HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

HOÀNG THỊ HUYỀN TRANG

NGHIÊN CỨU CÁC KỸ THUẬT VÀ CÔNG CỤ PHÂN TÍCH WEB LOG

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – 2022

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

HOÀNG THỊ HUYỀN TRANG

NGHIÊN CỨU CÁC KỸ THUẬT VÀ CÔNG CỤ PHÂN TÍCH WEB LOG

Chuyên ngành: **Hệ thống thông tin**

Mã số: **8.48.01.04**

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. HOÀNG XUÂN DẬU

HÀ NỘI – 2022

LỜI CAM ĐOAN

Tôi xin cam đoan các kết quả nghiên cứu trong luận văn này là sản phẩm của cá nhân tôi dưới sự hướng dẫn của thầy giáo PGS.TS. Hoàng Xuân Dậu. Các số liệu, kết quả được công bố là hoàn toàn trung thực. Những điều được trình bày trong toàn bộ luận văn này là những gì do tôi tự nghiên cứu hoặc là được tổng hợp từ nhiều nguồn tài liệu khác nhau. Các tài liệu tham khảo có xuất xứ rõ ràng và được trích dẫn đầy đủ, hợp pháp. Tôi xin hoàn toàn chịu trách nhiệm trước lời cam đoan của mình.

Tác giả luận văn

Hoàng Thị Huyền Trang

LỜI CẢM ƠN

Trong suốt quá trình học tập và hoàn thành luận văn tốt nghiệp, học viên đã nhận được rất nhiều sự giúp đỡ, động viên từ thầy cô, gia đình và bạn bè. Học viên xin chân thành cảm ơn sự giúp đỡ này.

Trước tiên em xin cảm ơn Ban giám đốc, Khoa sau Đại học – Học Viên Bưu Chính Viễn Thông đã giúp đỡ và tạo điều kiện tốt cho học viên học tập trong thời gian qua.

Học viên cũng xin cảm ơn các thầy cô trong khoa Công Nghệ Thông Tin 1- Học viện Bưu chính Viễn thông đã truyền đạt cho tôi những kiến thức chuyên sâu về chuyên ngành trong suốt thời gian học tập, để học viên có được nền tảng kiến thức hỗ trợ rất lớn cho học viên trong quá trình làm luận văn của mình.

Học viên cũng muốn bày tỏ sự biết ơn sâu sắc tới TS Hoàng Xuân Dậu, người đã định hướng cho học trong việc lựa chọn đề tài, đưa ra những nhận xét quý giá và trực tiếp hướng dẫn học viên trong suốt quá trình nghiên cứu và hoàn thành luận văn tốt nghiệp.

Học viên củng xin gửi lời cảm ơn chân thành đến tất cả các Thầy Cô của trường Học Viện Công Nghệ Bưu Chính Viễn Thông đã giảng dạy và dìu dắt chúng em trong suốt quá trình học tập tại Trường giúp Học viên vượt qua những giai đoạn khó khăn và tạo điều kiện thuận lợi cho học viên học tập tốt và hoàn thành luận văn này.

Xin chân thành cảm ơn tất cả mọi người!

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
API	Application Programming Interface	Giao diện lập trình ứng dụng
ASCII	American Standard Code for Information Interchange	Chuẩn mã trao đổi thông tin Hoa Kỳ
CSDL		Cơ sở dữ liệu
CSS	Cascading Style Sheets	Tập tin định kiểu theo tầng
DNS	Domain Name System	Hệ thống tên miền
GELF	Graylog Extended Log Format	Định dạng nhật ký mở rộng Graylog
HTTP	Hypertext Transfer Protocol	Giao thức truyền siêu văn bản
ISP	Internet Service Provider	Nhà cung cấp dịch vụ Internet
LAN	Local area network	Mạng máy tính cục bộ
PHP	Hypertext Preprocessor	Một ngôn ngữ lập trình kịch bản
SNMP	Simple Network Management Protocol	Giao thức quản lý mạng đơn giản
SQL	Structured Query Language	Ngôn ngữ truy vấn có cấu trúc
TCP	Transmission Control Protocol	Giao thức điều khiển truyền vận
UDP	User Datagram Protocol	Giao thức dữ liệu người dùng
URI	Uniform Resource Identifier	Mã định danh tài nguyên đồng nhất

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

MỞ ĐẦU

1. Lý do chọn đề tài:

Với sự phát triển của công nghệ kỹ thuật số, hành trình mua sắm của người tiêu dùng ngày càng phức tạp. Và với các doanh nghiệp kinh doanh trực tuyến, hiểu hành vi người dùng là điều càng quan trọng. Hiểu được hành vi của người dùng giúp doanh nghiệp xây dựng chiến lược marketing phù hợp, tiếp thị trên mạng xã hội, kích thích nhu cầu tiêu dùng của khách hàng.

Có rất nhiều công cụ giúp các doanh nghiệp phân tích hành vi người dùng, trong số đó phải kể đến các công cụ phân tích web log. Hành vi người dùng có thể được trích xuất từ việc phân tích các file web log. Log(còn được gọi là nhật ký, hoặc dấu vết) là các mục nhập thông tin được tạo ra bởi các ứng dụng hoặc hệ điều hành trong quá trình hoạt động. Hiện nay mọi ứng dụng, hệ thống lớn nhỏ đều có thực hiện ghi log. Mỗi nhật ký log thường được tạo bởi một hoạt động hoặc sự kiện, vì vậy nó còn được gọi là nhật ký sự kiện. Một số trình tạo nhật ký phổ biến là hệ điều hành, thiết bị mạng (như bộ định tuyến, tường lửa, v.v.), máy chủ dịch vụ (máy chủ web, máy chủ cơ sở dữ liệu, máy chủ DNS, máy chủ Email, v.v.) và các chương trình ứng dụng.Những lợi ích mà việc thu thập, xử lý và phân tích log mang lại bao gồm:

- Kiểm tra sự tuân thủ các chính sách an ninh;
- Hiểu các hành vi của người dùng trực tuyến, trên cơ sở đó tối ưu hóa hệ thống cho phục vụ tốt hơn cho người dùng hoặc quảng cáo trực tuyến.

Như vậy, việc xử lý và phân tích log đem lại nhiều lợi ích, đặc biệt trong việc đảm bảo an toàn thông tin và cải thiện chất lượng hệ thống và các dịch vụ kèm theo, như quảng cáo trực tuyến thông qua việc phân tích hành vi người dùng sử dụng log. Ngoài ra, khi hệ thống gặp sự cố, web log cũng là một nguồn cung cấp các dữ liệu quan trọng cho quản trị viên để tìm hiểu nguyên nhân và khắc phục sự cố.

Hiện nay có nhiều kỹ thuật và công cụ khác nhau được sử dụng trong thu thập và phân tích web log. Đây cũng là hướng nghiên cứu của luận văn này với đề tài "***Nghiên cứu các kỹ thuật và công cụ phân tích web log***". Mục đích của luận văn là nghiên cứu các kỹ thuật và công cụ xử lý và phân tích web log, sau đó áp dụng các

kiến thức tìm hiểu được để xây dựng mô hình và thử nghiệm ứng dụng phân tích web log nhằm trích xuất các báo cáo về hành vi truy nhập web của người dùng.

2. Tổng quan về vấn đề nghiên cứu:

Việc thu thập, xử lý và phân tích nhật ký truy cập hệ thống nói chung và nhật ký web log nói riêng là công việc không thể thiếu trong việc giám sát hệ thống, phân tích hành vi người dùng, phát hiện bất thường, phát hiện tấn công,phát hiện xâm nhập.[2] Từ dữ liệu nhật ký web thô được thu thập, thông qua quá trình xử lý và phân tích, chúng ta có thể trích xuất thông tin quan trọng về hành vi của người dùng trực tuyến, cũng như các tín hiệu hoặc khả năng xuất hiện của hành vi truy cập bất thường. Kết quả phân tích hành vi người dùng trực tuyến là cơ sở để tối ưu hóa và nâng cao chất lượng của hệ thống và các dịch vụ đi kèm, như quảng cáo trực tuyến ... nhằm đáp ứng tốt nhất yêu cầu của người dùng.

Nhiều giải pháp kỹ thuật và các nền tảng, công cụ xử lý, phân tích log đã được nghiên cứu và triển khai. Các giải pháp xử lý và phân tích log thường tập trung thực hiện các phần việc:

- Nhận dạng mẫu: nhận dạng các mẫu xuất hiện trong các bản ghi log.

- Chuẩn hóa: chuyển các dạng dữ liệu log về một dạng chuẩn chung cho các khâu xử lý tiếp theo.

- Phân loại và gán nhãn: phân loại các bản ghi log và gán nhãn chúng bằng các từ khóa.

- Phân tích tương quan: là kỹ thuật thu thập các thông điệp từ các hệ thống khác nhau và tìm tất cả các thông điệp thuộc về cùng một sự kiện.

- Phát hiện các bất thường nhân tạo: kỹ thuật cho phép nhận dạng, phát hiện các bất thường mới, hoặc hiếm gặp.

Các công cụ và nền tảng xử lý, phân tích log nói chung và web log nói riêng có thể được chia thành hai nhóm chính [1], bao gồm: nhóm các công cụ xử lý, phân tích log cho đảm bảo an toàn hệ thống, như hệ thống phát hiện xâm nhập OSSEC [6], Splunk [7], IBM Qradar SIEM [8] và các công cụ quản lý log, như Graylog [9], ELK Stack [10] và LOGalyze [11].

3. Mục đích nghiên cứu:

Luận văn nghiên cứu, khảo sát các kỹ thuật và công cụ phân tích web log và triển khai thử nghiệm một công cụ quản lý và phân tích web log. Các hệ thống quản lý và phân tích web log có thể được sử dụng cho phát hiện các bất thường và hành vi truy cập của người dùng trong quản trị hệ thống và đảm bảo an toàn thông tin.

4. Đối tượng và phạm vi nghiên cứu:

Đối tượng nghiên cứu

Đối tượng nghiên cứu của luận văn là các dạng web log và các kỹ thuật, công cụ phân tích web log.

Phạm vi nghiên cứu

Phạm vi nghiên cứu của luận văn là giới hạn một số dạng web log.

5. Phương pháp nghiên cứu:

Luận văn sử dụng kết hợp các phương pháp nghiên cứu sau:

Phương pháp nghiên cứu lý thuyết

Khảo sát các kỹ thuật và công cụ phân tích web log.

Phương pháp nghiên cứu thực nghiệm

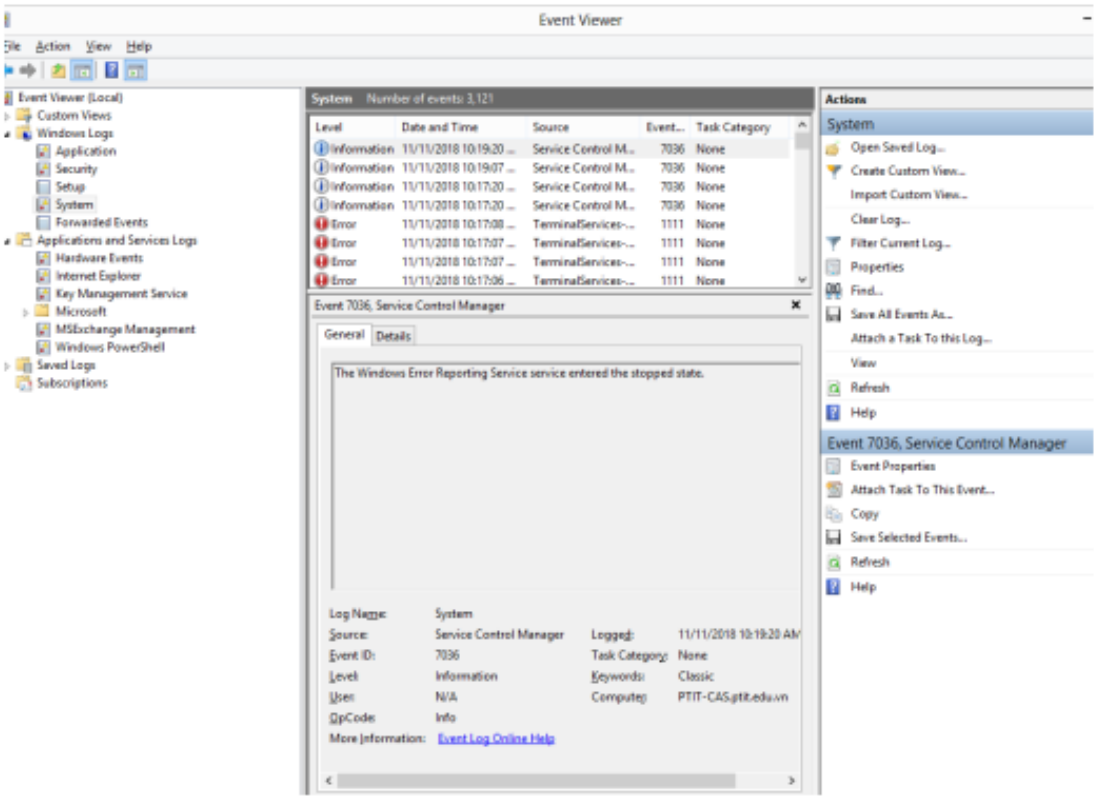
Triển khai thử nghiệm một hệ thống quản lý và phân tích log mã mở và đánh giá kết quả.

CHƯƠNG 1. TỔNG QUAN VỀ WEB LOG VÀ XỬ LÝ WEB LOG

1.1. Tổng quan về web log

1.1.1. Khái quát về web log

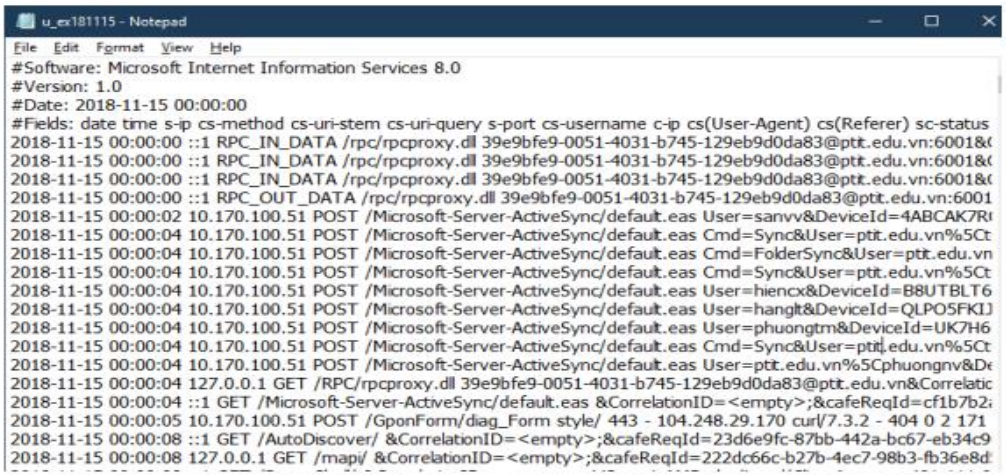
Nhật ký truy cập hay dấu vết truy cập, hay nhật ký (gọi tắt là log) là danh sách các bản ghi mà khi được yêu cầu truy cập tài nguyên hệ thống, hệ thống sẽ ghi lại. Ví dụ: nhật ký truy cập web (gọi tắt là nhật ký web log) chứa tất cả thông tin khi có yêu cầu truy cập tài nguyên của trang web.Tài nguyên của trang web có thể bao gồm các mẫu định dạng, tệp hình ảnh và tệp mã JavaScript. Khi người dùng truy cập trang web để tìm kiếm sản phẩm, máy chủ web sẽ tải xuống thông tin và hình ảnh của sản phẩm và nhật ký truy cập ghi lại các yêu cầu của người dùng đối với tài nguyên thông tin và ảnh của sản phẩm đó. Nhật ký web chứa các thông tin như tên người dùng, dấu thời gian, yêu cầu truy cập, địa chỉ IP, số byte được chuyển, trạng thái kết quả, URL.Các tệp nhật ký được duy trì bởi các máy chủ web.



