản phẩm trực quan dữ liệu tương tác tập trung vào BI(business intelligence). Công ty được thành lập tại Khoa Khoa học Máy tính của Đại học Stanford từ năm 1997 đến 2002. Tableau là một phần mềm hỗ trợ phân tích và trực quan hóa dữ liệu một cách nhanh chóng.

Tableau sử dụng ngôn ngữ truy vấn trực quan VizQL. VizQL là một ngôn ngữ truy vấn trực quan dịch các hoạt động kéo, thả vào các truy vấn dữ liệu và sau đó thể hiện dữ liệu đó một cách trực quan. VizQL mang lại lợi ích đáng kể về khả năng xem và hiểu dữ liệu của mọi người truy vấn và phân tích.

Tableau có nhiều sản phẩm như: Tableau Desktop, Tableau Server, Tableau Online, Table Prep. Trong đề tài luận văn này, tôi sử dụng Tableau Desktop. Tableau Desktop là ứng dụng trực quan hóa dữ liệu cho bất kỳ dữ liệu có cấu trúc nào để tạo ra báo cáo, biểu đồ có tương tác cao. Sau khi cài đặt bạn có thể liên kết với các dữ liệu và hiển thị thông tin theo một góc độ đồ họa.



Hình 2.8 Kết nối dữ liệu đầu vào của Tableau Desktop



Hình 2.9 Giao diện thực hiện trực quan hóa

Khi mở giao diện phần mềm thì chúng ta sẽ thêm dữ liệu vào để tiến hành phân tích như ở hình 2.8. Sau khi chèn dữ liệu vào chúng ta có thể thấy dữ liệu được phân ra các cột. Có các thuộc tính trong dữ liệu thêm vào. Bên phải có “Show Me” cho phép xem các loại biểu đồ sẽ hiển thị như ở hình 2.9. Kéo thả các trường của các cột vào dữ liệu để tạo ra các biểu đồ theo mục đích sử dụng.

Lợi ích của Tableau:

- + Xử lý kết quả nhanh chóng để có các thông tin hữu ích
- + Dễ sử dụng cho tất cả mọi người
- + Có thể chèn dữ liệu đầu vào ở nhiều dạng khác nhau:
 - Hệ quản trị cơ sở dữ liệu (Oracle, MSSQL,...)
 - Microsoft Access
 - Microsoft Excel
 - CSV/Flat Files

CHƯƠNG 3 : HIỆN THỰC

Chương này tôi sẽ nêu ra cách thức tìm hiểu và phân tích log sử dụng Apache Spark và trực quan hóa bằng Tableau. Cách thức thực hiện tạo ra các tập dữ liệu từ log file ban đầu thu được từ Server. Cách chia nhỏ và phân tích tạo ra các tập dữ liệu có ích hơn cho sau này. Sử dụng các tập dữ liệu để hiển thị lên các dạng biểu đồ cho một cách nhìn trực quan hơn về các khía cạnh có trong log. Phần mềm Tableau cung cấp cho ta nhiều cách hiển thị để cho phù hợp với từng mục đích.

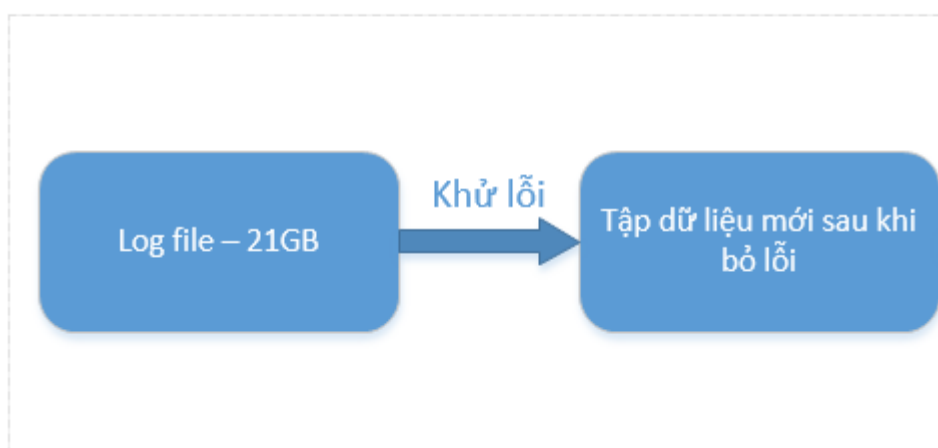
3.1 Phân tích log file sử dụng Spark

Như đã nói ở trên, tôi sẽ phân tích access log của Web Server: www.hcmut.edu.vn.

Log trên ghi lại các hoạt động truy cập của người dùng đến Server. Trong đề tài luận văn này tôi sẽ sử dụng công cụ hỗ trợ là Apache Spark. Sau khi cài đặt Spark cho Ubuntu thì chúng ta có thể sử dụng lệnh trên Terminal là spark-shell với ngôn ngữ Scala. Bên cạnh đó cài đặt Apache Zeppelin sử dụng trên các trình duyệt để thực hiện kết hợp với Spark.

3.1.1 Loại bỏ các lỗi trong tập dữ liệu ban đầu

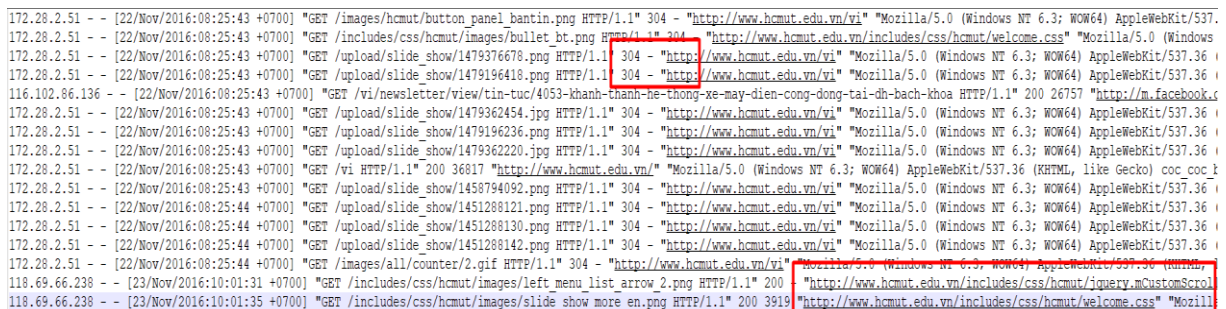
Log file hiện tại có dung lượng khá lớn với khoảng 21GB. Với dung lượng log file có kích thước khá lớn gây khó khăn cho việc phân tích toàn bộ file mà không có xảy ra lỗi gì.



Hình 3.1 Khử lỗi từ tập dữ liệu ban đầu

Log có kích thước lớn và nhìn vào bên trong như một kho dữ liệu mơ hồ và hầu như không có ngữ nghĩa, gây khó hiểu và không biết phải bắt đầu làm từ đâu để có thể phân tích ra. Sau đó, tôi thực hiện chia nhỏ file ra để có thể nghiên cứu chi tiết về từng dòng được ghi trong file. Qua đó có thể biết được cấu trúc bên trong như thế nào. Tôi cố gắng tìm hiểu rõ hơn về từng phần trong cấu trúc đó. Từng dòng trong file log được chia ra các phần nhỏ và chi tiết hơn được nêu ở Mục 2.2. Qua các phần chi tiết như thế, tôi sẽ tạo ra một lớp (class) chứa các trường trong log. Để có thể chuyển các thành phần trong log như thế thì cần phải chú ý hơn về giữa các thành phần, chúng được phân ra với nhau bởi một dấu cách. Như thế chúng ta chia ra theo dấu cách thì sẽ được các phần nhỏ hơn để đưa vào các trường có trong lớp mới đã được tạo ra. Chúng ta có thể tạo thành một tập dữ liệu mới được định dạng thành các trường cụ thể để cho việc truy xuất dữ liệu phục vụ cho các phần tiếp theo.