Hình 1.1: Windows log sử dụng công cụ Event Viewer

Có nhiều nguồn tạo nhật ký trong hệ thống, chẳng hạn như nhật ký được tạo bởi máy chủ dịch vụ mạng, nhật ký do ứng dụng tạo, nhật ký do hệ điều hành tạo và nhật ký do thiết bị mạng và thiết bị đảm bảo an toàn thông tin tạo ra. Nhật ký được tạo bởi hệ điều hành thường bao gồm nhật ký của sự kiện khởi động hệ thống, sự kiện đăng nhập, sự kiện đăng xuất của người dùng, yêu cầu truy cập tệp và thư mục cũng như yêu cầu kích hoạt ứng dụng, yêu cầu truy cập phần cứng, yêu cầu truy cập dịch vụ mạng, lỗi xảy ra trong quá trình hoạt động, v.v. Ví dụ, hình 1.1 cho thấy hệ điều hành Microsoft Windows sử dụng công cụ Event Viewer và hệ điều hành Unix / Linux sử dụng công cụ Syslog để quản lý và lưu trữ nhật ký được tạo bởi chính hệ điều hành và các mô-đun phụ trợ của nó.

Nguồn nhật ký được tạo bởi máy chủ dịch vụ mạng, chẳng hạn như máy chủ web, máy chủ DNS, máy chủ email và máy chủ cơ sở dữ liệu là một trong những nguồn phổ biến nhất của việc tạo nhật ký. Máy chủ web có thể ghi nhật ký truy cập trang web cho mỗi trang web dưới dạng tệp văn bản thuần túy với một nhật ký trên mỗi dòng. Thông tin trong mỗi nhật ký có thể khác nhau tùy thuộc vào phiên bản máy chủ web được sử dụng. Hình 1.2 minh họa các bản ghi được tạo bởi máy chủ web Microsoft IIS. Máy chủ tên miền DNS cũng tạo ra một lượng lớn nhật ký trong quá trình xử lý các yêu cầu phân giải tên miền thành địa chỉ IP và ngược lại từ người dùng. Tương tự như vậy, các máy chủ cơ sở dữ liệu và email cũng tạo ra rất nhiều nhật ký trong khi xử lý các yêu cầu từ người dùng cũng như từ các ứng dụng.



Hình 1.2: Các bản ghi log sinh ra bởi máy chủ web Microsoft IIS

Các thiết bị và hệ thống mạng đảm bảo an toàn thông tin cũng như nguồn tạo ra nhật ký. Các thiết bị mạng thông thường như bộ định tuyến, bộ chuyển mạch và hệ thống bảo đảm an toàn thông tin, chẳng hạn như tường lửa, hệ thống kiểm soát truy cập, hệ thống phát hiện và ngăn chặn các cuộc tấn công, xâm nhập cũng tạo ra nhiều bản ghi log trong quá trình xử lý các yêu cầu truy cập mạng. Nhật ký được tạo từ các hệ thống này có thể được lưu trữ cục bộ hoặc xuất sang hệ thống lưu trữ bên ngoài.

Nhật ký log có thể được đặt ở ba nơi khác nhau:

▪ *Máy chủ Web* : Các nhật ký này thường cung cấp dữ liệu sử dụng đầy đủ và chính xác nhất, nhưng có hai yếu tố chính của chúng hạn chế là:

- Các nhật ký này chứa các thông tin nhạy cảm, thông tin cá nhân, do đó máy chủ thường giữ chúng.

- Các bản ghi log không ghi lại các trang đã truy cập vào bộ nhớ đệm. Các các trang đã lưu trong bộ nhớ đệm được triệu hồi từ bộ nhớ cục bộ của trình duyệt hoặc máy chủ proxy, không phải từ máy chủ web.

▪ *Máy chủ proxy web*: Máy chủ proxy lấy yêu cầu HTTP từ người dùng và chuyển chúng đến máy chủ Web sau đó trả lại cho người dùng các kết quả được máy chủ Web chuyển cho họ. Hai nhược điểm là:

- Xây dựng proxy-server là một nhiệm vụ khó khăn. Lập trình mạng nâng cao, chẳng hạn như TCP / IP, là cần thiết cho việc xây dựng này.

- Việc chặn yêu cầu bị hạn chế.

- Triển khai trình ghi proxy trong Web Quilt, một hiệu suất hệ thống ghi nhật ký web giảm nếu nó được tuyển dụng vì mỗi yêu cầu trang cần phải được xử lý bởi trình mô phỏng proxy.

▪ *Trình duyệt máy khách*: Người tham gia kiểm tra từ xa một trang Web bằng cách tải xuống phần mềm đặc biệt ghi lại việc sử dụng Web hoặc bằng cách sửa đổi mã nguồn của trình duyệt hiện có. Cookie HTTP cũng có thể được sử dụng cho mục đích này. Đó là những mẩu thông tin được tạo ra bởi một máy chủ web và được lưu trữ trong máy tính của người dùng, sẵn sàng để truy cập trong tương lai.

Hạn chế của phương pháp này là:

- Nhóm thiết kế phải triển khai phần mềm đặc biệt và yêu cầu người dùng cuối cài đặt nó.

- Kỹ thuật khó đạt được khả năng tương thích với một loạt hệ điều hành và các trình duyệt Web.

Như vậy, có thể thấy rằng có rất nhiều nguồn dữ liệu nhật ký truy cập với nhiều hình thức khác nhau. Tùy theo mục đích sử dụng mà người quản trị có thể cấu hình hệ thống để lựa chọn thu thập, quản lý và lưu trữ các thông tin cần thiết cho từng loại nhật ký.

1.1.2. Giới thiệu một số dạng web log

Nhật ký truy cập được tạo bởi hệ điều hành và các ứng dụng thường có định dạng riêng. Vì thử nghiệm trong luận văn này được thực hiện trên nhật ký web nên phần này giới thiệu các định dạng nhật ký web thường được sử dụng bao gồm định dạng nhật ký web chuẩn NCSA (NCSA Common Log Format), định dạng nhật ký web kết hợp (NCSA Combined Log Format)và định dạng nhật ký web mở rộng W3C (W3C Extended Log Format) và định dạng nhật ký web máy chủ web Microsoft IIS (Microsoft IIS Log Format). Trên thực tế, mọi máy chủ web hiện nay đều hỗ trợ một số định dạng nhật ký web này. Ví dụ, máy chủ web Microsoft IIS hỗ trợ cả ba định dạng,đó là : , W3C Extended Log Format , Microsoft IIS Log Format và NCSA Common Log Format . Ngược lại, máy chủ web Apache hay Apache HTTP Server sử dụng chuỗi định dạng để hỗ trợ 2 định dạng nhật ký bao gồm: Định dạng nhật ký web chuẩn NCSA và Định dạng nhật ký web kết hợp NCSA. Quản trị viên có thể chọn định dạng nhật ký web để sử dụng cho máy chủ để tạo tệp nhật ký web.

NCSA Common Log Format

Định dạng nhật ký chuẩn NCSA, hay thường được gọi là , là một định dạng tệp nhật ký dựa trên văn bản ASCII với các trường cố định, vì vậy nó không thể được tùy chỉnh. Các trường phân vùng trong mỗi nhật ký được phân tách bằng dấu cách. Các trường không chứa dữ liệu sẽ được biểu diễn bằng dấu (-), các ký tự

không in được sẽ được biểu diễn bằng dấu (+). Định dạng nhật ký web này ghi lại thông tin cơ bản mà người dùng yêu cầu như sau:

- Địa chỉ máy chủ từ xa
- Tên nhật ký từ xa (Giá trị này luôn là dấu gạch ngang.)
- Tên tài khoản
- Chênh lệch ngày, giờ và giờ trung bình Greenwich (GMT)
- Yêu cầu và phiên bản giao thức
- Mã trạng thái HTTP trả về (Giá trị 200 cho biết rằng yêu cầu đã được thực hiện thành công.)
- Số byte được gửi đến máy chủ

Với Máy chủ Apache HTTP, Định dạng Nhật ký web chuẩn có thể được định cấu hình bằng chuỗi định dạng sau:

LogFormat "%h %l %u %t \"%r\" %>s %b" common CustomLog logs/access_log common

Ví dụ, với Common Log Format thì một đầu mục (entry) sẽ có dạng như sau:

216.67.1.91 - leon [01/Jul/2002:12:11:52 +0000] "GET /index.html HTTP/1.1" 200 431

Trong đó, các trường thông tin của đầu mục này gồm:

- 216.67.1.91 (tương ứng kí hiệu %h): Địa chỉ IP của máy khách gửi yêu cầu đến máy chủ.
- Trống (-) (tương ứng kí hiệu %l): Định danh của máy khách.
- leon (tương ứng kí hiệu %u): Định danh/tên của người dùng gửi yêu cầu được xác định nhờ thủ tục xác thực HTTP.
- [01/Jul/2002:12:11:52 +0000] (tương ứng kí hiệu %t): Thời gian máy chủ kết thúc xử lý yêu cầu, theo định dạng sau:

[day/month/year:hour:minute:second:zone],

nghĩa là: [ngày/tháng/năm:giờ:phút:giây:múi giờ].

Trong đó, day = 2*digit, month = 3*letter; year = 4*digit; hour = 2*digit; minute = 2*digit; second = 2*digit và zone = ('+' | '-') 4*digit.

▪ “GET /index.html HTTP/1.1” (tương ứng kí hiệu \”%r”\”): Yêu cầu của máy khách gửi lên máy chủ.

▪ 200 (tương ứng kí hiệu %>s): Mã trạng thái mà máy chủ gửi trả về cho máy khách.

▪ 431 (tương ứng kí hiệu %b): Kích thướcicủa gói tin trả về cho máy khách, không bao gồm header.

NCSA Combined Log Format

Định dạng nhật ký kết hợp NCSA được viết tắt là Combined Log Format về cơ bản giống với Định dạng nhật ký chuẩn Common Log Format, ngoại trừ nó có thêm hai trường thông tin bổ sung ở cuối là Referrer (Liên kết tham chiếu) và User Agent(Máy khách người dùng). Với Apache HTTP Server, định dạng này có thể được cấu hình bằng cách sử dụng chuỗi định dạng như sau:

LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-agent}i\"" combined CustomLog log/acces_log combined

Ví dụ, một đầu mục của định dạng Combined Log Format sẽ như sau:

127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] “GET /apache_pb.gif HTTP/1.0” 200 2326 “http://www.example.com/start.html” “Mozilla/4.08 [en] (Win98;I;Nav)”

Các trường được bổ sung bao gồm:

▪ http://www.example.com/start.html (tương ứng kí hiệu \”%{Referer}i\”): Cho biết trang web người dùng đã ghé thăm trước khi đến trang hiện tại.

▪ Mozilla/4.08 [en] (Win98; I ;Nav)” (tương ứng kí hiệu \”%{User- agent}i\”): Cho biết trình duyệt và hệ điều hành máy khách đang sử dụng.

W3C Extended Log Format

Hiện tại, Định dạng nhật ký mở rộng W3C Extended Log Format là định dạng được sử dụng rộng rãi nhất và được hầu hết các máy chủ web hỗ trợ , định dạng này do Tổ chức World Wide Web Consortium (W3C) đề xuất. Định dạng web log này có các khả năng:

▪ Hỗ trợ kiểm soát các thông tin sẽ được ghi trong web log.

- Hỗ trợ định dạng web log chung cho máy khách, máy chủ web và cả proxy.
- Cung cấp một cơ chế mạnh mẽ để xử lý các vấn đề và các ký tự thoát (character escaping).
- Cho phép trao đổi dữ liệu nhân khẩu học (demographic).
- Hỗ trợ tổng hợp dữ liệu.

Tệp nhật ký ở định dạng Nhật ký mở rộng W3C Extended Log chứa một tập hợp các dòng văn bản thuần túy bao gồm các ký tự ASCII (hoặc Unicode) tiêu chuẩn được phân tách bằng dấu xuống dòng (LF hoặc CRLF). Các Web log này có thể được tùy chỉnh bởi người quản trị viên, có thể thêm hoặc bớt các trường tùy thuộc vào thông tin muốn ghi lại.

Ví dụ về một định dạng W3C log như sau:

```
#Software: Microsoft Internet Information Services 7.5
#Version: 1.0
#Date: 2012-12-05 08:25:10
#Fields: 1998-11-19 22:48:39 206.175.82.5 - 208.201.133.173
GET/global/images/navlineboards.gif – 200 540 324 157
HTTP/1.0 Mozilla/4.0+(compatible;+MSIE+4.01;+Windows+95)
USERID=CustomerA;+IMPID=01234
http://www.loganalyzer.net
```

Trong đó,

#software - phiên bản IIS đang chạy

#version - định dạng tệp nhật ký

Date - ghi ngày và giờ của lần nhập nhật ký đầu tiên.

#Fields: ngày giờ - Địa chỉ IP máy khách – Địa chỉ IP máy chủ - Phương thức -uristem cs-uri-query sc-status sc-byte cs-byte time-time cs version cs (Tác nhân người dùng) cs (Cookie) cs (Liên kết giới thiệu)

Các file log khác nhau sẽ có ký tự kết thúc dòng khác nhau tùy thuộc vào quy ước kết thúc dòng của nền tảng hoạt động. Trên mỗi dòng thường có một chỉ thị (directive) hoặc một đầu mục (entry). Phần tiếp theo mô tả chi tiết về 2 thành phần này.

- *Các chỉ thị:*

Chỉ thị là các dòng chứa ký tự bắt đầu là ký tự "#". Chúng chứa các thông tin mô tả về file log. Định dạng W3C Extended Log bao gồm các chỉ thị như sau:

▪ Version: <interger>.<interger>: Phiên bản của định dạng log được sử dụng.

▪ Fields: [<specifier>...]: Chỉ ra danh sách các trường được ghi lại trong tệp nhật ký log.

▪ Software: string: Chỉ ra phần mềm tạo ra log:

▪ Start-Date: <date> <time>: Ngày và giờ bắt đầu ghi log.

▪ End-Date: <date> <time>: Ngày và giờ kết thúc ghi log.

▪ Date: <date> <time>: Ngày và giờ thêm vào các đầu mục trong log.

▪ Remark: <text>: Các thông tin chú thích. Các công cụ phân tích nhật ký log thông thường sẽ bỏ qua dữ liệu trong trường này.

Các chỉ thị Version và Fields là bắt buộc và đứng trước tất cả các trường khác trong file log. Chỉ thị Fields đưa ra một danh sách định danh của trường, xác định thông tin được ghi trong mỗi đầu mục. Các định danh trường có thể là một trong số các kiểu sau: tên nhận dạng(Identifier), tiền tố tên nhận dạng(Prefix-identifier) và tiền tố (đề mục)(Prefix (header)).

Bảng 1.1: Danh sách các tiền tố

Tiền tố	Ý nghĩa
c	Client
s	Server
r	Remote
s	Client đến Server
sc	Server đến Client
sr	Server đến Remote Server (được dùng bởi proxy)
rs	Remote Server đến Server (được dùng bởi proxy)
x	Định danh riêng của ứng dụng

Bảng 1.2: Các định danh không yêu cầu có tiền tố

Định danh	Ý nghĩa
date	Ngày giao dịch hoàn thành, kiểu <date>
time	Thời gian (giờ) giao dịch hoàn thành, kiểu <time>
time- taken	Thời gian để giao dịch được hoàn thành tính bằng giây, kiểu <fixed>
bytes	Số byte đã truyền, kiểu <integer>
cached	Ghi lại số lần cache hit, nếu bằng 0 thì tức là cache miss, kiểu <integer>

Bảng 1.1 cung cấp danh sách các tiền tố (Prefix) cho các định danh, Bảng 1.2 đưa ra danh sách các định danh không yêu cầu về tiền tố và Bảng 1.3 cung cấp danh sách các định danh bắt buộc phải có tiền tố. Ví dụ, định danh cs-method cho biết phương thức của gói tin gửi đi bởi client đến server, định danh c-ip cho biết địa chỉ IP của client ,sc(Referer) cho biết trường referer trong gói tin trả lời.

Bảng 1.3: Các định danh cần phải có tiền tố

Định danh	Ý nghĩa
ip	Địa chỉ IP và cổng, kiểu <address>
dns	Tên DNS, kiểu <name>
status	Mã trạng thái, kiểu <interger>
comment	Mô tả trạng thái trả về của mã trạng thái, kiểu <text>
method	Method, kiểu <name>
uri	URL, kiểu <uri>
uri-stem	Phần thân của URL (bỏ qua phần truy vấn), kiểu <uri>
uri- query	Phần truy vấn của URI, kiểu <uri>
host	DNS hostname được sử dụng, kiểu <name>

Các đầu mục

Mỗi đầu mục là một chuỗi các trường được liên kết với một giao dịch HTTP, bao gồm một loạt các trường được phân tách bằng dấu cách hoặc ký tự tab, không chứa ký tự điều khiển ASCII và kết thúc bằng CR hoặc CRLF. Các chỉ thị #Fields cho biết ý nghĩa của các trường và nếu một trường không có thông tin trong đầu mục, nó sẽ được hiển thị dưới dạng ký tự "-". Bảng 1.4 mô tả các định dạng dữ liệu được sử dụng trong Định dạng mở rộng W3C Extended Format và Bảng 1.5 liệt kê các ký trường có sẵn trong định dạng nhật ký web này.

Bảng 1.4. Các định dạng dữ liệu sử dụng trong W3C Extended Format

Định dạng dữ liệu	Mô tả
Integer	Định dạng: <integer> = 1*<digit> Trong đó, một số integer được biểu diễn như là một dãy các chữ số
Fixed Format	Định dạng: <fixed> = 1*<digit> [. *<digit>]
URI	Theo chuẩn RCF 1738 và không được phép chứa khoảng trắng hay ký tự điều khiển ASCII.

Định dạng dữ liệu	Mô tả
Date	Định dạng: <date> = 4<digit> "-" 2<digit> "-" 2<digit> Ngày tháng năm được ghi với định dạng YYYY-MM-DD. Với YYYY, MM, DD tương ứng là năm, tháng và ngày. Lựa chọn định dạng này giúp sắp xếp dễ dàng hơn.
Time	Định dạng: <time> = 2<digit> ":" 2<digit> [":" 2<digit> ["." *<digit>] Thời gian được ghi với định dạng HH:MM, HH:MM:SS hoặc HH:MM:SS.S với HH là giờ từ 00-23, MM là phút, SS là giây.
String	Định dạng: <string> = "" *<schar> "". Với <schar> = xchar |..... Các string được đặt trong dấu ngoặc kép, nếu một string chứa dấu ngoặc kép cũng không gây khó hiểu bởi vì các trường được phân tách bởi khoảng trắng.
Text	Định dạng: <text> = <char>* Trường text được sử dụng bởi các chỉ thị.
Address	Định dạng: <name> = <integer> ["." *<integer>] [":" <integer>] Địa chỉ IP và port (trường port là tuỳ chọn).

Bảng 1.5. Các trường khả dụng trong W3C Extended Format

Trường	Tên trong file log	Mô tả
Date	date	Ngày giao dịch xảy ra
Time	time	Thời gian giao dịch xảy ra (UTC)
Service Name and Instance Number	s-sitename	Tên dịch vụ và số tiến trình chạy
Server Name	s-computername	Tên của server được tạo trong tệp tin log

Trường	Tên trong file log	Mô tả
Server IP Address	s-ip	Địa chỉ của server được tạo trong tệp tin log
Method	cs-method	Là phương thức yêu cầu, ví dụ như phương thức GET
URI Stem	cs-uri-stem	Đối tượng mục tiêu của phương thức, ví dụ như Default.html
URI Query	cs-uri-query	Universal Resource Identifier, được dùng trong các trang động
Server Port	s-port	Cổng trên server mà đã được cấu hình cho dịch vụ.
User Name	cs-name	Tên của người dùng hợp lệ đã truy cập vào server. Người dùng ẩn danh thì được biểu diễn bởi dấu "-".
Client IP Address	c-ip	Địa chỉ IP của máy khách đã gửi yêu cầu
Protocol Version	cs-version	Phiên bản giao thức HTTP được máy khách sử dụng.
User Agent	cs(User-Agent)	Trình duyệt mà máy khách đã sử dụng
Cookie	cs(Cookie)	Nội dung của cookie được gửi hoặc nhận, nếu có.
Referrer	cs(Referrer)	Trang web mà người dùng truy cập lần cuối, trang này cung cấp một đường link đến trang web hiện tại.
Host	cs-host	Host header name, nếu có
HTTP Status	sc-status	Mã trạng thái HTTP
Protocol Substatus	sc-substatus	Mã trạng thái phụ giao thức
Win32 Status	sc-win32-status	Mã trạng thái Windows

Trường	Tên trong file log	Mô tả
Bytes Sent	sc-bytes	Số lượng byte được gửi bởi server
Bytes Received	cs-bytes	Số lượng byte nhận và xử lý bởi server
Time Taken	time-taken	Độ dài khoảng thời gian diễn ra hành động (mili giây)

Microsoft IIS Log Format

Microsoft IIS là một máy chủ web chạy trên hệ điều hành Microsoft Windows Server. Như đã đề cập, IIS hỗ trợ nhiều định dạng nhật ký web khác nhau như: Định dạng nhật ký web chuẩn NCSA Common Log Format , Định dạng nhật ký web mở rộng W3C Extended Log Format, định dạng nhật ký Microsoft IIS Log Format.

Định dạng tệp nhật ký IIS là định dạng dựa trên văn bản ASCII cố định định dạng, vì vậy không thể tùy chỉnh nó.

Định dạng tệp nhật ký IIS ghi lại dữ liệu sau:

- Địa chỉ IP của máy khách
- Tên tài khoản
- Ngày tháng
- Thời gian
- Dịch vụ và phiên bản
- Tên máy chủ
- Địa chỉ IP máy chủ
- Mất thời gian
- Đã gửi các byte khách hàng
- Đã gửi byte máy chủ
- Mã trạng thái dịch vụ (Giá trị 200 cho biết rằng yêu cầu đã được thực hiện thành công.)
- Mã trạng thái Windows (Giá trị 0 cho biết rằng yêu cầu đã được thực hiện thành công.)

- Loại yêu cầu
- Mục tiêu hoạt động

Các trường trong mỗi nhật ký được phân tách bằng dấu phẩy, các trường không chứa thông tin được thay thế bằng dấu '-' và các ký tự không in được thay bằng dấu '+'. Ví dụ, đầu mục của nhật ký web log sẽ trông giống như sau với định dạng Microsoft IIS Log Format:

```
192.168.114.201, -, 03/20/01, 7:55:20, W3SVC2, SALES1, 172.21.13.45,
4502, 163, 3223, 200, 0, GET, /DeptLogo.gif, -,
```

Trong đó:

- 192.168.114.201 là địa chỉ IP máy khách
- 03/20/01, 7:55:20 là ngày và giờ thực hiện yêu cầu
- W3SVC2 chỉ tiến trình chạy dịch vụ web
- SALES1 là tên máy chủ web
- 172.21.13.45 là địa chỉ IP máy chủ web
- 4502 là thời gian xử lý tính bằng mili giây
- 163 là số byte của yêu cầu
- 3223 là số byte phản hồi (kết quả) của máy chủ gửi đến máy khách
- 200 là mã trạng thái thực hiện yêu cầu (thành công)
- GET là phương thức yêu cầu
- /DeptLogo.gif là file được yêu cầu.

1.2. Tổng quan về xử lý web log

1.2.1. Khái quát về xử lý web log

Hệ thống phân tích nhật ký log bao gồm ba bước cơ bản: thu thập, xử lý và phân tích nhật ký log. Hình 1.3 mô tả các bước của quá trình thu thập, xử lý và phân tích nhật ký log được áp dụng trong thực tế.Các bước xử lý cụ thể của quá trình bao gồm:

Hình 1.3: Các khâu của quá trình thu thập, xử lý và phân tích log

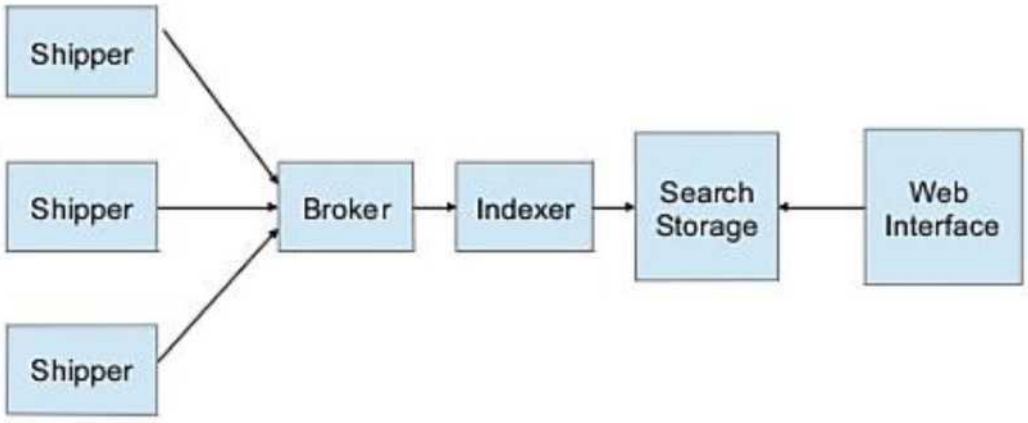
- *Thu thập dữ liệu* : Đây là bước đầu tiên trong quá trình thu thập, xử lý và phân tích nhật ký log. Thu thập dữ liệu nhật ký log là việc thu thập các bản ghi nhật ký thô từ các nguồn tạo nhật ký log và chuyển chúng đến một trung tâm xử lý.

- *Làm sạch dữ liệu*: Các bản ghi log thô có thể bao gồm một số dữ liệu vô dụng không mong muốn, không có gì để làm với thủ tục khai thác. Đây là khâu để loại bỏ những mục không liên quan hoặc dư thừa ra khỏi tệp nhật ký. Các bản ghi log thô được làm sạch để giảm bớt dữ liệu nhiễu. Khi những dữ liệu không mong muốn này được loại bỏ, kích thước của tệp được giảm thiểu đến một mức độ lớn hơn.

- *Định dạng dữ liệu* : Đây là bước chuẩn hóa dữ liệu log. Nhật ký log được tạo ra từ nhiều nguồn khác nhau với nhiều định dạng khác nhau. Do đó, nhật ký log cần được chuẩn hóa theo định dạng yêu cầu và chuyển đổi sang dạng có cấu trúc bằng các thuật toán khai thác dữ liệu. Bước này tạo dữ liệu nhật ký làm đầu vào để phân tích nhật ký.

- *Phân tích dữ liệu:* Đây là bước quan trọng nhất trong quá trình phân tích nhật ký log. Phân tích nhật ký log là việc trích xuất thông tin quan trọng và đưa ra kết luận về trạng thái bảo mật từ nhật ký thống kê. Phân tích nhật kýlog ứng dụng để bảo mật thông tin và các ứng dụng khác.

- *Kết xuất kết quả*: Đây là khâu kết xuất kết quả ra giao diện của người dùng.



Hình 1.4: Kiến trúc điển hình của hệ thống thu thập, xử lý và phân tích log

Hình 1.4 mô tả một kiến trúc điển hình của một hệ thống thu thập, xử lý và phân tích nhật ký log. Hệ thống gồm các thành phần như sau:

- *Shipper* là mô-đun thu thập nhật ký log từ các nguồn tạo nhật ký log khác nhau. Các Shipper thường được cài đặt trên hệ thống được giám sát. Shipper chỉ cần thu thập nhật ký thô và gửi lại cho Broker hoặc thực hiện các công việc làm sạch và chuẩn hóa dữ liệu nhật ký log.

- *Broker* là mô đun nhận dữ liệu nhật ký log từ nhiều nguồn. Sau khi nhận được, dữ liệu nhật ký log được làm sạch, chuẩn hóa và chuyển tiếp sang giai đoạn tiếp theo.

- Indexer là một mô-đun lập chỉ mục cho dữ liệu nhật ký. Lập chỉ mục là một bước quan trọng để tìm kiếm và trích xuất dữ liệu nhật ký trong bước tiếp theo.

- *Search & Strorage* là mô đun cuối cùng cho phép tìm kiếm, trích xuất các dữ liệu log quan trọng và lưu trữ dữ liệu.

- *Web Interface* là giao diện người dùng dựa trên nền web cho hệ thống quản lý và phân tích nhật ký log.

1.2.2. Ứng dụng của xử lý web log

Phân tích nhật ký truy cập thường được thực hiện cho các mục đích sau:

- (1) Đảm bảo an toàn thông tin hệ thống
- (2) Hỗ trợ khắc phục sự cố hệ thống
- (3) Hỗ trợ điều tra kỹ thuật số
- (4) Hỗ trợ hiểu hành vi của người dùng trực tuyến.

Để đảm bảo an toàn thông tin cho hệ thống là một trong những mục đích chính của phân tích nhật ký truy cập. Có thể thấy, phân tích nhật ký truy cập có thể hỗ trợ giám sát và kiểm tra việc tuân thủ các chính sách bảo mật và chính sách kiểm toán của các cơ quan, tổ chức. Hơn nữa, phân tích nhật ký truy cập có thể hỗ trợ ứng phó với các sự cố an toàn thông tin bằng cách hỗ trợ xác định các nguyên nhân và yếu tố bảo mật. Theo dõi log truy cập có thể cho thấy các hành vi độc hại, điều này có thể giúp xác định các cách cải thiện bảo mật và ngăn chặn các cuộc tấn công. Khi một trang web bị vi phạm bảo mật hoặc bị tấn công, nhật ký truy cập sẽ hiển thị chính xác tất cả dữ liệu được yêu cầu.Nhiều công cụ an toàn thông tin dựa trên nhật ký giám sát, thu thập, xử lý và phân tích log đã được nghiên cứu, phát triển và triển khai trên thực tế như IBM QRadar SIEM, VNCS Web Monitoring và hệ thống. Hệ thống phát hiện xâm nhập OSSEC. Các công cụ này sẽ giám sát và thu thập các loại nhật ký log được tạo ra bởi hệ điều hành, bởi các dịch vụ và bởi các ứng dụng trong hệ thống được giám sát để phát hiện hành vi bất thường và các cuộc tấn công, xâm nhập.