Bên trong các dòng log có khá nhiều lỗi như: dung lượng chuyển file là dấu “-“; các dòng tham khảo đến từ nguồn nào thì thay vào đó là chuỗi kí tự không rõ hay cũng có thể là đường dẫn đến một file định dạng trang web,.....



```

172.28.2.51 - - [22/Nov/2016:08:25:43 +0700] "GET /images/hcmut/button_panel_bantin.png HTTP/1.1" 304 - "http://www.hcmut.edu.vn/vi" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.
172.28.2.51 - - [22/Nov/2016:08:25:43 +0700] "GET /includes/css/hcmut/images/bullet_bt.png HTTP/1.1" 304 - "http://www.hcmut.edu.vn/includes/css/hcmut/welcome.css" "Mozilla/5.0 (Windows
172.28.2.51 - - [22/Nov/2016:08:25:43 +0700] "GET /upload/slide_show/1479376678.png HTTP/1.1" 304 - "http://www.hcmut.edu.vn/vi" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36
172.28.2.51 - - [22/Nov/2016:08:25:43 +0700] "GET /upload/slide_show/1479196418.png HTTP/1.1" 304 - "http://www.hcmut.edu.vn/vi" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36
116.102.86.136 - - [22/Nov/2016:08:25:43 +0700] "GET /vi/newsletter/view/tin-tuc/4053-khanh-thanh-he-thong-xe-may-dien-cong-dong-tai-dh-bach-khoa HTTP/1.1" 200 26757 "http://m.facebook.c
172.28.2.51 - - [22/Nov/2016:08:25:43 +0700] "GET /upload/slide_show/1479362454.jpg HTTP/1.1" 304 - "http://www.hcmut.edu.vn/vi" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36
172.28.2.51 - - [22/Nov/2016:08:25:43 +0700] "GET /upload/slide_show/1479196236.png HTTP/1.1" 304 - "http://www.hcmut.edu.vn/vi" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36
172.28.2.51 - - [22/Nov/2016:08:25:43 +0700] "GET /upload/slide_show/1479362220.jpg HTTP/1.1" 304 - "http://www.hcmut.edu.vn/vi" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36
172.28.2.51 - - [22/Nov/2016:08:25:43 +0700] "GET /vi HTTP/1.1" 200 36817 "http://www.hcmut.edu.vn/vi" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) coc coc k
172.28.2.51 - - [22/Nov/2016:08:25:43 +0700] "GET /upload/slide_show/1458794092.jpg HTTP/1.1" 304 - "http://www.hcmut.edu.vn/vi" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36
172.28.2.51 - - [22/Nov/2016:08:25:44 +0700] "GET /upload/slide_show/1451288121.png HTTP/1.1" 304 - "http://www.hcmut.edu.vn/vi" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36
172.28.2.51 - - [22/Nov/2016:08:25:44 +0700] "GET /upload/slide_show/1451288130.png HTTP/1.1" 304 - "http://www.hcmut.edu.vn/vi" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36
172.28.2.51 - - [22/Nov/2016:08:25:44 +0700] "GET /upload/slide_show/1451288142.png HTTP/1.1" 304 - "http://www.hcmut.edu.vn/vi" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36
172.28.2.51 - - [22/Nov/2016:08:25:44 +0700] "GET /images/all/counter/2.gif HTTP/1.1" 304 - "http://www.hcmut.edu.vn/vi" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML,
118.69.66.238 - - [23/Nov/2016:10:01:31 +0700] "GET /includes/css/hcmut/images/left_menu_list_arrow_2.png HTTP/1.1" 200 "http://www.hcmut.edu.vn/includes/css/hcmut/jquery.mCustomScrol
118.69.66.238 - - [23/Nov/2016:10:01:35 +0700] "GET /includes/css/hcmut/images/slide_show_more_en.png HTTP/1.1" 200 3919 "http://www.hcmut.edu.vn/includes/css/hcmut/welcome.css" "Mozilla

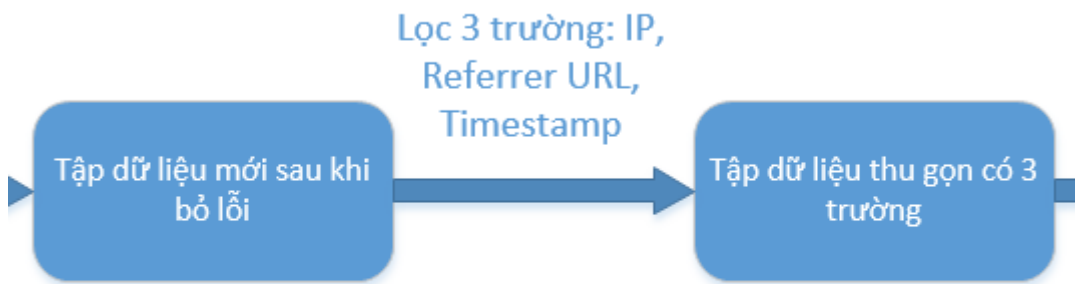
```

Hình 3.2 Các lỗi trên log file

Để có thể lọc bỏ các phần bị lỗi, tôi dùng đến các điều kiện so sánh ở các phần đó trước khi được thực hiện tiếp. Ngoài ra còn sử dụng các biểu thức chính quy để loại bỏ nhanh hơn về các đường dẫn là các file định dạng web. Khi phân tích từ log file thô ban đầu, tôi lọc bỏ đi các phần bị lỗi làm hữu ích hơn dữ liệu được tạo ra tránh gây nhiễu cho các bước làm tiếp. Trong phần thực hiện này tôi chú ý đến các phần chính như: địa chỉ IP, đường dẫn tham khảo, thời gian của từng dòng log,...

Qua bước xử lý này, chúng ta sẽ giảm bớt đi các dòng lỗi bên trong log.

3.1.2 Rút gọn tập dữ liệu



Hình 3.3 Lọc lại các thông tin sử dụng

Từ tập dữ liệu sau khi khử lỗi, tôi tiến hành loại bỏ bớt đi các trường không sử dụng cho các bước tiếp theo. Hình 3.2 nói về quá trình ở bước này tôi thực hiện lấy các trường là: IP, Referrer URL, Timestamp. Ở bước khử lỗi, tôi đã chuyển các trường trong log thành các phần trong class. Với các thành phần trong class, tôi chọn ra 3 thông tin tôi muốn như trên. Trong Spark, tôi sử dụng hàm *select* để chọn ra các cột từ tập dữ liệu ban đầu. Như vậy tập dữ liệu sẽ được rút gọn rất nhiều với các thông tin tôi muốn. Bước này làm tập dữ liệu nhỏ gọn hơn để làm tiếp các bước phía dưới.

3.1.3 Tập dữ liệu truy cập thật sự

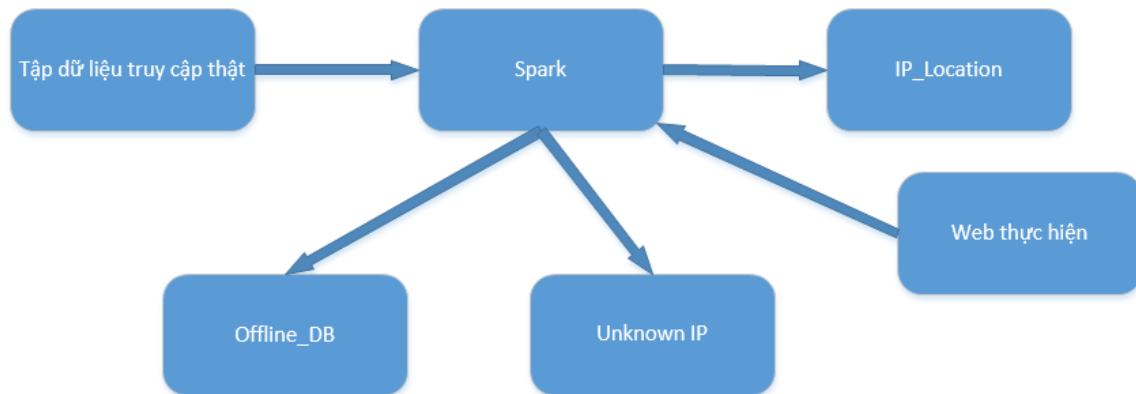


Hình 3.4 Lọc lại lượng truy cập thật sự

Sau khi chọn ra các trường để tiếp tục thực hiện, tôi bắt đầu xem xét lại các thông tin về lượng truy cập trong tập dữ liệu mới được tạo ra. Hình 3.3 nêu lên công việc tôi sẽ làm ở phần này. Trong log có rất nhiều dòng truy cập các thông tin giống nhau, cùng một IP và thời gian lại giống nhau thì sẽ làm cho dữ liệu của chúng ta gây nhiễu về lượng truy cập thật sự. Tôi sẽ kiểm tra nếu các cùng một IP mà truy cập với các nội dung khác nhau thì xem đó là những lần truy cập khác nhau. Nếu cùng IP và có thời gian truy cập liên tục thì phải xét lại. Qua tìm hiểu về thời gian truy cập giữa các lần với nhau thì 3 giây là phù hợp. Như vậy nếu cùng IP, cùng nội dung thì thời gian truy cập phải lớn hơn 3

giây mới tính là hai lần truy cập. Trong Spark, tôi sử dụng các phép so sánh để kiểm tra các điều kiện trên. Như vậy ta sẽ có tập dữ liệu về lượng truy cập thật sự.