Việc phân tích log truy cập giúp quản trị hệ thống hỗ trợ khắc phục sự cố hệ thống . Các vấn đề khắc phục sự cố thường bắt đầu bằng việc xem qua log truy cập và nhật ký lỗi. Vì log truy cập chứa một mục hiển thị trạng thái của từng yêu cầu đến, nó giúp làm sáng tỏ những vấn đề cần chú ý ở đâu. Phân tích nhật ký truy cập giúp loại bỏ dữ liệu nhiễu, tổng hợp các thông báo lỗi riêng lẻ, giúp xác định nguyên nhân gây ra sự cố hệ thống rõ ràng và chính xác hơn, trên cơ sở đó, người quản trị có thể đưa ra biện pháp khắc phục phù hợp.

Phân tích nhật ký truy cập cũng có thể hỗ trợ điều tra số thông qua truy tìm và xâu chuỗi các sự kiện nhật ký riêng lẻ bằng cách sử dụng kỹ thuật khai thác dữ liệu và phân tích tương quan. Nó cung cấp cho người dùng khả năng phân tích khối lượng lớn của dữ liệu luồng nhấp chuột hoặc luồng nhấp chuột, tích hợp liền mạch dữ liệu, với các bản dịch và dữ liệu nhân khẩu học từ nguồn ngoại tuyến ..Theo dõi nhật ký truy cập cũng có thể cung cấp thông tin chi tiết về kỹ thuật SEO của trang web bằng cách hiển thị quá nhiều mã lỗi HTTP hoặc chuyển hướng và xem cách

trình thu thập thông tin tìm kiếm truy cập các trang. Quá nhiều chuyển hướng hoặc lỗi có thể cản trở hiệu suất SEO tổng thể của trang web. Việc xem các mẫu thu thập thông tin có thể cảnh báo cho chủ sở hữu trang web về cách các công cụ tìm kiếm xem nội dung trang web và xác định các khu vực cần cải thiện cho các trang cụ thể. Từ đó, dựa vào kết quả phân tích log có thể sử dụng để tạo các bằng chứng số cho các sự cố mất an toàn thông tin.

Hỗ trợ sự hiểu biết về hành vi của người dùng trực tuyến là một trong những mục đích chính của phân tích nhật ký truy cập, đặc biệt là phân tích trang web hoặc nhật ký web. Phân tích nhật ký web có thể tạo báo cáo sử dụng các trang web của người dùng, bao gồm lưu lượng truy cập, các trang tham chiếu, phân bố địa lý của người dùng và tải xuống dữ liệu. Đồng thời, phân tích nhật ký truy cập cũng giúp rút ra nhiều thông tin quan trọng về hành vi người dùng trực tuyến và trên cơ sở đó có thể hỗ trợ tối ưu hóa website, nhằm nâng cao chất lượng dịch vụ cung cấp và trải nghiệm người dùng. Bằng cách xác định hành vi truy cập của người dùng, các liên kết cần thiết có thể được xác định để cải thiện tổng thể hiệu suất của các quyền truy cập trong tương lai. Cá nhân hóa cho người dùng có thể đạt được bằng các theo dõi các trang đã truy cập trước đó. Các trang này có thể được sử dụng để xác định các hành vi duyệt web của người dùng và sau đó dự đoán các trang mong muốn. Để hỗ trợ sự hiểu biết về hành vi của người dùng trực tuyến ,một số công cụ phân tích log được phát triển và triển khai có thể liệt kê bao gồm: Sumo Logic, Logstash, Graylog và Webalizer.

1.3. Kết luận chương

Chương 1 giới thiệu tổng quan về web log, một số định dạng của web log, bao gồm các dạng Apache web log, Microsoft IIS log. Chương này cũng giới thiệu về vấn đề phân tích web log và các ứng dụng của phân tích web log.

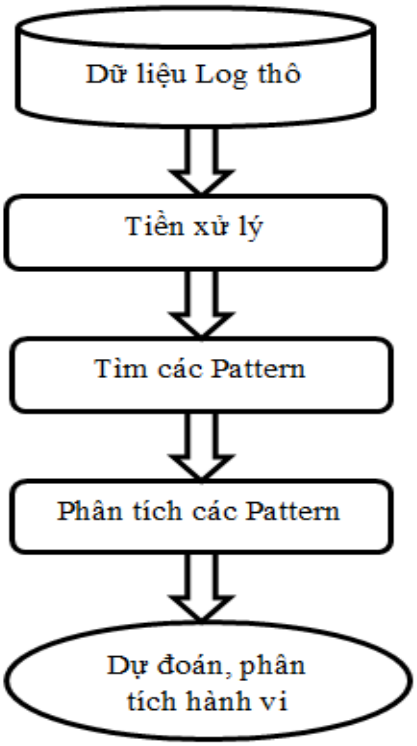
CHƯƠNG 2. CÁC KỸ THUẬT VÀ CÔNG CỤ PHÂN TÍCH WEB LOG

2.1. Các kỹ thuật phân tích web log

2.1.1. Mô hình xử lý web log

Hình 2.1 mô tả mô hình hệ thống xử lý web log điển hình. Theo đó, một hệ thống xử lý web log sẽ phải thực hiện các bước sau:

- Bước tiền xử lý và chuẩn hóa dữ liệu log
- Bước tìm các pattern
- Bước phân tích pattern
- Bước dự đoán, phân tích hành vi người dùng



Hình 2.1: Mô hình xử lý web log

Bước tiền xử lý và chuẩn hóa dữ liệu log

Trong bước này, hệ thống nhận dữ liệu nhật ký thô từ các nguồn khác nhau, trích xuất thông tin cần thiết và đưa nó về một định dạng thống nhất. Ngoài ra, giai đoạn này còn có nhiệm vụ tiền xử lý một số thông tin như: người dùng, phiên làm việc…Giai đoạn này bao gồm các bước: Làm sạch và gộp dữ liệu; nhận dạng người dùng; nhận dạng phiên; nhận dạng số lần xem trang pageview; hoàn thành đường dẫn.

Bước tìm các pattern

Giai đoạn này sử dụng các phương pháp và thuật toán như thống kê, học máy, khai thác dữ liệu, nhận dạng mẫu để xác định các mẫu người dùng. Trong phân tích nhật ký, các mẫu pattern cơ bản cần được xác định bao gồm: Các trang web yêu thích, thời gian xem trung bình trên mỗi trang, các lĩnh vực quan tâm… Trong giai đoạn này, các kỹ thuật phân tích dữ liệu có thể được sử dụng. dữ liệu như: phân tích thống kê; phân cụm; phân lớp; luật kết hợp; các mẫu tuần tự hoặc mô hình hóa phụ thuộc.

Bước phân tích pattern

Giai đoạn này chịu trách nhiệm phân tích các mẫu pattern được tìm thấy trong giai đoạn trước, xác định các mẫu pattern không có nhiều giá trị và loại bỏ chúng khỏi quá trình phân tích nhật ký. Giai đoạn này được thực hiện bởi các truy vấn SQL, sử dụng phân tích xử lý trực tuyến hoặc cũng bằng các kỹ thuật trực quan hóa dữ liệu để lọc và phân tích các mẫu pattern.

Bước dự đoán, phân tích hành vi người dùng

Sau khi phân tích và lọc các mẫu pattern, các mẫu pattern còn lại sẽ được sử dụng để đưa ra kết luận về hành vi của người dùng như: Các trang web thường truy cập, lĩnh vực quan tâm, thời gian trung bình đã xem mỗi trang web. Pha này có thể đưa ra các báo cáo thống kê, các biểu đồ hình vẽ về hành vi của một người dùng cụ thể hoặc tổng quan của cả trang web.

2.1.2. Thu thập và tiền xử lý

Thu thập web log

Nhật ký web có thể được tạo tại nhiều vị trí khác nhau trong mạng, vì vậy có nhiều cách để thu thập nhật ký web. Nhật ký web có thể được nhận từ nhiều nguồn

khác nhau như: từ tệp, từ Internet hoặc từ đầu ra của các ứng dụng khác.. Một số nguồn cụ thể có thể kể ra như:

- Nhận các sự kiện từ framework Elastic Beats.
- Đọc các kết quả truy vấn từ một cụm Elasticsearch.
- Lấy các sự kiện từ file log.
- Nhận đầu ra của các công cụ dòng lệnh như là một sự kiện.
- Tạo các sự kiện dựa trên các bản tin SNMP.
- Đọc các bản tin syslog.
- Đọc sự kiện từ một TCP socket.
- Đọc sự kiện thông qua UDP.
- Đọc sự kiện thông qua một UNIX socket.

Nhật ký có thể được lưu trên chính hệ thống hoặc chuyển sang hệ thống khác. Quá trình chuyển các bản ghi được tạo trong tất cả các hệ thống đến một môi trường duy nhất được gọi là lưu trữ nhật ký. Tuy nhiên, khi kết quả được phân tích, tất cả các sự cố máy tính được ghi lại trong hình thức của một số lượng lớn các đống đã làm cho việc điều tra tội phạm có chủ đích hoặc sai sót trở nên rất phức tạp.

Việc thu thập web log gặp khó khăn vì những lý do sau:

- Nhật ký được tạo ra từ nhiều hệ thống với số lượng và kích thước lớn,
- Tạo các loại nhật ký khác nhau từ các hệ thống khác nhau,
- Nội dung nhật ký khác xa nhau.

Trong phạm vi luận văn này ta sử dụng phương pháp lấy các sự kiện để xử lý từ file log, ví dụ như file log của Apache server hay IIS server.

Tiền xử lý

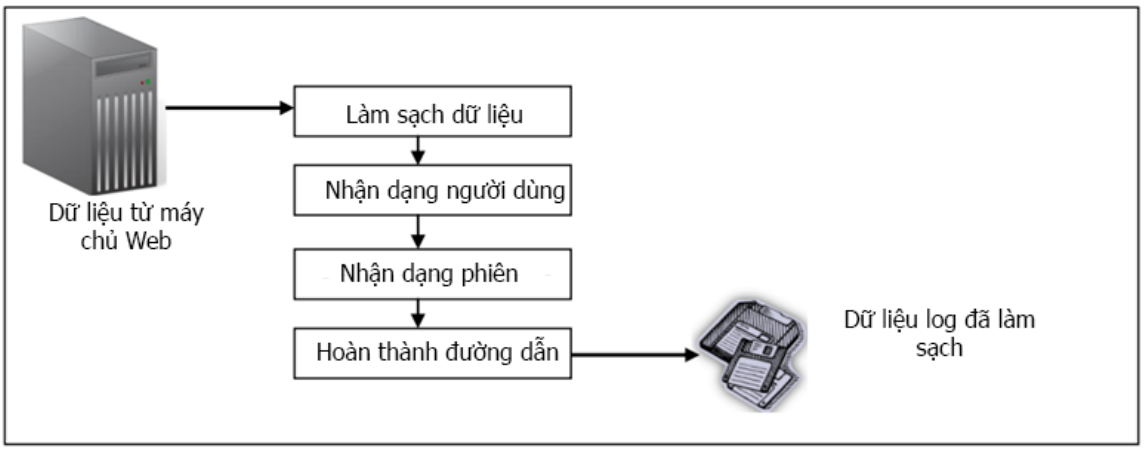
Thông tin được truy cập thông qua web là không đồng nhất và bán cấu trúc hoặc không cấu trúc về bản chất. Do sự không đồng nhất này, một tệp nhật ký web có thể bao gồm một số các mục nhật ký không mong muốn, mà sự hiện diện của chúng không quan trọng để khai thác sử dụng web. Điều này làm cho xử lý trước tệp nhật ký, một điều kiện tiên quyết quan trọng đối với khám phá các mô hình hiểu biết. Mục đích của tiền xử lý là chuyển đổi dữ liệu luồng nhấp chuột thô thành bộ

hồ sơ người dùng . Xử lý trước dữ liệu trình bày một số những thách thức độc đáo dẫn đến nhiều thuật toán và kỹ thuật heuristic để xử lý trước các tác vụ như hợp nhất và làm sạch, nhận dạng người dùng và phiên, …

Tiền xử lý dữ liệu là một trong những giai đoạn phức tạp nhất của quy trình khai thác sử dụng web.

Tiền xử lý dữ liệu bao gồm bốn giai đoạn phụ :

- Làm sạch dữ liệu
- Nhận dạng người dùng
- Nhận dạng phiên
- Hoàn thành đường dẫn



Hình 2.2: Các nhiệm vụ của tiền xử lý dữ liệu log

Làm sạch dữ liệu

Trong quá trình này, tệp nhật ký web có thể bao gồm một số dữ liệu vô dụng không mong muốn nhất định không có gì để làm với thủ tục khai thác.Ví dụ có thể kể đến như: hình ảnh, đồ họa, đa phương tiện…. Do đó, bắt buộc phải loại bỏ những mục không liên quan khỏi tệp nhật ký. Khi những dữ liệu này được loại bỏ, kích thước của tệp nhật ký được giảm thiểu khá nhiều. Có ba loại dữ liệu không liên quan hoặc dư thừa cần thiết để làm sạch:

- Tài nguyên phụ trợ được nhúng trong tệp HTML: Tài nguyên phụ trợ không gì khác ngoài đồ họa, tập lệnh hoặc một số video, có thể xuất hiện trong một trang HTML có thể không có mối quan hệ với nội dung của trang html trên mà

chúng được nhúng. Đúng hơn họ có thể là một phần của một số quảng cáo. Khi người dùng yêu cầu một trang web cụ thể, những thứ này cũng có thể được tải xuống cùng với tệp HTML và tạo thành một số mục nhật ký. Như đã thảo luận trước đó, mục tiêu của khai thác sử dụng Web là để nắm bắt hành vi của người dùng; vì thế những mục này như đồ họa, hình ảnh và tập lệnh là vô dụng. Vì lý do này, việc loại bỏ các mục không liên quan này dường như cần thiết.

▪ Các yêu cầu của rô bốt: Robot web là công cụ phần mềm được sử dụng để tự động trích xuất nội dung của một trang Web bằng cách theo dõi tất cả các siêu liên kết từ một trang Web. Các công cụ tìm kiếm như Google định kỳ sử dụng các trình thu thập dữ liệu để lấy tất cả các trang từ một trang web để cập nhật các chỉ mục tìm kiếm của họ. Các chỉ mục này cũng không quan trọng từ quan điểm khai thác và do đó phải loại bỏ.

▪ Các yêu cầu lỗi: Tất cả các truy cập và lỗi xảy ra trong quá trình truy cập đến trang Web đều được ghi lại trong các tập nhật ký log. Các yêu cầu lỗi này không có tác dụng với việc phân tích, nên chúng cần phải được loại bỏ để giảm thiểu kích thước tập nhật ký log.

Dưới đây là một thuật toán được sử đụng để làm sạch tệp nhật ký web để truy xuất thông tin hữu ích và loại bỏ dữ liệu không cần thiết. Đầu vào là tệp nhật ký web thô để xử lý và cuối cùng đầu ra được tạo là tệp nhật ký web đã xử lý và dữ liệu của nó là được chèn vào bảng của cơ sở dữ liệu.

Input: tệp nhật ký web thô.

Output: tệp nhật ký web đã xử lý.

Bước 1. For với mỗi dòng trong tệp nhật ký web do

Bước 2. nếu độ dài của dòng nhiều hơn thì một ký tự thì

Tránh Dòng Trống

Bước 3. nếu dòng không bắt đầu bằng ‘#’ thì #Avoid

Bình luận

Bước 4. nếu tên liên kết chứa tên miền thì #Consider

Chỉ các liên kết dành riêng cho ứng dụng

Bước 5. nếu phần mở rộng trang là aspx hoặc html thì #E Loại bỏ

các liên kết không phải trang như hình ảnh, pdf

chèn truy vấn để thêm dữ liệu nhật ký trong cơ sở dữ liệu

Nhận dạng người dùng

Người dùng là được xác định, là người liên hệ với máy chủ web yêu cầu một số tài nguyên trên web. Các phương pháp khác nhau được đề xuất để nhận dạng người dùng. Điều đơn giản nhất một là gán id người dùng khác nhau cho địa chỉ IP khác nhau. Trong quá trình xác định người dùng, sự cố do bộ nhớ đệm có thể xảy ra. Vấn đề bộ nhớ đệm có thể được khắc phục bằng cách chỉ định một đoạn ngắn hời gian hết hạn đối với các trang HTML buộc trình duyệt phải truy xuất mọi trang từ máy chủ.

Nhận dạng người dùng có nghĩa là xác định cá nhân người dùng bằng cách quan sát địa chỉ IP của họ. Để xác định duy nhất người dùng, chúng tôi đề xuất một số quy tắc: Nếu có địa chỉ IP mới, thì có một người dùng mới, nếu địa chỉ IP giống nhau nhưng hệ điều hành hoặc phần mềm duyệt web khác nhau, một giả định hợp lý là mỗi loại tác nhân khác nhau cho một địa chỉ IP đại diện cho một người dùng khác.

Trong trường hợp trang web được truy cập không có cơ chế xác thực, phương pháp được sử dụng để phân biệt khách truy cập là dựa vào cookie. Phương pháp này cho kết quả chính xác cao, tuy nhiên, do lo ngại về quyền riêng tư, không phải người dùng nào cũng cho phép trình duyệt lưu trữ cookie.

Chỉ sử dụng địa chỉ IP là không đủ để nhận dạng duy nhất người dùng. Nguyên nhân chính là do máy chủ ISP proxy server sẽ gán lại địa chỉ IP cho người dùng sau một khoảng thời gian nhất định. Ngoài ra, có thể có nhiều người dùng trong một mạng LAN sẽ sử dụng cùng một địa chỉ IP public. Do đó, trường hợp hai truy cập khác nhau với cùng một địa chỉ IP nhưng từ hai người dùng khác nhau là hoàn toàn có thể xảy ra.

Để tăng độ chính xác của nhận dạng người dùng dựa trên địa chỉ IP, chúng tôi có thể kết hợp các thông tin khác nhau như tác nhân người dùng user agent hoặc refferer.

Bảng 2.1 mô tả một ví dụ về nhận dạng người dùng sử dụng kết hợp địa chỉ IP và User agent. Bảng 2.2, 2.3, 2.4 cho kết quả sau khi nhận dạng được người dùng riêng biệt.

Bảng 2.1: Kết hợp địa chỉ IP và User Agent

STT	Địa chỉ IP	URL	Ref	Agent
1	1.2.3.4	A	-	Mozilla; Windows NT
2	1.2.3.4	B	A	Mozilla; Windows NT
3	2.3.4.5	C	-	Mozilla; Linux
4	2.3.4.5	B	C	Mozilla; Linux
5	2.3.4.5	E	C	Mozilla; Linux
6	1.2.3.4	C	A	Mozilla; Windows NT
7	2.3.4.5	D	B	Mozilla; Linux
8	1.2.3.4	A	-	Mozilla; Linux
9	1.2.3.4	E	C	Mozilla; Windows NT
10	1.2.3.4	C	A	Mozilla; Linux
11	1.2.3.4	B	C	Mozilla; Linux
12	1.2.3.4	D	B	Mozilla; Linux
13	1.2.3.4	E	D	Mozilla; Linux
14	1.2.3.4	A	-	Mozilla; Windows NT
15	1.2.3.4	C	A	Mozilla; Windows NT

16	1.2.3.4	F	C	Mozilla; Linux
17	1.2.3.4	F	C	Mozilla; Windows NT
18	1.2.3.4	B	A	Mozilla; Windows NT
19	1.2.3.4	D	B	Mozilla; Windows NT
20	1.2.3.4	B	A	Mozilla; Windows NT

Bảng 2.2: Kết quả nhận dạng được người dùng 1

STT theo thời gian	**Địa chỉ IP**	**URL**	**Ref**
1	1.2.3.4	A	-
2	1.2.3.4	B	A
6	1.2.3.4	C	A
9	1.2.3.4	E	C
14	1.2.3.4	A	-
17	1.2.3.4	F	C
18	1.2.3.4	B	A
19	1.2.3.4	D	B
20	1.2.3.4	B	A

Bảng 2.3: Kết quả nhận dạng được người dùng 2

STT theo thời gian	Địa chỉ IP	URL	Ref
3	2.3.4.5	C	-
4	2.3.4.5	B	C
5	2.3.4.5	E	C
7	2.3.4.5	D	B

Bảng 2.4: Kết quả nhận dạng được người dùng 3

STT theo thời gian	Địa chỉ IP	URL	Ref
8	1.2.3.4	A	-
10	1.2.3.4	C	A
11	1.2.3.4	B	C
12	1.2.3.4	D	B
13	1.2.3.4	E	D
16	1.2.3.4	F	C

Nhận dạng phiên

Sau khi xác định từng người dùng, phiên của từng người dùng được làm. Phương pháp đơn giản nhất để xác định phiên sử dụng cơ chế thời gian chờ. Ý nghĩa của thời gian chờ là nếu thời gian giữa các yêu cầu trang vượt quá giới hạn nhất định, cho biết người dùng đang bắt đầu một phiên mới.

Dữ liệu được xử lý lọc bỏ bớt các thông tin như trạng thái không thành công, các file ảnh, file robot. Người dùng được định danh, phân tích hành vi. Trong file log, phần xác thực của user sẽ được ghi lại, từ đó xác định được người dùng. Tuy nhiên, đối với các site khác, không yêu cầu về đăng nhập, thì hoàn toàn không có thông tin này. Trong trường hợp này, người ta thường dựa vào trường cookies. Nhưng không phải trong mọi trường hợp đều có thể sử dụng, bởi một số người dùng disable cookie trên trình duyệt. Khi đó, trường thông tin về IP, user agent và site topology sẽ được dùng đến để xác định một người dùng mới bằng các link.

Nhận dạng phiên làm việc là phân chia các bản ghi theo hoạt động của người dùng thành các phiên, mỗi phiên đại diện cho một lượt truy cập vào website của người dùng đó. Đối với những website không có các cơ chế xác thực người dùng cũng như là các cơ chế bổ sung khác như nhúng thêm định danh phiên (session id) thì cần dùng các phương pháp dựa trên kinh nghiệm - để xác định phiên làm việc ta dùng heuristics methods. Chúng ta coi tập hợp các phiên thực tế của người dùng trên website là R. Một bộ phân loại phiên dựa trên kinh nghiệm - sessionization heuristic h được cố gắng để ánh xạ R thành tập hợp các phiên C_h. Thông thường, các phân loại phiên dựa trên kinh nghiệm gồm hai loại chính: Dựa vào thời gian hoặc dựa vào cấu trúc của trang web.

Phân loại dựa vào thời gian dựa vào việc ước lượng khoảng thời gian giữa các yêu cầu để phân biệt các phiên liên tiếp. Trong khi phân loại dựa trên cấu trúc của website dựa trên cấu trúc của trang web và trường referrer trong web log để phân biệt các phiên.

Với hai loại trên thì một log của máy chủ web có thể được chia thành các phiên dựa trên các phương pháp phân loại cụ thể như sau:

- h1: Tổng thời gian của một phiên thường không vượt quá một ngưỡng θ nhất định. Cho t0 là thời gian của yêu cầu đầu tiên trong phiên S, yêu cầu với thời gian là t sẽ được gán vào phiên S nếu nó thoả mãn: $t - t0 \leq \theta$.

▪ h2: Khoảng thời gian mà người dùng xem một trang web thường không quá một giới hạn δ. Với t1 là thời gian của yêu cầu đã được gán cho phiên S, yêu cầu tiếp theo với thời gian t2 sẽ được gán cho phiên S nếu như nó thoả mãn t2 -t1 $\leq \delta$.

▪ h-ref: Một yêu cầu q sẽ được gán cho phiên S nếu trường referr của q có liên quan đến S. Nếu không thì q được xem như là yêu cầu đầu tiên của một phiên mới. Chú ý rằng, với phương pháp này có thể dẫn tới trường hợp là một yêu cầu q có thể thuộc nhiều phiên khác nhau bởi vì nó có thể cùng lúc liên quan tới nhiều phiên trước đó. Trong trường hợp này, các thông tin khác sẽ được bổ sung để tránh việc mơ hồ khi nhận dạng các phiên. Ví dụ, q có thể được gán cho phiên thoả mãn điều kiện ở trên và được cập nhật mới gần đây nhất.

User 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C

Session 2

1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Hình 2.3: Một ví dụ về nhận dạng phiên dựa trên thời gian

User 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:26	1.2.3.4	F	C

Session 2

1:15	1.2.3.4	A	-
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Hình 2.4 : Một ví dụ về nhận dạng phiên dựa trên thời gian

Hình 2.3 mô tả một ví dụ về nhận dạng phiên dựa trên kinh nghiệm theo phương pháp h1, với θ = 30 phút. Nếu ta áp dụng phương pháp h2 với θ = 10 phút thì kết quả có thể chia thành 3 phiên như sau: A - B - C - E; A và F - B - D.

Hình 2.4 mô tả một ví dụ về nhận dạng phiên sử dụng phương pháp h-ref có cùng tập dữ liệu đầu vào với ví dụ ở Hình 2.3. Trong ví dụ này, với yêu cầu F có thời gian là 1:26 thì sẽ phân làm hai phiên là A-B-C-E và A. Yêu cầu F được thêm vào sesion đầu với vì trường Ref của nó là C có liên quan đến phiên 1. Yêu cầu B với thời gian là 1:30 có thể thuộc cả hai phiên, bởi vì trường Ref của nó là A đều liên quan đến cả hai phiên. Trường hợp này B được thêm vào phiên 2, bởi vì đó là phiên mới nhất được cập nhật mới.

Nhận dạng PageView

Việc xác định các trang mà người dùng xem - pageview phụ thuộc rất nhiều vào cấu trúc cũng như nội dung của trang web. Mỗi lần xem trang có thể được xem như một tập hợp các đối tượng hoặc sự kiện web. Ví dụ: nhấp vào liên kết, xem trang sản phẩm, thêm sản phẩm vào giỏ hàng. Với các trang web tĩnh, mỗi tệp HTML tương ứng với một lần xem trang. Tuy nhiên, với các trang web động, một lần xem trang có thể kết hợp nội dung tĩnh và động do máy chủ tạo ra dựa trên một tập hợp các tham số đầu vào. Ngoài ra, chúng ta có thể xem số lần xem trang dưới dạng tập hợp các trang và đối tượng liên quan đến cùng một lĩnh vực. Ví dụ, với các website thương mại điện tử, lượt xem trang có thể tương ứng với các sự kiện phát sinh khác nhau như xem sản phẩm, đăng ký tài khoản, đổi giỏ hàng, thanh toán ... Các thuộc tính cơ bản cần phải có của một pageview bao gồm: pageview id (thường là một URL), loại pageview (ví dụ như: trang chủ, trang sản phẩm, trang thanh toán.) và các metadata khác (ví dụ như các từ khoá hay các thuộc tính của sản phẩm).

Hoàn thành đường dẫn

Có khả năng bị thiếu các trang sau khi xây dựng giao dịch do máy chủ proxy và sự cố bộ nhớ đệm. Trong điều kiện như vậy, nó trở nên cần thiết xác định đường dẫn truy cập của người dùng và thêm phần còn thiếu những con đường.

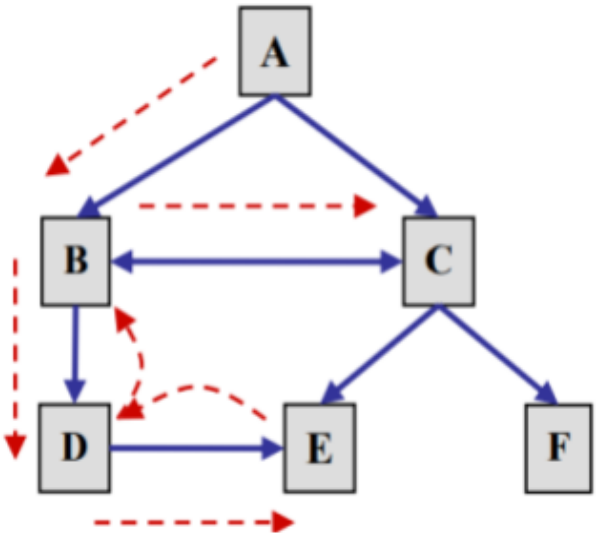
Trong trường hợp sử dụng nút “BACK”, thì các thông tin có thể không được ghi log. Do đó để tìm kiếm các thông tin bị thiếu, thì bước Path complete là cần thiết. Path complete được thực hiện bằng cách phân tích URLs và trường Refferer trong một phiên của người dùng. Nếu một trang nào đó được request không trực tiếp từ trang cuối cùng, thì lịch sử phiên sẽ được tìm kiếm và nếu trang đó là có trong trường referrer URL thì nó sẽ được thêm vào để hoàn thiện log của truy cập.

Ví dụ, trong cùng một phiên làm việc, người dùng truy cập một trang web A 2 lần thì sau lần đầu truy cập, trang web A được proxy server lưu lại trong cache của nó. Trong lần truy cập thứ hai, yêu cầu được gửi đi, máy chủ proxy sẽ trả lại cho máy khách trang web A mà nó đã lưu trước đó và không gửi yêu cầu truy cập đến máy chủ web, dẫn đến yêu cầu truy cập trang web A lần thứ hai không được ghi trên nhật ký máy chủ. Hình 2.5 mô tả một ví dụ về việc tham chiếu- referrer bị thiếu. Quá trình truy cập các trang web của người dùng: A->B->D- >E->D->B->C. Sau khi truy cập trang E, người dùng quay trở về trang D rồi trở về trang B, sau đó chuyển sang trang C. Bảng 2.5 mô tả trường URL và Referrer trong file log của server trong trường hợp này:

Bảng 2.5: Ví dụ trường hợp referrer sai

URL	Referrer
A	-
B	A
D	B
E	D
C	B

Việc quay trở về trang D và trang B sẽ không có trong server log do chúng đã được lưu lại trong cache của proxy server giữa client và máy chủ web. Trong file log, ta sẽ thấy sau khi truy cập trang E, người dùng sẽ truy cập vào trang C với tham chiếu là trang B.