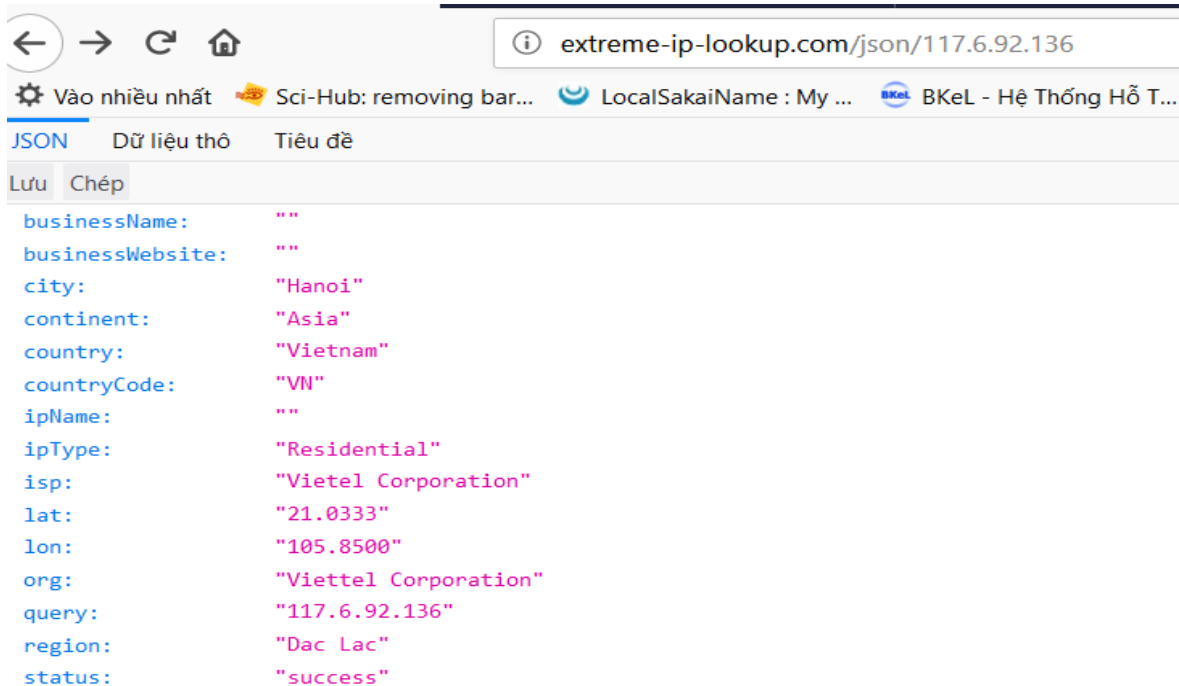
3.1.4 Xây dựng tập dữ liệu vị trí các IP



Hình 3.5 Mô hình lấy vị trí từ IP

Từ tập dữ liệu ở Mục 3.1.3, tôi muốn lấy các thông tin về vị trí của các IP theo mô hình ở hình 3.4. Từng IP chúng ta có thể tra cứu thêm về các thông tin vị trí như quốc gia, thành phố, tổ chức phân phối nào, vị trí trên bản đồ của thành phố đó. Tôi cố gắng tạo ra một tập dữ liệu mới chứa các thông tin vị trí đó.

Để có được các thông tin đó tôi phải tìm kiếm tra cứu trên mạng Internet. Có một số trang web hỗ trợ chúng ta về trả về các thông tin đó khi chỉ cần nhập IP đó vào. Nhưng tra cứu bằng tay, chủ công cho từng IP là việc khó xảy ra cho một dữ liệu log khá lớn.



Hình 3.6 Tra cứu thông tin IP trên một trang web

Chúng ta có một số trang web có các API hỗ trợ cho việc lấy thông tin về dưới dạng chuỗi. Tra cứu trên các trang web đó để lấy được các dữ liệu về máy. Nhưng cũng có giới hạn truy cập cho từng web khác nhau. Mỗi trang web trả về những thông tin dữ liệu ở dạng chuỗi có cấu trúc khác nhau. Bên cạnh đó có những trang trả về cho ta không đủ nội dung mong muốn, nó không xác định được vị trí IP đó vậy phải thực hiện trên nhiều API web khác nhau. Nhưng mỗi web cũng có giới hạn số lần lấy thông tin trong một thời gian nhất định theo giờ, ngày. Ngoài ra còn có một số web khi chúng ta là thành viên mới được phép tra cứu thông qua các key nhưng cũng có giới hạn như trên. Có một file web thực thi chứa các web mà chúng ta sẽ thực hiện tìm kiếm vị trí từ IP.

Từ chuỗi thông tin lấy được từ trên mạng chúng ta sẽ phân tích ra thành phần để lấy thông tin cần. Chuỗi trả về có cấu trúc dạng json.

Dữ liệu lấy về sẽ có rất nhiều thông tin có ích. Theo mô hình ở trên thì tôi có thể tạo thêm các tập dữ liệu cụ thể về IP và vị trí. Bên cạnh đó có nhiều IP không thể tra cứu được vị trí, tôi sẽ lưu lại các thông tin vào một file dữ liệu khác là Unknown IP. Còn Output của tôi là có ghép lại các thông tin từ log với số lần truy cập trên và kèm theo vị trí, ISP của IP đó. Tách chuỗi ra theo từng phần và lấy thông tin về quốc gia, thành phố

và tổ chức nhà phân phối mạng internet. Ghép từng phần đó vào file vừa rút trích kết hợp với IP để thành một tập dữ liệu mới. Như vậy ta tổng hợp được tập dữ liệu mới chứa các thông tin về IP, Referrer URL, Timestamp, quốc gia, thành phố, tổ chức của từng ISP. Và thêm 2 tập dữ liệu về thông tin vị trí của IP đó và các IP không xác định được.

3.1.5 Xây dựng tập dữ liệu thống kê truy cập

Tôi muốn tạo ra tập chứa các nội dung với số lượng truy cập. Mỗi người dùng có thể truy cập vào web đó nhiều lần khác nhau, có thể cùng nội dung hoặc khác nội dung. Tôi sẽ thống kê lại từng nội dung với số lần truy cập của nó. Từ dữ liệu mới có, tôi quét qua các dòng trong dữ liệu nếu phần đường dẫn nào giống nhau thì sẽ tăng lên.

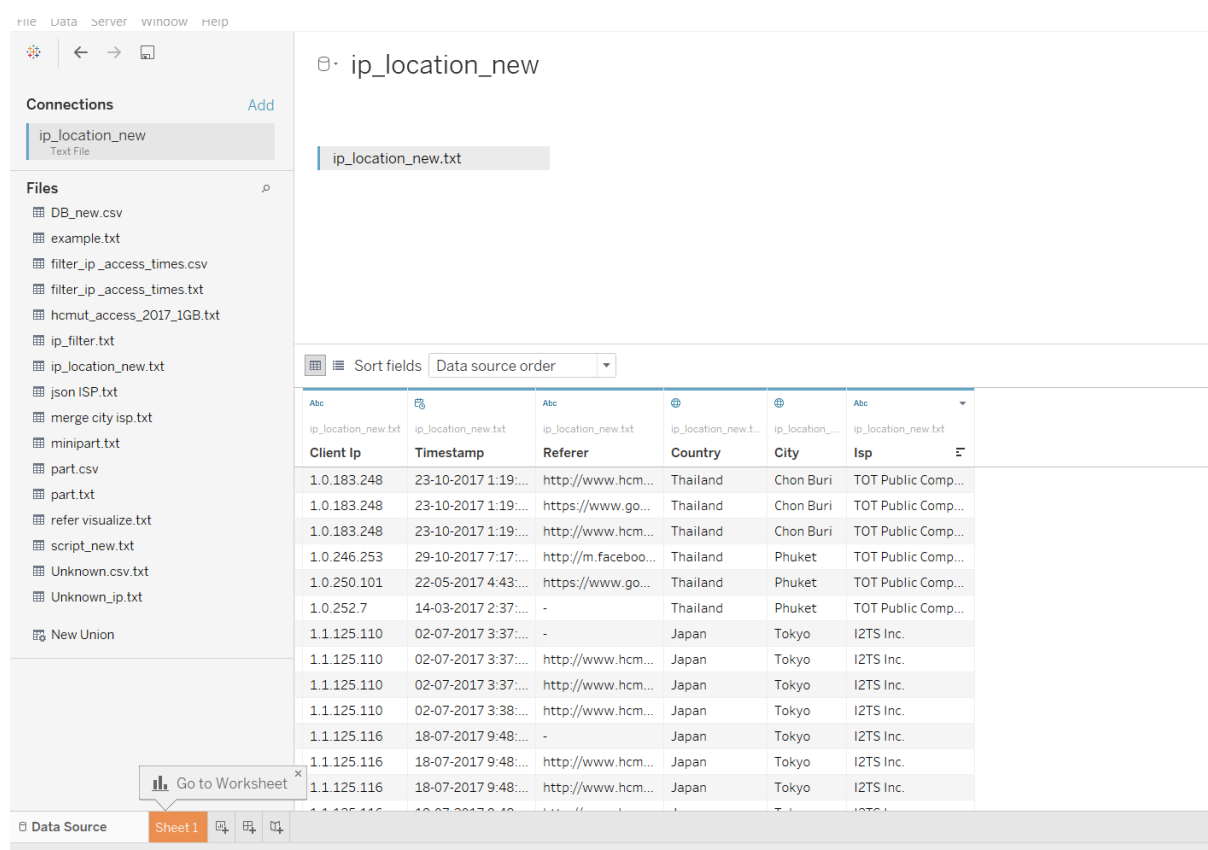
Nhưng khi quét qua các dòng như thế đối với các dữ liệu nhỏ khoảng vài trăm, vài ngàn dòng thì không sao hết, nhưng log này có kích thước lớn với khoảng đến cả chục triệu dòng sẽ gây khó khăn hơn về thời gian xử lý. Qua phần tìm hiểu Spark sử dụng ngôn ngữ Scala, tôi nhận thấy có nhiều hàm hỗ trợ cho việc phân tích nhanh hơn để đạt được mục đích là có tập về số lượng truy cập của các nội dung. Tôi sử dụng hàm *map* được tích hợp sẵn trong Scala và thực thi theo cơ chế song song phân chia lại các phần trong tập dữ liệu mới có đó. Tôi chỉ lấy phần đường dẫn tham khảo ở các dòng trong log để đem đi so sánh và đếm số lần truy cập. Mỗi khi duyệt một dòng mới thì phải kiểm tra xem nó đã có trước đó chưa là tăng lên hay là thêm nó vào nội dung mới. Dựa trên quan sát này, tôi sử dụng hàm *reduceByKey* để duyệt lại các nội dung nhanh chóng và tăng lên số lần nếu nội dung đó đã được nêu trước đó. Chúng ta có thể sắp xếp lại tập dữ liệu để hiển thị xem nội dung nào được tham khảo nhiều nhất và giảm dần theo sau đó. Định danh có các nội dung để có thể hiển thị gọn hơn cho các đường dẫn. Chúng ta đã xếp theo số lần truy cập giảm dần nên sẽ biết rõ định danh đầu là nội dung được tham khảo nhiều nhất và giảm dần theo đó còn định danh sẽ tăng dần theo thứ tự. Như vậy ta có được tập dữ liệu mới là thống kê số lần truy cập của các nội dung.