Hình 2.5: Ví dụ về tham chiếu sai do cache

Với vấn đề này, chúng ta có thể sử dụng phương pháp dựa trên thực nghiệm kết hợp với cấu trúc của trang web để phát hiện những tham chiếu bị thiếu hoặc sai để đưa ra giải pháp phù hợp.

Xây dựng cấu trúc của Transactions

Mục tiêu của xác định phiên là tạo ra các trường tham chiếu có ý nghĩa cho mỗi một user. Để xác định lịch sử duyệt web và biết mối quan tâm của một người dùng, thì lưu ý tới giao dịch travel path và giao dịch nội dung. Phiên travel path là một sự kết hợp giữa các page được truy cập thường xuyên và nội dung trang web đó.

Quá trình tiền xử lý và chuẩn hóa làm các công việc như: làm sạch và hợp nhất dữ liệu từ nhiều nguồn khác nhau; nhận dạng người dùng; nhận dạng phiên; xác định số lần xem trang ... kết hợp dữ liệu dòng nhấp chuột với nội dung trang web hoặc dữ liệu cá nhân người dùng. Quá trình này cung cấp dữ liệu tối ưu và nhất quán để phân tích nhật ký web log.

Trên các trang web lớn, nội dung nhật ký log được lưu trữ trong nhiều nguồn khác nhau. Hợp nhất dữ liệu cho phép tổng hợp dữ liệu từ các tệp nhật ký có định dạng khác nhau. Trong trường hợp các nguồn dữ liệu này không có cơ chế chia sẻ mã nhận dạng phiên để hợp nhất dữ liệu, có thể sử dụng phương pháp tiếp cận theo kinh nghiệm học như dựa trên trường "liên kết giới thiệu" trong nhật ký máy chủ, kết quả là nó tương thích với các phương pháp nhận dạng người dùng và nhận dạng phiên để dữ liệu có thể được hợp nhất. Làm sạch dữ liệu loại bỏ các tham chiếu không liên quan hoặc không quan trọng cho mục đích phân tích nhật ký như: tệp CSS trang web, tệp biểu tượng, âm thanh trang web. Quá trình này cũng xóa các trường tệp nhật ký không cung cấp nhiều thông tin quan trọng cho việc phân tích nhật ký như: Phiên bản giao thức HTTP. Ngoài ra, việc dọn dẹp dữ liệu sẽ xóa các tham chiếu là kết quả của lỗi trình thu thập thông tin hoặc công cụ tìm kiếm. Danh sách trình thu thập thông tin của các công cụ tìm kiếm phổ biến có thể được duy trì để các kết quả nhật ký của chúng có thể được phát hiện và loại bỏ.Một phương pháp khác để phát hiện các crawler là dựa vào giao thức hoạt động của chúng, đó là bắt đầu phiên làm việc trên một website, nó đầu tiên sẽ truy cập vào file “robot.txt” của trang web. Dựa vào đặc điểm này, ta cũng có thể xoá bỏ các phiên làm việc của crawler trên website.

2.1.3. Các kỹ thuật phân tích web log

Quá trình đánh giá các bản ghi log sẽ khiến nhân viên của công ty gặp khó khăn trong việc xử lý và quản lý bằng tay các dữ liệu trong các trường hợp như nơi đó diễn ra một số lượng lớn các bản ghi sự kiện . Không chỉ các cuộc tấn công mạng, mà còn các lỗi khác trong hệ thống có thể được giải quyết bằng cách phân tích nhật ký. Các tệp nhật ký hiện có cung cấp một lượng khả năng hiển thị nhất định về hệ thống. Khi diễn giải các tệp nhật ký, điều quan trọng là nhận ra các sự kiện mới theo các cách tiếp cận khác nhau đối với các hành động phát triển trên hệ thống và ứng dụng. Tích hợp thông tin có thể thu được thông qua việc phân tích và so sánh đầy đủ các bản ghi. Nếu các sự kiện không được coi là tổng thể và không

được phân tích đầy đủ, tác động và tầm quan trọng của những sự kiện này có thể không được tiết lộ.

Các kỹ thuật nhận dạng mẫu

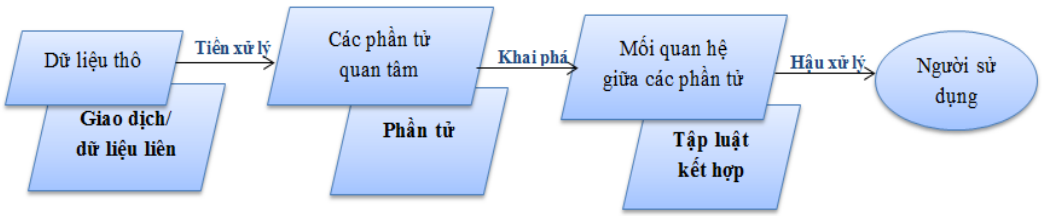
Phân tích thống kê

Thống kê là kỹ thuật phổ biến nhất trong phân tích nhật ký log. Bằng việc phân tích tệp nhận dạng người dùng, phiên làm việc, chúng ta có thể thực hiện các phương pháp thống kê khác nhau như tính tần suất,trung bình... với các biến số khác nhau như: số trang đã xem, số lượt xem, thời gian xem trên mỗi trang. Một số công cụ phân tích hiện nay cũng báo cáo định kỳ về thống kê trang web như: trang web được truy cập nhiều nhất, thời gian xem trang web trung bình, số lượt truy cập trung bình mỗi trang web.

Loại phân tích thống kê này có rất nhiều thông tin hữu ích để cải thiện hiệu suất hệ thống hoặc để tiếp thị.

Luật kết hợp

Các quy tắc kết hợp được sử dụng để khám phá các quy tắc kết hợp giữa các phần tử dữ liệu trong cơ sở dữ liệu. Mẫu đầu ra của thuật toán khai thác dữ liệu là tập hợp các quy tắc kết hợp được tìm thấy.



Hình 2.6: Quá trình sử dụng luật kết hợp

Hình 2.6 đưa ra ví dụ về việc chúng ta sử dụng luật kết hợp, để minh họa về luật kết hợp, ta có thể sử dụng một ví dụ đơn giản như sau: Phân tích cơ sở dữ liệu bán hàng để lấy thông tin về những khách hàng mua card màn hình cũng có xu hướng mua một quạt tản nhiệt trong cùng một lần mua được mô tả trong quy tắc liên kết sau:

“Mua card màn hình Mua quạt tản nhiệt”

[Độ tin cậy: 70%, Độ hỗ trợ: 4%,]

Hai thước đo của sự quan tâm của luật là độ tin cậy và độ hỗ trợ. Chúng lần lượt phản ánh tính hữu ích và tính chắc chắn của định luật được khám phá.

Với độ hỗ trợ là 4% có nghĩa là 4% của tất cả các tác vụ được phân tích chỉ ra rằng card màn hình và quạt tản nhiệt đã được mua cùng nhau. Và với độ tin cậy 70% có nghĩa là 70% khách hàng mua card màn hình cũng mua quạt tản nhiệt.

Phân lớp- Classification

Bài toán phân lớp là quá trình phân lớp một đối tượng dữ liệu thành một hoặc nhiều lớp cho trước bằng cách sử dụng một mô hình phân lớp(model). Mô hình này được xây dựng dựa trên một tập dữ liệu đã xây dựng trước đó với các nhãn (hay còn gọi là tập huấn luyện). Phân lớp là quá trình gán nhãn cho các đối tượng dữ liệu.

Nhiệm vụ của bài toán phân lớp là tìm ra mô hình phân lớp để khi có dữ liệu mới có thể xác định được dữ liệu đó thuộc lớp nào. Có nhiều vấn đề về phân lớp dữ liệu như phân lớp nhị phân, phân loại đa lớp và phân lớp đa trị.

Trong phân tích nhật ký truy cập, phân lớp thường được sử dụng để ánh xạ người dùng đến một lớp hoặc một kiểu cụ thể. Việc phân lớp trong phân tích nhật ký web có thể được thực hiện bằng các thuật toán học máy có giám sát như: cây quyết định, thuật toán Naive Bayes, thuật toán K láng giềng gần nhất ... Ví dụ, phân lớp nhật ký máy chủ có thể giúp phân loại 46% người dùng đặt mua sản phẩm trên trang 'laptop dell’có độ tuổi từ 18-23 và chủ yếu sống ở miền Bắc.

Phân cụm – Clustering

Phân cụm là một kỹ thuật khá quan trọng trong khai phá dữ liệu, nó thuộc về lớp phương pháp Unsupervised Learning(Học không đào tạo) trong Học máy. Về định nghĩa về kỹ thuật này có khá nhiều định nghĩa khác nhau song về bản chất chúng ta có thể hiểu phân cụm là quá trình tìm kiếm các nhóm đối tượng đã cho thành từng cụm - cluster, sao cho các đối tượng trong cùng một cụm là tương tự nhau , và các đối tượng trong các cụm khác nhau là không tương tự.

Thực chất việc phân cụm là để tìm ra bản chất các nhóm dữ liệu bên trong. Các thuật toán phân cụm tạo ra các cụm. Tuy nhiên, không có tiêu chí tốt nhất duy

nhất để đánh giá hiệu quả của phân tích cụm, nó phụ thuộc vào mục đích của phân nhóm như: data redution, , “useful” clustes, “natural clustes”,outlier detection.

Trong phân tích nhật ký log , có thể thực hiện hai kiểu phân cụm: usage cluster và page cluster.

Phân cụm nhóm người dùng có các mẫu pattern tương tự có nhiều thông tin có giá trị cho tiếp thị và thương mại điện tử. Ví dụ, với một số nhóm người nhất định, chỉ có thể đưa ra các đề xuất mua hàng phù hợp với sở thích của nhóm người dùng đó.

Mặt khác, phân cụm trang giúp xác định các nhóm trang có nội dung liên quan. Thông tin này đặc biệt hữu ích cho các công cụ tìm kiếm, có thể tạo các trang được đề xuất phù hợp với truy vấn của người dùng.

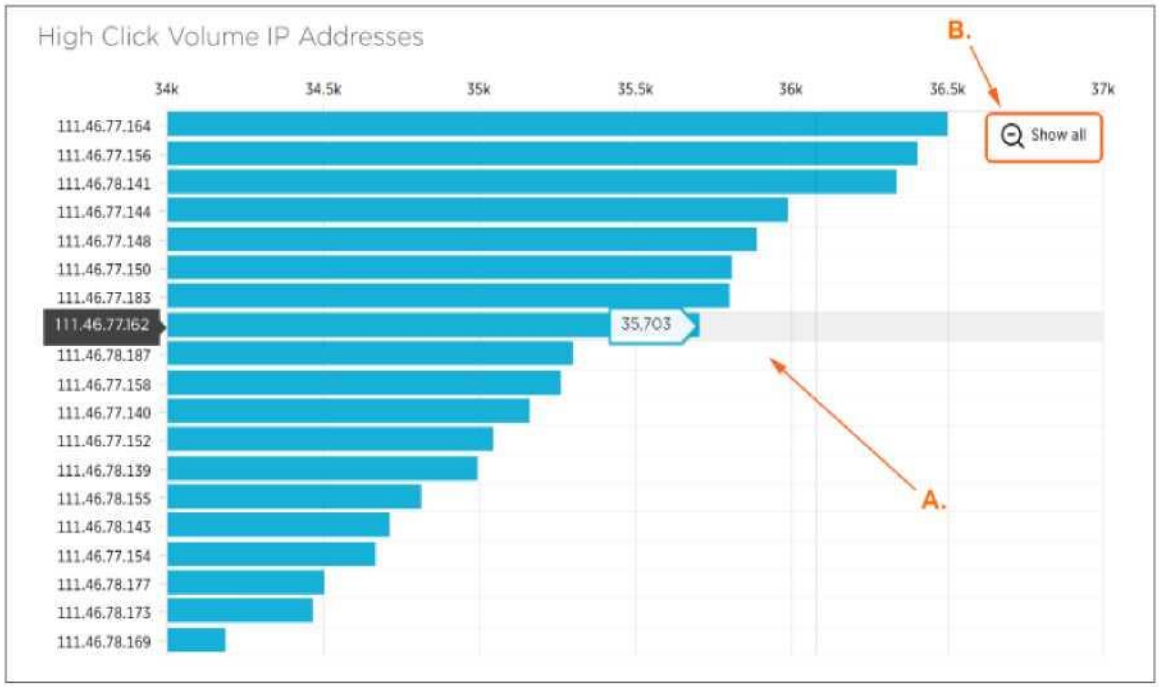
Phân tích mẫu

Đây là bước cuối cùng của quá trình phân tích nhật ký log truy cập. Quá trình này để lọc ra các luật hoặc mẫu pattern không có nhiều giá trị đã được tạo trong bước khám phá mẫu (Pattern Discovery).

Có nhiều phương pháp để thực hiện việc phân tích mẫu, song phương pháp phổ biến và được sử dụng nhiều nhất là thông qua truy vấn SQL hoặc cũng có thể sử dụng phân tích xử lý trực tuyến - OLAP.

Ngoài ra, ở bước này, chúng ta còn áp dụng các kỹ thuật trực quan hóa dữ liệu như sơ đồ, biểu đồ thống kê để phục vụ cho việc phân tích các mẫu pattern.

Hình 2.7 cho thấy một ví dụ sử dụng trực quan hóa dữ liệu. Chúng ta thấy rằng việc biểu diễn dữ liệu bằng biểu đồ, đồ thị thống kê giúp chúng ta dễ dàng nhận ra sự tương quan của dữ liệu cũng như nhận biết được xu hướng phát triển của dữ liệu.

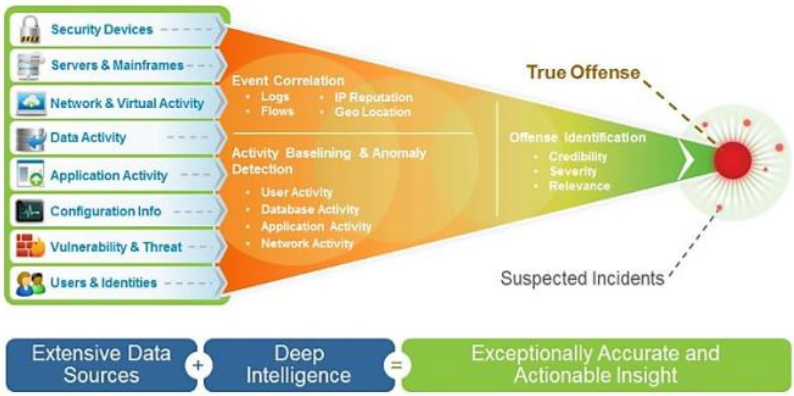


Hình 2.7: Ví dụ sử dụng data visualization

2.2. Các công cụ và nền tảng phân tích web log

2.2.1. IBM QRadar SIEM

IBM QRadar SIEM (Security Information and Event Management) là một hệ thống được thiết kế để cung cấp cho các nhóm bảo mật khả năng hiển thị tập trung vào các doanh nghiệp để bảo vệ dữ liệu. Nó quản lý thông tin và các sự cố bảo mật an ninh do IBM, Hoa Kỳ phát triển và cung cấp. QRadar SIEM (IBM QRadar, 2017) cho phép phát hiện các bất thường và mối đe dọa với độ chính xác cao và tỷ lệ cảnh báo sai thấp thông qua xử lý và phân tích dữ liệu nhật ký log và luồng mạng từ hàng nghìn thiết bị và ứng dụng phân tán trong mạng, như minh họa trong Hình 2.8.[8]



Hình 2.8 : Mô tả thu thập dữ liệu và xử lý của Qradar SIEM

QRadar SIEM có sẵn cho các môi trường tại chỗ và đám mây.Các tính năng tiêu biểu của QRadar SIEM bao gồm (IBM QRadar, 2017):

▪ Cảm nhận, phát hiện giả mạo, các nguy cơ bên trong: Khám phá hoạt động đáng ngờ của người dùng có thể cho thấy thông tin đăng nhập bị xâm phạm hoặc mối đe dọa từ nội bộ.