3.2 Cách trực quan hóa dữ liệu

Các tập dữ liệu được tạo ra ban đầu chỉ các file dưới dạng text sẽ gây khó hiểu khi xem toàn bộ dữ liệu. Nhìn toàn bộ là chữ sẽ gây cho chúng ta khó hiểu rõ các nội dung trong log thay đổi như thế nào. Qua đó nếu chúng ta mô phỏng lên dưới dạng các biểu đồ, ta có thể nhìn một cách khách quan toàn bộ dữ liệu một cách nhanh chóng. Cùng

một tập dữ liệu, có thể tách nhỏ ra từng kiểu khác nhau để thể hiện các ý nghĩa khác có mục đích hơn.

Tôi sử dụng Tableau Desktop phục vụ cho việc trực quan hóa các dữ liệu. Tableau hỗ trợ dữ liệu đầu vào với nhiều loại file khác nhau. Sau khi kết nối với dữ liệu, Tableau sẽ hiển thị với dạng bảng cho từng cột dữ liệu là từng trường.



Hình 3.7 Giao diện sau khi kết nối với dữ liệu

Chúng ta có thể tạo ra các Sheet phục vụ cho từng biểu đồ mình muốn hiển thị. Tạo ra các Dashboard hiển thị nhiều biểu đồ tạo thành bảng thống kê nhiều mục.

Tableau cung cấp cho ta rất nhiều loại biểu đồ như : biểu đồ cột, biểu đồ hình đa giác, biểu đồ tròn, biểu đồ miền, biểu đồ đường và nhiều loại biểu đồ khác.

3.2.1 Biểu đồ trực quan truy cập toàn thế giới

Chúng ta thực hiện hiển thị với biểu đồ thế giới với vị trí các quốc gia, các thành phố. Bản đồ thế giới chúng ta sẽ hiển thị vị trí quốc gia theo kinh độ và vĩ độ của thủ đô. Khung nhìn của Tableau sẽ phân chia theo cột và hàng, ta đưa trường quốc gia vào mục

hiển thị, kèm theo đó là kinh độ và vĩ độ theo cột và hàng để có thể nhìn tổng quan toàn bản đồ thế giới.

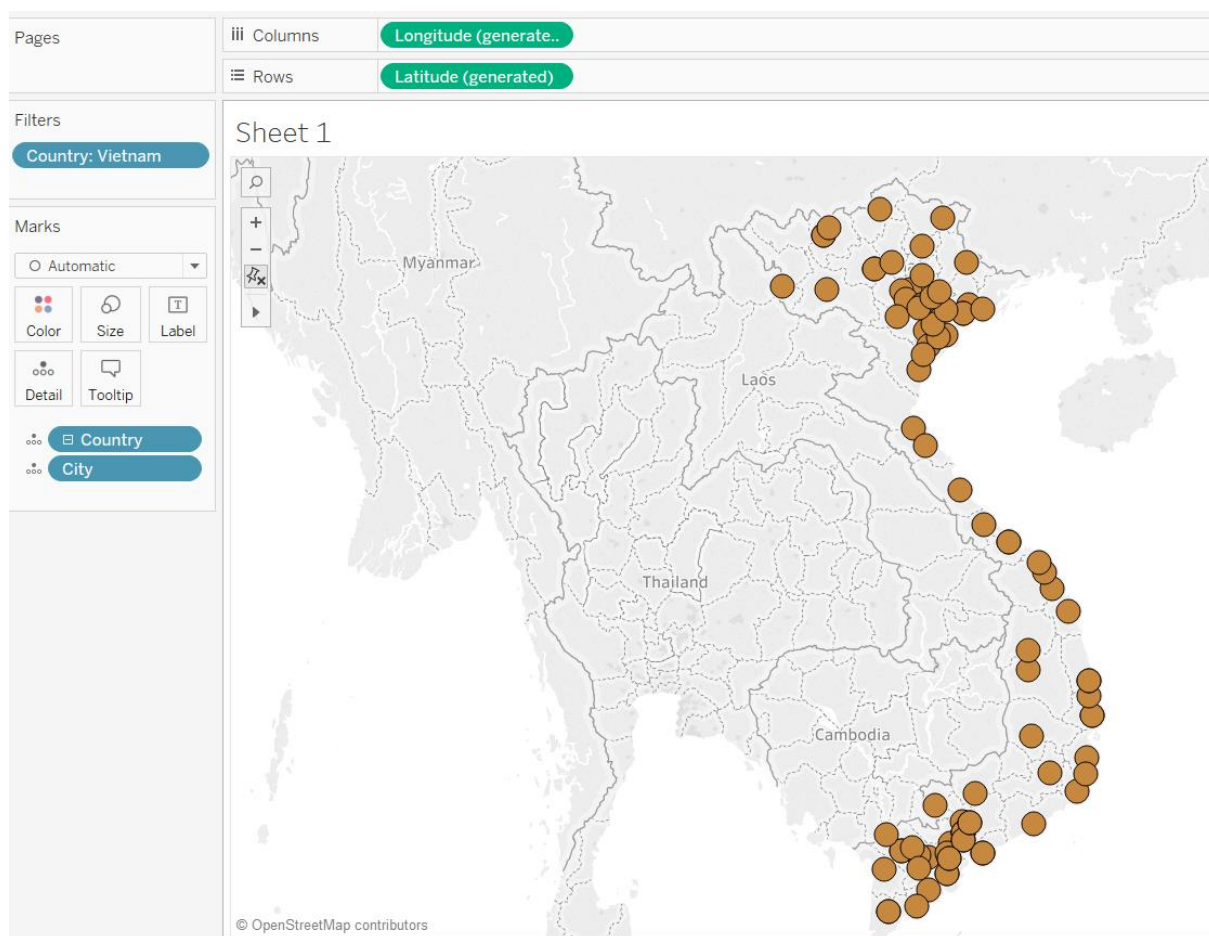


Hình 3.8 Các quốc gia truy cập đến Server

Chúng ta có thể biểu thị thêm kích thước số lượng truy cập của quốc gia đó bằng kích thước vòng tròn. Phần mềm cũng cung cấp cho ta thay đổi kích thước lớn nhỏ của vòng tròn đó, thay đổi màu sắc theo số lượng truy cập. Chú thích thêm cho từng mục hiển thị trong bên bản đồ cho người dùng hiểu rõ hơn. Sau khi thực hiện ta sẽ thấy được sự phân bố các truy cập của các quốc gia trên thế giới. Như vậy ta sẽ đánh giá chi tiết về nước nào truy cập nhiều hay ít, chủ yếu ở các quốc gia nào đến Server của trường.

3.2.2 Biểu đồ truy cập ở Việt Nam

Chúng ta sẽ chọn loại bản đồ thế giới để biểu thị ở Việt Nam. Thêm kinh độ và vĩ độ vào thuộc tính cột và hàng trong Tableau. Phần mềm Tableau Desktop cung cấp cho ta cách để lọc lại dữ liệu theo mong muốn theo các trường có trong dữ liệu. Biểu thị cụ thể ở Việt Nam nên sẽ đưa trường quốc gia vào phần Filters có trong Tableau. Ở mục quốc gia chúng ta sẽ chọn là Việt Nam thì công cụ sẽ hỗ trợ hiển thị chủ yếu và chi tiết hơn tại Việt Nam. Đưa thêm trường thành phố vào mục Marks sẽ giúp hiển thị chi tiết hơn về vị trí cụ thể của từng thành phố ở nước ta.