▪ Cảm nhận, phát hiện giả mạo, các mối đe dọa nâng cao: Nhận được khả năng phát hiện mối đe dọa chính xác, theo thời gian thực để kết hợp một số sự kiện có vẻ rủi ro thấp để tìm ra cuộc tấn công mạng có nguy cơ cao đang diễn ra.

▪ Thực hiện việc chuẩn hóa và tương quan các sự kiện tức thời;

▪ Cảm nhận, theo dõi và liên kết các sự cố và nguy cơ;

▪ Hỗ trợ cài đặt tại hệ thống mạng của khách hàng, hoặc có thể truy cập như một dịch vụ SIEM trên nền điện toán đám mây;

▪ Giám sát bảo mật OT và Iot: Giám sát tập trung cho các giải pháp OT và IoT để xác định hoạt động bất thường và các mối đe dọa tiềm ẩn.

▪ Có thể bổ sung dung lượng lưu trữ và năng lực xử lý nhanh chóng và rẻ tiền;

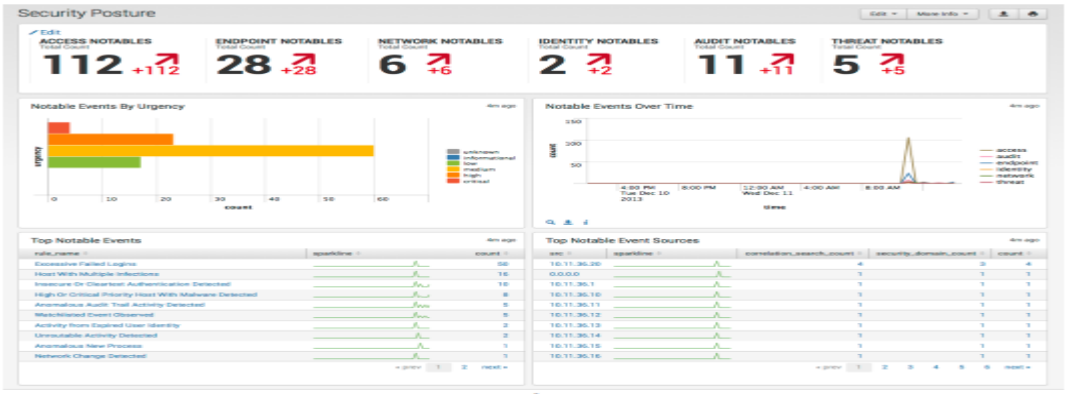
▪ Tích hợp tri thức phát hiện nguy cơ từ IBM X-Force: Danh mục đầu tư, được hỗ trợ bởi nghiên cứu và phát triển X-Force nổi tiếng thế giới, cung cấp thông tin bảo mật để giúp các tổ chức bảo vệ toàn diện con người, cơ sở hạ tầng, dữ liệu và ứng dụng của họ, cung cấp giải pháp quản lý truy cập và nhận dạng, bảo mật cơ sở dữ liệu, quản lý rủi ro, quản lý rủi ro , an ninh mạng và nhiều hơn nữa.

▪ Hỗ trợ quản lý và phối hợp phòng chống rủi ro ;

▪ Khả năng tích hợp với các sản phẩm khác của IBM, hoặc các công ty khác.

IBM QRadar SIEM đã được triển khai, sử dụng và được đánh giá cao tại các cơ quan, tổ chức chính phủ, ngân hàng, doanh nghiệp có hệ thống mạng lớn. Tuy nhiên, hạn chế lớn nhất của QRadar SIEM là chi phí cài đặt ban đầu và phí bản quyền khá lớn nên chưa thực sự phù hợp với các cơ quan, tổ chức có mạng lưới quy mô vừa và nhỏ và nguồn lực bị giới hạn.

2.2.2. Splunk



Hình 2.9: Giao diện tổng hợp của Splunk

Splunk (Splunk, 2017) là một phần mềm giám sát an ninh mạng dựa trên phân tích nhật ký log. Đây là công cụ phân tích và xử lý nhật ký log rất mạnh mẽ, được cung cấp bởi Splunk Inc., Hoa Kỳ.[7] Splunk với hàng trăm công cụ tích hợp sẵn, cho phép xử lý nhiều loại nhật ký log khác nhau với khối lượng lớn theo thời gian thực. Splunk có thể xử lý và phân tích nhật ký log để đảm bảo an toàn thông tin, cũng như trích xuất thông tin để hỗ trợ các hoạt động kinh doanh. Splunk cung cấp các công cụ tìm kiếm và vẽ đồ thị cho phép biểu diễn đầu ra ở nhiều định dạng khác nhau. Hình 2.9 cho thấy màn hình tổng hợp an ninh (Sercurity Posture Dashboard) của Splunk.

Splunk có ba phiên bản, bao gồm:

- Splunk Enterprise : Phiên bản dành cho các khách hàng có nhu cầu xử lý nhật ký log tại chỗ với khối lượng lớn;
- Splunk Cloud : Phiên bản dành cho các khách hàng tải nhật ký log lên nền tảng đám mây của Splunk để xử lý;
- Splunk Light: Phiên bản dành cho các khách hàng có nhu cầu xử lý nhật ký log tại chỗ với khối lượng vừa và nhỏ.

Các tính năng tiêu biểu của Splunk bao gồm:

- Cập nhật dữ liệu: Splunk cập nhật dữ liệu liên tục khi có thay đổi trong thời gian thực. Giúp cho việc phát hiện và cảnh báo chính xác trong thời gian thực.

▪ Lập chỉ mục dữ liệu: Splunk được xây dựng dựa trên Lucence, có thể lập chỉ mục dữ liệu với một lượng dữ liệu rất lớn trong một khoảng thời gian ngắn. Giúp tìm kiếm nhanh chóng và thuận tiện.

▪ Tìm kiếm thông tin: Splunk hoạt động rất tốt với dữ liệu lớn và được cập nhật liên tục. Nó cung cấp cho công cụ tìm kiếm một "Splunk Language" cực kỳ thông minh bao gồm từ khóa, hàm và cấu trúc tìm kiếm giúp người dùng có thể truy xuất mọi thứ, theo nhiều tiêu chí từ tập dữ liệu rất lớn. Để nói lên sức mạnh của Splunk, Splunk còn được các nhà quản trị mạng cao cấp và chuyên nghiệp gọi với cái tên "Splunk toàn năng" hay "Splunk as Google for Log files".

▪ Giám sát và cảnh báo an ninh mạng: Splunk cung cấp cho người dùng một cơ chế cảnh báo dựa trên việc tìm kiếm thông tin do người dùng thiết lập. Khi có sự cố liên quan đến hệ thống phù hợp với tiêu chí người dùng đặt ra, hệ thống sẽ cảnh báo ngay cho người dùng (cảnh báo trực tiếp qua giao diện, gửi đến Email).

▪ Xử lý sự cố: Splunk cũng cung cấp cơ chế tự động khắc phục các sự cố xảy ra bằng cách cấu hình tự động chạy các tập tin Script do người dùng tạo (Ví dụ: Chặn IP, đóng port…) khi có báo động.

▪ Hiển thị thông tin: Splunk cung cấp cơ chế hiển thị trực quan giúp người dùng dễ dàng hình dung tình trạng của hệ thống, đưa ra các đánh giá về hệ thống. Splunk cũng tự động xuất ra các báo cáo chuyên nghiệp với nhiều định dạng khác nhau.

Phát triển: Đồng thời cung cấp các API hỗ trợ người dùng tạo ứng dụng trên Splunk. Một số bộ API tiêu biểu như Splunk SDK (Cung cấp SDK trên nền tảng Python, Java, JS, PHP), Shep (Splunk Hadoop Integration - Đây là sự kết hợp giữa Splunk và Hadoop), Shuttl (Là sản phẩm hỗ trợ backup dữ liệu trong Splunk), Splunkgit (Giúp bạn hình dung dữ liệu của mình tốt hơn), Splunk power shell resource Kit (Bộ công cụ hỗ trợ mở rộng và quản lý hệ thống).

Hạn chế lớn nhất của Splunk là chi phí lắp đặt cao, do mức đầu tư ban đầu cho hệ thống thiết bị chuyên dụng rất phức tạp. Một vấn đề nữa là phí bản quyền hàng năm của Splunk cũng rất đắt (ước tính hàng chục nghìn USD / năm) nên

Splunk không thực sự phù hợp với các cơ quan, tổ chức có hệ thống mạng quy mô vừa và nhỏ với tài nguyên hạn chế.

2.2.3. ELK Stack

ELK Stack là tập hợp 3 phần mềm đi chung với nhau, phục vụ cho công việc logging.[10] Ba phần mềm này lần lượt là:

Elasticsearch: Cơ sở dữ liệu để lưu trữ, tìm kiếm và query log.

Logstash: Tiếp nhận log từ nhiều nguồn, sau đó xử lý log và ghi dữ liệu và Elasticsearch.

Kibana: Giao diện để quản lý, thống kê log. Đọc thông tin từ Elasticsearch.

Hình 2.10: Cơ chế hoạt động của ELK Stack

Hình 2.10 mô tả cơ chế hoạt động của ELK Stack. Đầu tiên, log sẽ được đưa đến Logstash. (Thông qua nhiều con đường, ví dụ như server gửi UDP request chứa log tới URL của Logstash, hoặc Beat đọc file log và gửi lên Logstash). Logstash sẽ đọc những log này, thêm những thông tin như thời gian, IP, parse dữ liệu từ log (server nào, độ nghiêm trọng, nội dung log) ra, sau đó ghi xuống database là Elasticsearch. Khi muốn xem log, người dùng vào URL của Kibana. Kibana sẽ đọc thông tin log trong Elasticsearch, hiển thị lên giao diện cho người dùng query và xử lý.

Elasticsearch

ElasticSearch là một công cụ tìm kiếm cấp doanh nghiệp. Mục tiêu của nó là tạo ra một công cụ, nền tảng hoặc kỹ thuật phân tích và tìm kiếm trong thời gian thực và cách nó có thể dễ dàng áp dụng hoặc triển khai cho nhiều nguồn dữ liệu.

Nguồn dữ liệu trên bao gồm các cơ sở dữ liệu nổi tiếng như MS SQL, PostgreSQL, MySQL, ...

Logstash

Logstash là một công cụ thu thập dữ liệu mã nguồn mở với khả năng tổng hợp theo thời gian thực. Logstash có thể tự động thu thập dữ liệu từ nhiều nguồn khác nhau và chuẩn hóa dữ liệu đó tùy thuộc vào đích dữ liệu.

Ban đầu logstash chỉ hoạt động như một công cụ thu thập nhật ký, nhưng các khả năng của logstash hiện đã vượt qua vai trò đó. Bất kỳ loại sự kiện nào cũng có thể được logstash thu thập thông qua các plugin đầu vào và đầu ra, cùng với mã được đơn giản hóa giúp tăng khả năng nhập, xử lý và khai thác hiệu quả nhiều loại dữ liệu khác nhau.

Kibana

Kibana là một nền tảng phân tích và trực quan mã nguồn mở được thiết kế để hoạt động với Elasticsearch. Chúng tôi sử dụng Kibana để tìm kiếm, xem và tương tác với dữ liệu được lưu trữ trong Elasticsearch. Từ đó, dễ dàng thực hiện phân tích dữ liệu và trực quan hóa dữ liệu của bạn thông qua biểu đồ và bảng.

Beats

Beats là một tập hợp các công cụ thu thập thông tin chuyên biệt, được gọi là Shippers, thu thập và gửi dữ liệu từ máy khách đến máy chủ ELK. Ngoài ra, những nhịp này có thể được gửi trực tiếp đến Elasticsearch vì bản thân các công cụ đã được tiêu chuẩn hóa, việc kết nối nhịp với logstash thường có ý nghĩa bảo mật đối với các hệ thống quy mô lớn khi họ muốn. bảo vệ dịch vụ tìm kiếm đàn hồi.

ELK Stack là một công cụ tiện dùng và được nhiều công ty sử dụng. Và lý do là:

- Đọc log từ nhiều nguồn: Logstash có thể đọc được log từ rất nhiều nguồn, từ log file cho đến log database cho đến UDP hay REST request.

- Dễ tích hợp.

- Hoàn toàn miễn phí.

- Khả năng tìm kiếm mạnh mẽ: nhờ có Elasticsearch mà việc tìm kiếm dữ liệu trở lên nhanh chóng hơn, so với nhiều công cụ khác thì Elasticsearch có thể nói là rất nhanh và mạnh mẽ dựa trên Apache Lucene. Nó tìm kiếm gần với thời gian thực – Near-Real Time Searching, điều này cho thấy tốc độ tìm kiếm của nó rất nhanh.

- Khả năng phân tích dữ liệu.

2.2.4. Graylog

Graylog là một nền tảng mã nguồn mở được tích hợp đầy đủ để thu thập, lập chỉ mục và phân tích dữ liệu có cấu trúc và phi cấu trúc từ hầu như bất kỳ nguồn nào.[9] Nó đã được phát triển từ năm 2010.

Các thành phần của ứng dụng Graylog:

- Máy chủ Graylog
- Giao diện web Graylog
- Mango DB (Thống kê và đồ thị)
- ElasticSearch (Tin nhắn và tìm kiếm).

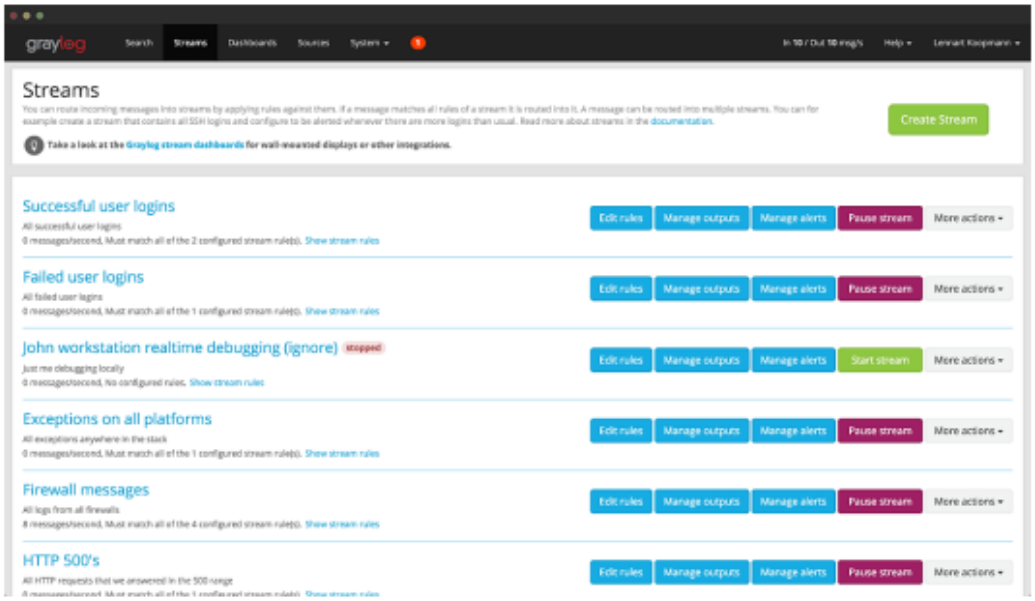
Việc thu thập dữ liệu nhật ký log được thực hiện rất linh hoạt nhờ sự hỗ trợ của các công cụ thu thập nhật ký của bên thứ ba, chẳng hạn như beats, fluentd và nxlog. Hình 2.11 minh họa màn hình quản lý các nguồn nhật ký của Graylog. Graylog có khả năng phân tích hành vi của người dùng, ứng dụng cho phép phát hiện và cảnh báo những lượt truy cập bất thường cũng như trích xuất các mẫu hành vi truy cập để tối ưu hóa trang web. Graylog cũng cho phép ánh xạ từ ID đến tên người dùng và ánh xạ từ địa chỉ IP đến vị trí địa lý.