Hình 3.9 Các thành phố ở Việt Nam truy cập đến Server

Về lượng truy cập ở các thành phố biểu thị bằng kích thước vòng và màu sắc cũng đại diện cho mức độ truy cập. Vòng tròn càng lớn thì truy cập càng nhiều và vòng tròn nhỏ thì truy cập ít hơn. Màu sắc có thể thay đổi theo nhiều dạng ở phần Marks. Biểu thị bằng mức độ tăng độ đậm dần của màu sắc. Khi đó biểu đồ chúng ta sẽ rõ hơn về phân bố lượng truy cập ở từng khu vực thế nào trên toàn Việt Nam.

3.2.3 Biểu đồ lượng truy cập của các ISP tại Việt Nam

Từ tập dữ liệu thống kê về các quốc gia, thành phố và các ISP ở trên thế giới, chúng ta sẽ phân tích về cụ thể tại Việt Nam. Biểu thị bản đồ ở Việt Nam như trên. Lượng truy cập chủ yếu tại Hà Nội và Hồ Chí Minh. Mỗi thành phố có lượng truy cập khác nhau theo các ISP. Khi hiển thị đến bản đồ tại Việt Nam, chúng ta sẽ dùng thêm tính năng lọc dữ liệu cho trường thành phố. Biểu thị ở hai thành phố lớn nên phần đó ta chọn Hà Nội và Hồ Chí Minh.

Để biểu thị độ truy cập ISP ta phải dùng thêm biểu đồ hình tròn để phục vụ chia ra các phần khác nhau. Tableau cung cấp cho ta hiển thị kết hợp các bản đồ với nhau. Ở phần Marks sẽ có hai biểu đồ kế tiếp nhau cho chúng ta cùng hiển thị. Kéo thêm lượng truy cập vào Size để thay đổi kích thước vòng tròn. Có nhiều ISP tại hai thành phố nên ta sẽ chia ra theo màu sắc khác nhau để phân biệt dễ hơn. Thêm trường ISP vào phần color để cho thể chia theo nhiều màu sắc. Khi đó sẽ thấy được các IP đến từ các ISP nào truy cập đến trường chúng ta. Về phần muốn hiển thị chi tiết hơn ta có thể rê chuột đến từng phần hoặc có mục Details trình bày cho ta biết phần đó thuộc quốc gia, thành phố và lượng truy cập của ISP đó như thế nào.

3.2.4 Biểu đồ độ phổ biến nội dung trong một năm

Các nội dung trong log được ghi lại trong một năm, nội dung nào được truy cập nhiều hay ít theo từng ngày, từng tháng. Nên phần này tôi xin biểu thị về độ phổ biến nội dung theo từng tháng. Tôi sẽ sử dụng biểu đồ miền để trực quan hóa dữ liệu log. Trong phần mềm Tableau, ta sẽ lọc ra các tháng trong cột dữ liệu timestamp và đưa vào Columns. Biểu đồ sẽ hiển thị theo 2 cột tung và hoành. Cột tung – thẳng đứng sẽ biểu diễn số lượng truy cập theo thời gian và cụ thể là theo tháng. Cột hoành là phân chia ra theo từng tháng. Để hiển thị từng nội dung ta thêm vào trường chữ thông tin truy cập vào phần Marks trong Tableau. Phần mềm sẽ hỗ trợ ta biểu thị lên giao diện về sự thay đổi của các nội dung trong log.

CHƯƠNG 4 : ĐÁNH GIÁ

Chương này tôi sẽ thực hiện đánh giá lại các kết quả sau khi phân tích ra các tập dữ liệu và hiển thị bằng công cụ Tableau. Đánh giá về các tập dữ liệu mới được tạo ra từ phân tích bằng Spark sẽ có ích và hỗ trợ như thế nào cho bài toán CDN. Các biểu đồ Tableau sẽ cho ta thấy khái quát hơn về toàn bộ dữ liệu để có thể tìm hiểu và nhận xét về các cái mới cho sau này.

4.1 Môi trường thực nghiệm

Các phiên bản của các phần mềm được sử dụng thực hiện trong nghiên cứu:

Apache Spark : 2.2.1

Scala : 2.11.6

Java : 1.8.0_171

Tableau :10.0.2648

Trong phần thực hiện đề tài luận văn này tôi sử dụng máy tính cá nhân với cấu hình là :

- Ubuntu :16.04 LTS
- Ram : 16 GB
- CPU : Intel Core i7 Kabylake, 7700HQ, 2.80 GHz
- CPU core : 8

4.2 Phân tích dữ liệu dùng Spark

4.2.1 Tập dữ liệu về thống kê lượng truy cập

```

1, (http://www.hcmut.edu.vn/vi,4199841)
2, (-,3425097)
3, (http://www.hcmut.edu.vn/,1908131)
4, (https://www.google.com.vn/,745344)
5, (http://m.facebook.com/,138145)
6, (http://hcmut.edu.vn/,93884)
7, (http://www.hcmut.edu.vn/en,86549)
8, (http://www.hcmut.edu.vn/vi/welcome/view/menu-chinh/gioi-thieu/tong-quan/lich-su-hinh-thanh-va-phat-trien-,66080)
9, (http://www.hcmut.edu.vn/vi/event/view/thong-bao,62903)
10, (http://coccoc.com/search,55841)
11, (https://www.google.com/,55175)
12, (http://www.hcmut.edu.vn/vi/newsletter/view/tin-tuc/3808-dhbk-cong-bo-diem-trung-tuyen-dh-cd-chinh-quy-2016,51259)
13, (http://www.hcmut.edu.vn/vi/newsletter/view/su-kien,44683)
14, (http://www.hcmut.edu.vn/vi/welcome/tieudiem/336,42627)
15, (http://hcmut.vnptwifi.vn/login?username=vnpt&password=vnpt&dst=http://www.hcmut.edu.vn/,41392)
16, (http://10.28.232.1:8002/index.php?zone=wifi_public&redirurl=http%3A%2F%2Fconnectivitycheck.gstatic.com%2Fgenerate_204,39840)
17, (http://www.fme.hcmut.edu.vn/fme/index.php,38085)
18, (https://www.facebook.com/,35017)
19, (http://m.facebook.com,32876)
20, (http://www.hcmut.edu.vn/vi/welcome/view/cac-don-vi-truc-thuoc/phong--ban-chuc-nang/thu-vien,32346)
21, (http://10.28.232.1:8002/index.php?zone=wifi_public&redirurl=http%3A%2F%2Fconnectivitycheck.android.com%2Fgenerate_204,31434)
22, (http://www.hcmut.edu.vn/vi/welcome/tieudiem/508,23552)
23, (http://www.hcmut.edu.vn/vi/welcome/view/cac-don-vi-truc-thuoc/-khoa--trung-tam-dao-cao/khoa-ky-thuat-xay-dung,22656)
24, (http://www.hcmut.edu.vn/vi/event/view/hoc-bong--tuyen-dung,22628)
25, (https://private.hcmut.edu.vn:8003/index.php?zone=wifi_240&redirurl=http%3A%2F%2Fwww.windowsphone.com%2F,22575)
26, (http://fme.hcmut.edu.vn/fme/index.php,22003)
27, (http://10.28.232.1:8002/index.php?zone=wifi_public&redirurl=http%3A%2F%2Fapple.com%2Fphotospot-detect.html,20268)
28, (http://www.hcmut.edu.vn/vi/newsletter/view/tin-tuc/4214-116-truong-thpt-duoc-uu-tien-xet-tuyen-vao-bach-khoa,19085)
29, (http://www.google.com.vn/,16918)
30, (http://www.hcmut.edu.vn/vi/welcome/view/cac-don-vi-truc-thuoc/phong--ban-chuc-nang/phong-dao-cao,16031)
31, (http://www.hcmut.edu.vn/vi/welcome/view/menu-chinh/gioi-thieu/bo-may-to-chuc/ban-giam-hieu,15086)
32, (http://10.28.232.1:8002/index.php?zone=wifi_public&redirurl=http%3A%2F%2Fclients3.google.com%2Fgenerate_204,15042)
33, (android-app://com.google.android.googlequicksearchbox,14720)
34, (http://10.28.232.1:8002/index.php?zone=wifi_public&redirurl=http%3A%2F%2Fgo.microsoft.com%2Ffwlink%2F%3FLinkId%3D219472,14589)
35, (http://www.hcmut.edu.vn/en/welcome/view/menu-intro/introduction/overview/message-from-the-rector,14487)
36, (http://www.hcmut.edu.vn/vi/newsletter/view/tin-tuc,13965)
37, (http://www.hcmut.edu.vn/vi/welcome/view/cac-don-vi-truc-thuoc/-khoa--trung-tam-dao-cao/khoa-co-khi,13393)
38, (http://www.hcmut.edu.vn/en/,13289)
39, (http://www.hcmut.edu.vn/vi/welcome/view/cac-don-vi-truc-thuoc/-khoa--trung-tam-dao-cao/khoa-dien--dien-tu,13165)
40, (http://www.hcmut.edu.vn/vi/welcome/view/cac-don-vi-truc-thuoc/-khoa--trung-tam-dao-cao/khoa-ky-thuat-hoa-hoc,13129)
41, (http://www.hcmut.edu.vn/vi/event/view/van-ban-moi,12142)

```

Hình 4.1 Thống kê lượng truy cập của từng nội dung

Tập dữ liệu hiển thị thông tin về các trường: Id, nội dung, đếm lượng truy cập. Tập dữ liệu hiển thị được nội dung theo số lượng giảm dần. Nội dung được truy cập nhiều nhất là: <http://www.hcmut.edu.vn/vi> với số lượng truy cập là 4.199.841. Và các nội dung khác với số lượng khá nhiều cũng có sự chênh lệch khá lớn.

4.2.2 Tập dữ liệu về vị trí và ISP của IP

```

1.0.183.248;2017-10-23 01:19:28.0;http://www.hcmut.edu.vn/en/;Thailand;Chon Buri;TOT Public Company Limited
1.0.183.248;2017-10-23 01:19:47.0;https://www.google.co.th/;Thailand;Chon Buri;TOT Public Company Limited
1.0.183.248;2017-10-23 01:19:48.0;http://www.hcmut.edu.vn/en/;Thailand;Chon Buri;TOT Public Company Limited
1.0.246.253;2017-10-29 19:17:49.0;http://m.facebook.com;Thailand;Phuket;TOT Public Company Limited
1.0.250.101;2017-05-22 16:43:03.0;https://www.google.com/;Thailand;Phuket;TOT Public Company Limited
1.0.252.7;2017-03-14 02:37:32.0;-;Thailand;Phuket;TOT Public Company Limited
1.1.125.110;2017-07-02 15:37:35.0;-;Japan;Tokyo;I2TS Inc.
1.1.125.110;2017-07-02 15:37:36.0;http://www.hcmut.edu.vn/;Japan;Tokyo;I2TS Inc.
1.1.125.110;2017-07-02 15:37:38.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.110;2017-07-02 15:38:08.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.116;2017-07-18 09:48:31.0;-;Japan;Tokyo;I2TS Inc.
1.1.125.116;2017-07-18 09:48:32.0;http://www.hcmut.edu.vn/;Japan;Tokyo;I2TS Inc.
1.1.125.116;2017-07-18 09:48:33.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.116;2017-07-18 09:48:40.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.116;2017-07-18 09:50:35.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.116;2017-07-18 09:50:42.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.116;2017-07-18 09:50:51.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.116;2017-07-18 09:51:00.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.116;2017-07-18 09:52:16.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.116;2017-07-18 09:55:05.0;http://www3.hcmut.edu.vn/index.php/vi/noibo/;Japan;Tokyo;I2TS Inc.
1.1.125.71;2017-07-04 13:50:58.0;-;Japan;Tokyo;I2TS Inc.
1.1.125.71;2017-07-04 13:50:58.0;http://www.hcmut.edu.vn/;Japan;Tokyo;I2TS Inc.
1.1.125.71;2017-07-04 13:50:58.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.71;2017-07-04 13:51:15.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.71;2017-07-04 13:51:50.0;http://www3.hcmut.edu.vn/index.php/vi/noibo/;Japan;Tokyo;I2TS Inc.
1.1.125.71;2017-07-04 13:56:22.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.79;2017-07-24 09:13:41.0;-;Japan;Tokyo;I2TS Inc.
1.1.125.79;2017-07-24 09:13:42.0;http://www.hcmut.edu.vn/;Japan;Tokyo;I2TS Inc.
1.1.125.79;2017-07-24 09:13:47.0;http://www.hcmut.edu.vn/;Japan;Tokyo;I2TS Inc.
1.1.125.79;2017-07-24 09:14:51.0;http://www.hcmut.edu.vn/;Japan;Tokyo;I2TS Inc.
1.1.125.79;2017-07-24 09:14:53.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.79;2017-07-24 09:15:10.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.79;2017-07-24 09:15:21.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.79;2017-07-24 09:15:28.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.79;2017-07-24 09:15:48.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.
1.1.125.79;2017-07-24 09:16:10.0;http://www3.hcmut.edu.vn/index.php/vi/noibo/;Japan;Tokyo;I2TS Inc.
1.1.125.79;2017-07-24 09:30:00.0;http://www.hcmut.edu.vn/vi;Japan;Tokyo;I2TS Inc.

```

Hình 4.2 Thông tin về vị trí và ISP của từng IP

Sau khi thực hiện phân tích ở bước 3.1.3 ta sẽ thu được output có chứa thêm các thông tin về quốc gia, thành phố, ISP. Qua đó ta thấy các IP sẽ gắn với các quốc gia cụ thể. Ta sẽ thấy như IP : 1.0.183.248 ở nước Thái Lan với thành phố(tỉnh) là Chon Buri, với tổ chức là TOT Public Company Limited. Như vậy tập dữ liệu mới này nhỏ gọn hơn và chứa các thông tin có liên quan đến IP như chúng ta muốn cụ thể về các trường là IP, thời gian, nội dung, quốc gia, thành phố, và tổ chức của ISP.

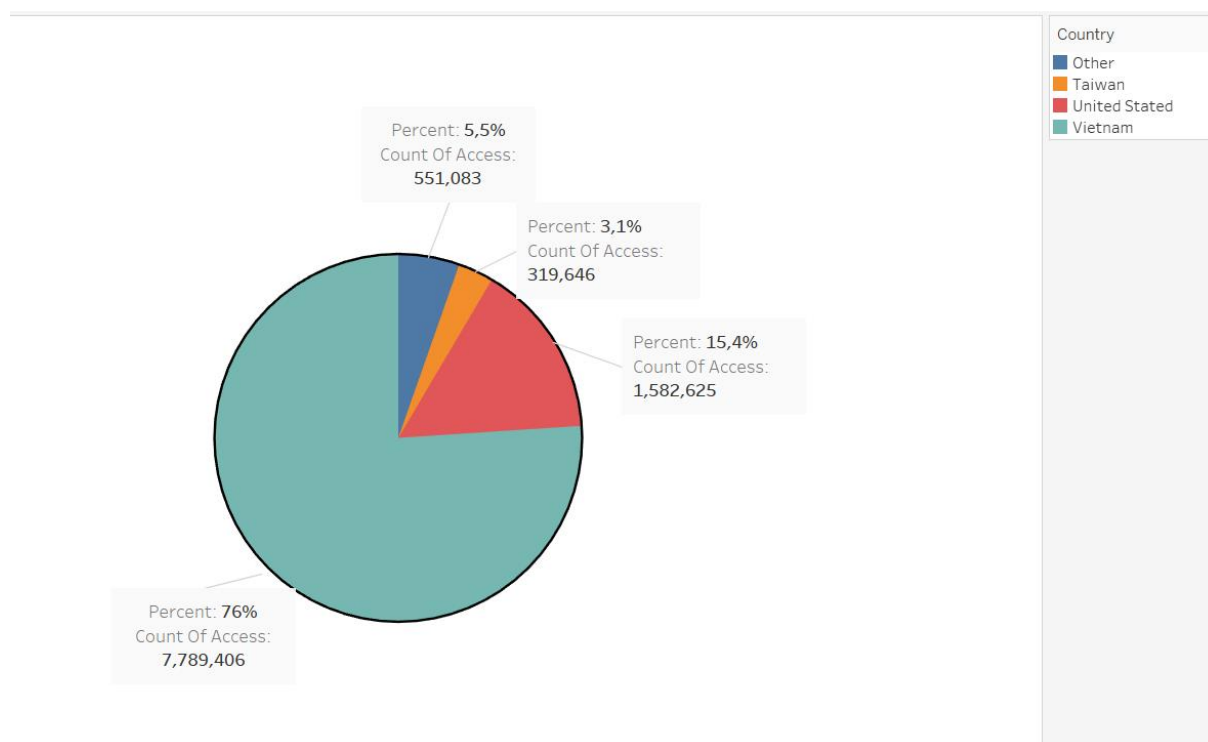
4.3 Trực quan hóa dữ liệu dùng Tableau

Trong phần này tôi sẽ hiển thị các kết quả có được sau khi phân tích bằng Spark và trực quan bằng công cụ Tableau.

4.3.1 Lượng truy cập của các quốc gia trên thế giới

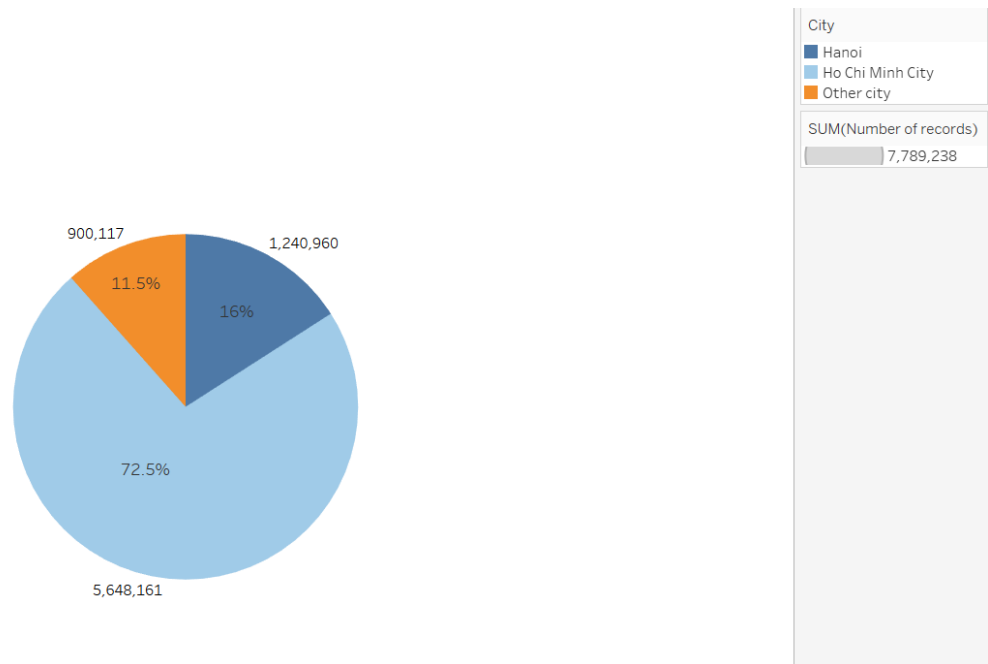


Hình 4.3 Số lượng truy cập của các quốc gia trên thế giới



Hình 4.4 Tỷ lệ phần trăm lượng truy cập các nước trên thế giới

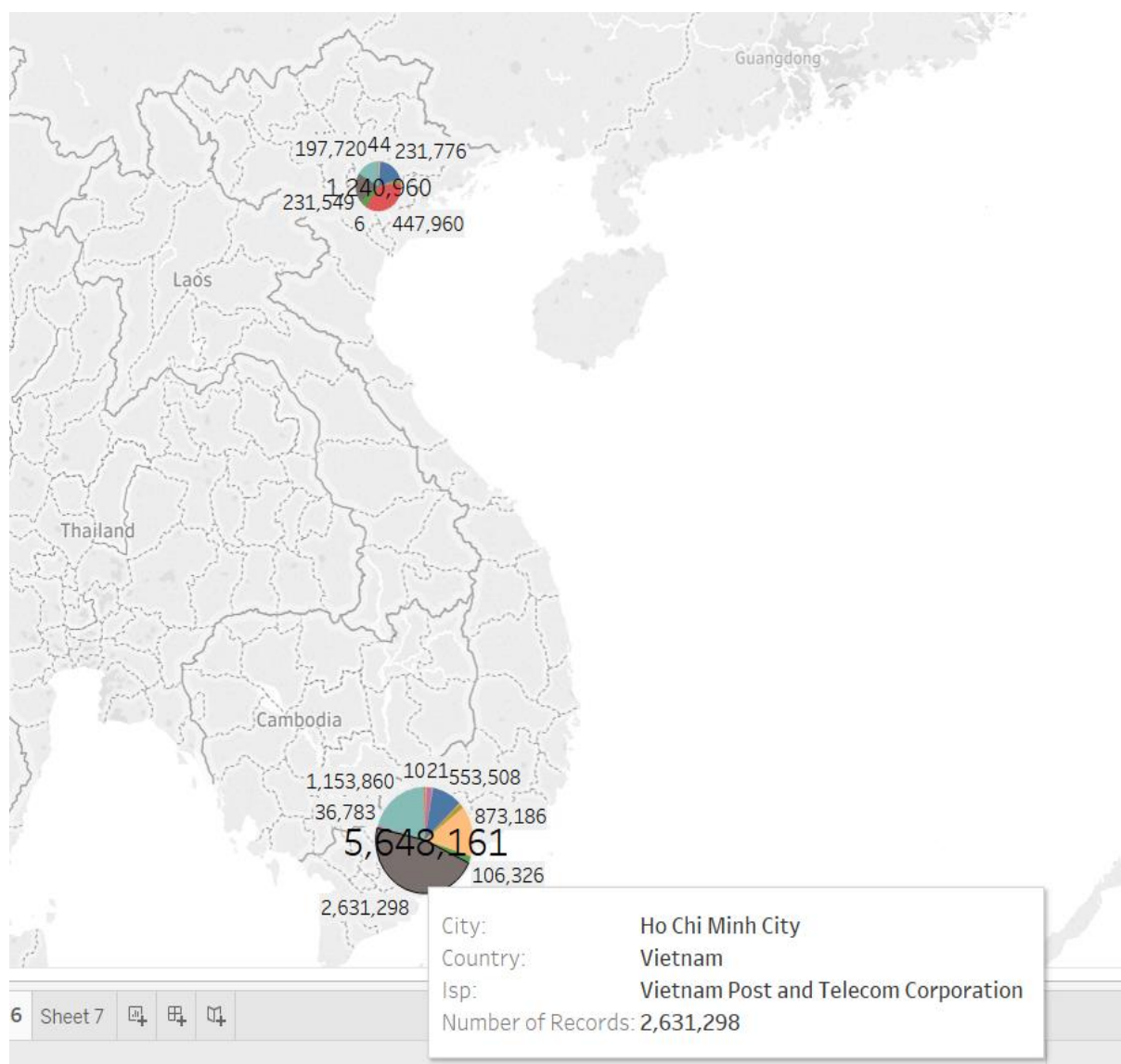
Từ tập dữ liệu về vị trí các quốc gia và có số lượng truy cập, tôi hiển thị trên bản đồ thế giới về vị trí và số lượng truy cập của các quốc gia. Mỗi chấm đỏ bên trên bản đồ thế



Hình 4.6 Tỷ lệ phần trăm chiếm nội dung của 2 thành phố lớn

Ở hình 4.3 xét tình hình chung về các quốc gia trên thế giới để xem về lượng truy cập của các nước. Biểu đồ 4.5 sẽ hiển thị cụ thể hơn về chi tiết các thành phố trong nước Việt Nam như thế nào. Qua phân tích các số liệu cụ thể ở Việt Nam để hiển thị lên biểu đồ, thấy về vị trí các thành phố và lượng truy cập. Cho ta thấy được về lượng truy cập chủ yếu ở các thành phố nào cụ thể lớn nhất là Hồ Chí Minh, sau đó là Hà Nội,... Bản đồ có hiển thị vài số liệu cụ thể ở một số thành phố. Kích thước của vòng đại diện cho số lượng truy cập nhiều hay ít làm cho nó to hay nhỏ. Bên cạnh đó màu sắc cũng có mức độ tăng dần từ nhạt đến đậm cũng cho ta nhìn rõ từng khu vực truy cập nhiều ít khác nhau. Màu càng đậm thì tương ứng với truy cập càng nhiều. Kích thước của thành phố Hồ Chí Minh là lớn nhất với số lượng truy cập lên đến 5.648.161 và chiếm đến 72.5% trên tổng toàn bộ lượng truy cập ở Việt Nam. Thành phố nổi bật thứ hai là Hà Nội với lượng truy cập là 1.240.960 và chiếm 16% trên toàn bộ lượng truy cập. Phân bố lượng truy cập rải trên toàn nước Việt Nam.

4.3.3 Thống kê theo ISP về lượng truy cập ở Việt Nam

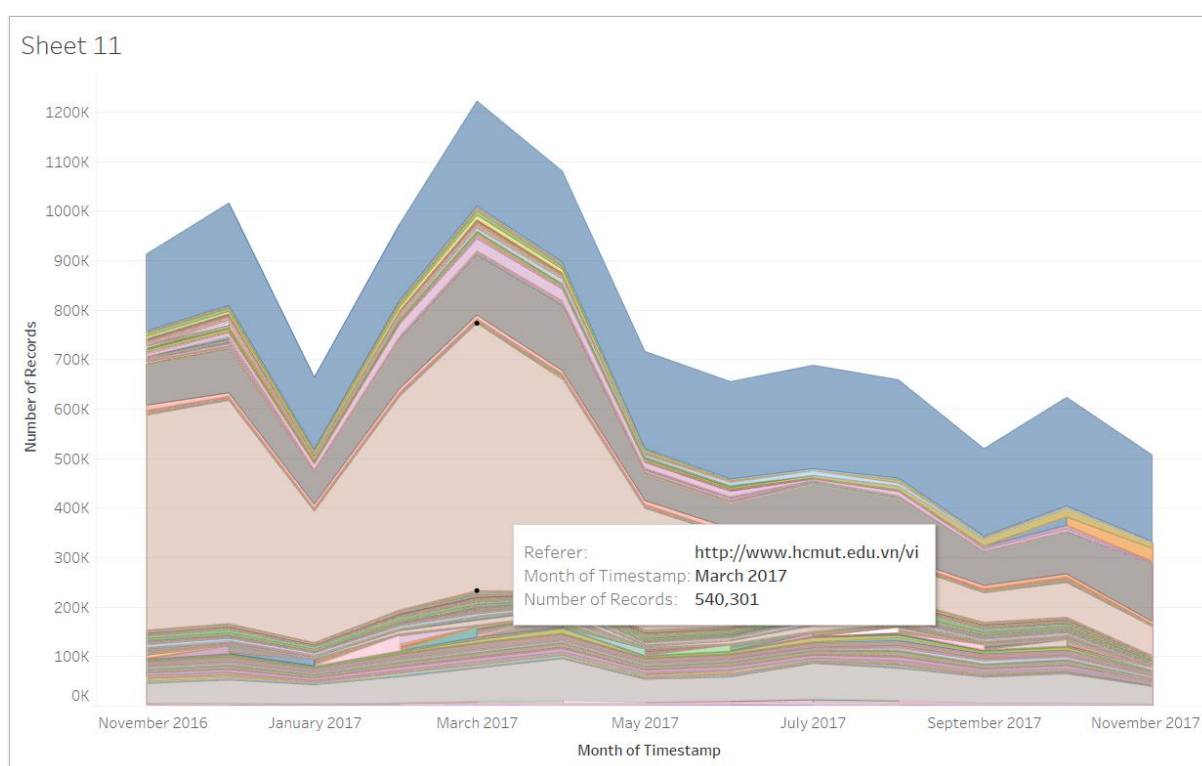


Hình 4.7 Thống kê lượng truy cập của các ISP ở Hà Nội và Hồ Chí Minh

Chúng ta có thể thấy được tổ chức của các ISP ở từng khu vực. Hình trên biểu thị về thông tin truy cập của các nhà phân phối mạng internet. Biểu diễn lượng truy cập ở hai thành phố có lượng xem nhiều nhất là Hồ Chí Minh và Hà Nội. Thống kê lại từng nhà mạng có lượng truy cập phân bố chủ yếu như thế nào. Đánh giá chung khác nhau về nhà mạng ở từng khu vực lượng xem khác nhau. Các dãy IP thuộc các nhà mạng đó tham khảo nội dung nào cũng sẽ có thể giúp cho các nhà mạng quan tâm đến hơn, chú ý lưu

lại các nội dung chính và được nhiều người xem. Với các số liệu nổi bật thì toàn thành phố Hồ Chí Minh 5.648.161 thì các ISP của tổ chức Vietnam Post and Telecom Corporation đã chiếm đến 2.631.298, gần 50% của toàn thành phố Hồ Chí Minh. Ở Hà Nội thì các nhà mạng cũng chiếm khá lớn trên tổng thể lượng truy cập. Như thế có ích cho nhà phân phối mạng và Server của trường. Trường có thể quan tâm hơn nhà phân phối mạng nào chủ yếu truy cập đến trường.

4.3.3 Độ phổ biến nội dung



Hình 4.8 Độ phổ biến nội dung trên web trường trong một năm

Qua hình ta thấy được độ phát triển hay giảm dần bớt của từng nội dung. Nội dung nào được truy cập như thế nào theo từng tháng trong năm. Tìm hiểu rõ hơn về nội dung đó theo thời gian. Nội dung được truy cập nhiều nhất tại tháng 3 năm 2017 là "http://www.hcmut.edu.vn/vi" với lượng thống kê là 540.301. Nhưng thấy nội dung đó giảm dần sau đó tới tháng 10 năm 2017 là: 69.688.

CHƯƠNG 5 : TỔNG KẾT VÀ HƯỚNG PHÁT TRIỂN

Trong chương này, tôi tóm tắt lại những kết quả đạt được và những đóng góp của đề tài vào bài toán CDN. Cuối cùng, tôi đưa ra những đề xuất và hướng mở rộng để tiếp tục phát triển theo hướng luận văn này.

5.1 Tổng kết

Xây dựng các tập dữ liệu là một bước thực hiện tại các Cache Server để thực hiện tính toán và dự đoán cho các thời gian ở phía sau này. Phân tích dữ liệu lớn là một khó khăn cho quá trình thực hiện.

Tìm hiểu và phân tích ý nghĩa trong mối liên quan giữa các thành phần trong log file. Khai hết toàn bộ ý nghĩa trong log là một vấn đề lớn để thực hiện và tìm ra ý nghĩa của chúng. Trong phần đề luận văn này tôi có giới hạn phạm vi nghiên cứu là tập trung vào khai thác IP và một số mối liên quan đến IP.

Với việc IP và các mối liên quan để phân tích và xây dựng các tập dữ liệu mới có ích hơn từ log file thô ban đầu. Tôi tìm hiểu và tạo ra các tập dữ liệu về thống kê các lượng truy cập của các nội dung. Từ IP có thể xác định vị trí đến thành phố và quốc gia để tạo thêm tập dữ liệu về vị trí các IP.

Bên cạnh đó qua các tập dữ liệu tôi có trực quan hóa tạo ra các biểu đồ có ý nghĩa. Biểu diễn các quốc gia trên thế giới cũng như thành phố truy cập đến Server của trường. Thống kê lại các phân bố nội dung ở trong và ngoài nước. Đến chi tiết hơn là về cụ thể các thành phố nổi trội và chiếm chủ yếu lượng truy cập ở nước ta. Bản đồ phân bố nội dung ra theo từng tháng từ dữ liệu trong log file.

Tóm lại, quá trình thực hiện đề tài có tạo ra một số kết quả có ý nghĩa phục vụ cho bài toán CDN:

- Tập dữ liệu về thống kê lại số lượng truy cập các nội dung
- Tập dữ liệu về quốc gia, thành phố, và tổ chức của ISP
- Tập dữ liệu về các IP không tìm kiếm được trong log
- Biểu đồ lượng truy cập của các nước trên thế giới
- Biểu đồ lượng truy cập cụ thể các thành phố ở Việt Nam
- Biểu đồ phân bố nội dung theo từng tháng trong log

5.2 Hướng phát triển

Trong đề tài luận văn còn rộng phía sau để xây dựng các tập dữ liệu khác có ích cho bài toán mạng phân phối nội dung CDN.

Phân tích dữ liệu trong log file sẽ có nhưng khó khăn như dữ liệu ngày càng lớn hơn nữa, nội dung phong phú hơn, có những lỗi cấu trúc khác,...

Trong log file có nhiều dữ liệu chưa được xử lý và tìm hiểu. Phần trên tôi chủ yếu tìm hiểu và tìm các mối liên quan đến IP trong log. Còn nhiều phần như nội dung, các đường dẫn tham khảo, thời gian, các phương thức truy cập,... Các nội dung nào có liên quan sẽ nên được sắp xếp và chia ta theo từng phần khác nhau. Kết hợp giữa các nội dung với IP theo một hướng mới nào khác để tìm tập dữ liệu mới.

Tìm hiểu các mối sự liên quan với nhau trong đường dẫn, nội dung có thể thống kê các nguồn truy cập từ đâu đến đâu. Sẽ chú ý hơn đến thời gian truy cập và các sự liên quan đến nội dung sau này.

Từ các vấn đề về nội dung và thời gian sẽ có thể tạo ra các tập dữ liệu mới nhiều hơn, có ích cho qua các giai đoạn tính toán thực hiện ở Cache Server.

TÀI LIỆU THAM KHẢO

- [1] Nguyễn Đăng Thế, (2010). *Tìm hiểu mạng phân phối nội dung (Content Delivery Networks: CDN)*. Luận văn thạc sĩ, Trường Đại học Bách khoa Hà Nội
- [2] Saif Muttair, (2015) *CONTENT DELIVERY NETWORK* , Last accessed date:10/06/2018
- [3] Spark , <https://spark.apache.org/>, Last accessed date:10/06/2018
- [4] Đức (2016), *Giải pháp xếp hạng và tính toán song song trên nền tảng Apache Spark*. Luận văn Thạc sĩ, Trường Đại học Công Nghệ(Hà Nội)
- [5] BIGDATA ANALYTICS, <https://dataartblog.wordpress.com/2014/11/20/gioi-thieu-ve-apache-spark/>, Last accessed date:10/06/2018
- [6] Matei Zaharia, Patrick Wendell, Andy Konwinski, Holden Karau(2015). *Learning Spark: Lightning-Fast Big Data Analysis*
- [7] Resilient Distributed Datasets, <https://spark.apache.org/docs/latest/rdd-programming-guide.html>, Last accessed date:10/06/2018
- [8] Components, <https://spark.apache.org/docs/latest/cluster-overview.html>, Last accessed date:10/06/2018
- [9] Zeppelin, <https://zeppelin.apache.org/>, Last accessed date:10/06/2018
- [10] Tableau, <https://www.tableau.com/>, Last accessed date:10/06/2018