Một số tính năng của Graylog, bao gồm:

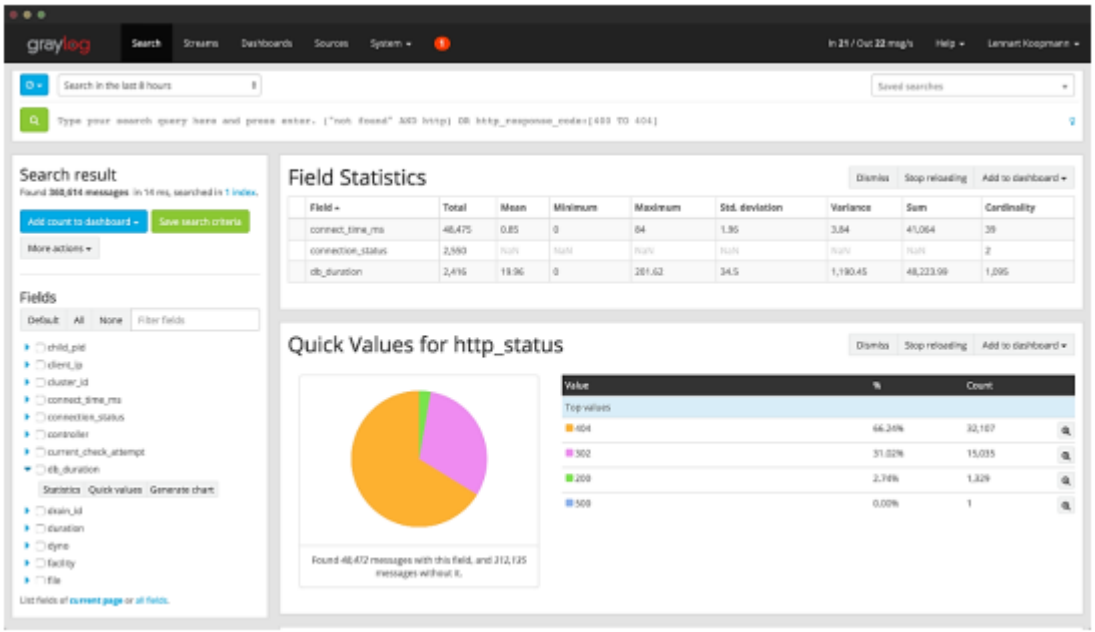
- Đây là một giải pháp quản lý nhật ký mã nguồn mở.
- Nó hỗ trợ nhiều đầu vào như Syslog, GELF, TCP, UDP, AMQP cho quá trình ghi nhật ký.
- Lưu trữ tin nhắn trên ElasticSearch cho phép tìm kiếm nhanh trên các kho lưu trữ.
- Sử dụng MongoDB cho các hoạt động thống kê.

- Có thể phân loại nhật ký và thực hiện các thao tác đồ họa.
- Có thể dễ dàng tạo ra các cảnh báo chủ động theo các điều kiện mong muốn.

Hình 2.12 biểu diễn màn hình báo cáo tổng hợp của Graylog. Mặc dù Graylog có khả năng xác định các hành vi truy cập bất thường nhưng nó không cho phép phân tích sâu về các mối đe dọa an toàn thông tin, chẳng hạn như sự xuất hiện của mã độc và các kiểu tấn công vào hệ thống, tài nguyên và dịch vụ mạng.



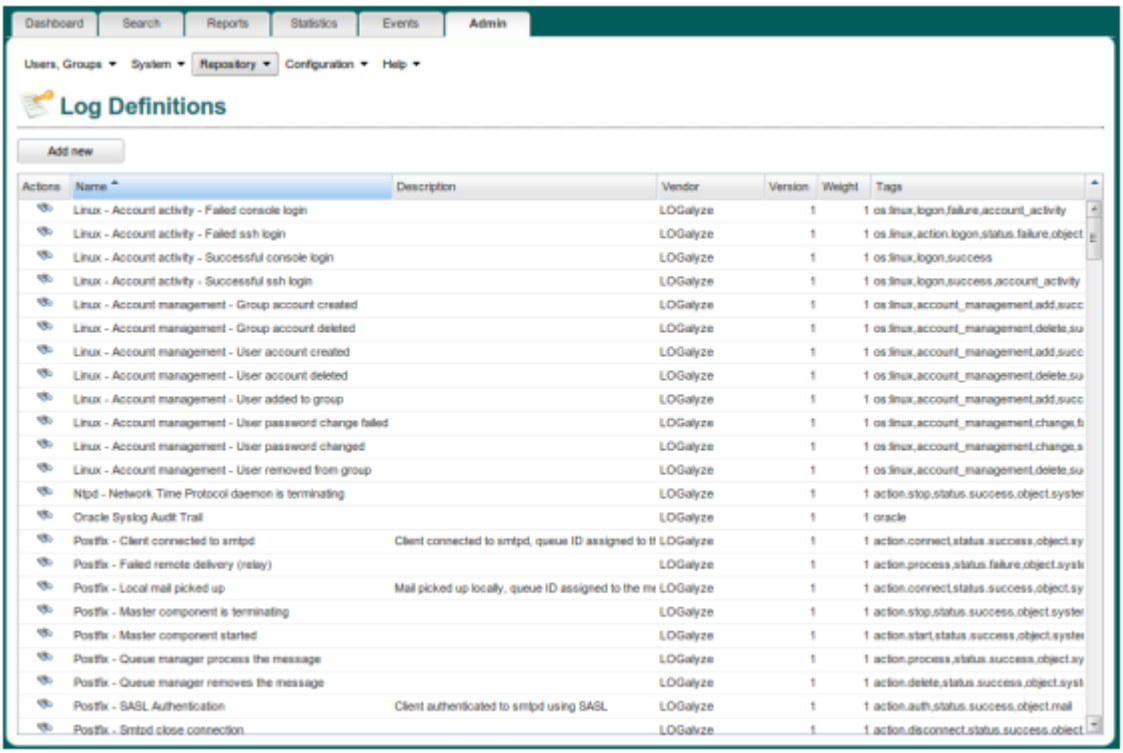
Hình 2.11: Màn hình quản lý các nguồn thu thập log của GrayLog



Hình 2.12: Màn hình báo cáo tổng hợp của Graylog

2.2.5. LOGalyze

LOGalyze (LOGalyze, 2017) là một phần mềm mã nguồn mở cho phép giám sát mạng tập trung và quản lý nhật ký log[11]. LOGalyze hỗ trợ xử lý nhật ký log từ nhiều nền tảng, bao gồm nhật ký từ máy chủ Unix / Linux, Windows và các thiết bị mạng với xử lý, tìm kiếm và phát hiện bất thường trong thời gian thực. LOGalyze cũng cho phép người dùng xác định các sự kiện và cảnh báo dựa trên dữ liệu nhật ký log được thu thập và xử lý. Hình 2.13 minh họa màn hình quản lý các dạng nhật ký log của LOGalyze. Ngoài ra, LOGalyze còn là công cụ quản lý và giám sát mạng, giúp phát hiện các cấp độ truy cập bất thường và các sự cố mạng. Tuy nhiên, tương tự như Graylog, LOGalyze khó có khả năng phân tích sâu về các nguy cơ mất an toàn thông tin, chẳng hạn như dấu hiệu của mã độc và các cuộc tấn công vào các dịch vụ và tài nguyên mạng.



Hình 2.13: Màn hình quản lý các dạng log của LOGalyze

2.2.6. So sánh các công cụ và nền tảng phân tích web log

Phân tích web giúp chúng ta hiểu những gì đang xảy ra trong các quy trình công nghệ thông tin và hướng dẫn giám sát, đánh giá và giải pháp của các vấn đề. Các dịch vụ ghi nhật ký log cung cấp sự bảo vệ đối với dữ liệu nhật ký chứa thông tin có giá trị như hệ thống, mạng và các ứng dụng . Quản lý nhật ký là một quá trình phức tạp và các tổ chức thường mắc sai lầm trong khi đánh giá nó. Công cụ quản lý log giúp chúng ta dễ dàng lấy thông tin từ dữ liệu nhật ký có kích thước lớn. Những công cụ này kết hợp tất cả dữ liệu và cho phép chúng ta quản lý nó bằng một giao diện trung tâm, dễ tiếp cận và dễ sử dụng. Do đó, chúng ta có thể thu thập, lưu trữ và quản lý dữ liệu trong một cửa hàng. Với nhật ký các công cụ quản lý, xu hướng hữu ích có thể được trích xuất từ dữ liệu nhật ký hiện tại. Công cụ phù hợp với bạn sẽ phụ thuộc vào số lượng hệ thống đang được giám sát và nhu cầu tuân thủ của tổ chức bạn. Khi tổ chức gặp khó khăn tình huống như tấn công mạng, sẽ dễ dàng hơn khi sử dụng công cụ quản lý nhật ký thay vì xử lý các tệp TXT hiện có trong môi trường.

Với một truy vấn duy nhất, nguyên nhân gốc rễ của sự cố của bất kỳ ứng dụng hoặc phần mềm nào có thể được xác định với sự trợ giúp của các công cụ này. Nhật ký ban quản lý có thể giảm thiểu thiệt hại mà nó sẽ gây ra trước khi xảy ra một cuộc tấn công mạng.

Bảng 2.6 cho thấy những ưu điểm và nhược điểm của các nền tảng, công cụ xử lý, phân tích log truy cập ở trong và ngoài nước.[1]

Bảng 2.6: So sánh các công cụ và nền tảng phân tích web log

Nền tảng	Ưu điểm	Nhược điểm
IBM QRadar SIEM	- IBM Qradar thu thập và xử lýđược nhiều loại log khác nhau với khối lượng lớn và dữ liệu từ luồng mạng khác nhau - Hỗ trợ thu thập dữ liệu từ hàng ngàn thiết bị mạng - IBM Qradar SIEM phát hiện các bất	- Chi phí cài đặt ban đầu và phí bản quyền khá lớn - Đòi hỏi thiết bị chuyên dụng - Khó khăn trong vận hành và bảo trì.

Nền tảng	Ưu điểm	Nhược điểm
	thườngvới tỷ lệ cảnh báo sai thấp ,độ chính xác cao .	
Splunk	- Splunk hỗ trợ xử lý nhiều dạng nhật ký log khác nhau với khối lượng lớn theo thời gian thực - Hỗ trợ phân tích nhật ký để đảm bảo an toàn thông tin. - Hỗ trợ trích xuất thông tin hỗ trợ hoạt động kinh doanh.	- Chi phí bản quyền, cài đặt và vận hành cao - Đòi hỏi thiết bị chuyên dụng - Khó khăn trong vận hành và bảo trì
ELK Stack	- Mã mở, miễn phí - ELK Stack thu thập log từ rất nhiều nguồn khác nhau: log snmp, log ứng dụng, log hệ thống, log thiết bị mạng, log từ các hệ thống API (Application Programming Interface).	- Không phù hợp cho những trường hợp mà dữ liệu được ghi nhiều (create, update, delete). - Không hỗ trợ transaction, không có ràng buộc quan hệ giữa các dữ liệu dẫn tới việc dữ liệu có thể bị sai.
Graylog	- Nguồn mở và miễn phí - Các luồng cho phép xác định các sự kiện trong thời gian thực và thực hiện các hành động. - Cài đặt dễ dàng - Chức năng phía máy chủ có thể được mở rộng thông qua các trình cắm thêm - Nhật ký có thể được bổ sung và phân tích cú pháp bằng cách sử dụng thuật toán quy trình toàn diện. - Bảng điều khiển đặc biệt để xuất nhật ký trực quan dữ liệu và truy vấn.	- Hỗ trợ số lượng ít các loại nhật ký

Nền tảng	Ưu điểm	Nhược điểm
	- Giao diện tìm kiếm trực quan	
LOGalyze	- Mã mở, miễn phí - Cho phép quản lý log và giám sát mạng tập trung - Hỗ trợ xử lý log từ nhiều nền tảng - Hỗ trợ phát hiện bất thường, sự cố theo thời gian thực	- Không được cập nhật và hỗ trợ từ 2013

2.3. Kết luận chương

Chương 2 giới thiệu chi tiết các kỹ thuật xử lý, phân tích log, bao gồm mô hình xử lý web log, vấn đề thu thập và tiền xử lý web log và các kỹ thuật phân tích web log. Đồng thời, chương cũng khảo sát và so sánh các ưu và nhược điểm của một số nền tảng và công cụ phân tích log phổ biến hiện nay, bao gồm IBM Qradar SIEM, Splunk, ELK Stack, GrayLog và Logalyze.

CHƯƠNG 3. THỬ NGHIỆM TRIỂN KHAI GIẢI PHÁP PHÂN TÍCH WEB LOG SỬ DỤNG ELK STACK

3.1. Mô hình thử nghiệm xử lý và phân tích web log

3.1.1. Giới thiệu mô hình hệ thống

Hình 3.1 mô tả mô hình một hệ thống xử lý và phân tích log dựa trên ELK Stack và đây cũng là mô hình triển khai hệ thống xử lý và phân tích web log thử nghiệm thực hiện trong luận văn.

Hình 3.1. Mô hình hệ thống xử lý và phân tích log dựa trên ELK

Theo mô hình trên, hệ thống xử lý và phân tích log dựa trên ELK Stack gồm các thành phần chính sau:

- Beats là các mô đun thu thập dữ liệu log tại các hệ thống cần giám sát và vận chuyển dữ liệu log về mô đun Logstash. ELK Stack hỗ trợ nhiều dạng beat cho thu thập nhiều dạng dữ liệu khác nhau, như filebeat cho thu thập các dạng log của hệ điều hành và các ứng dụng, dịch vụ, metricbeat cho thu thập các dữ liệu về hoạt động của hệ thống như tình hình sử dụng CPU, bộ nhớ RAM, packetbeat cho thu thập dữ liệu lưu lượng mạng... ELK Stack cũng hỗ trợ thu thập và xử lý dữ liệu từ các công cụ và thiết bịu bảo mật như tường lửa, các hệ thống IDS/IPS...

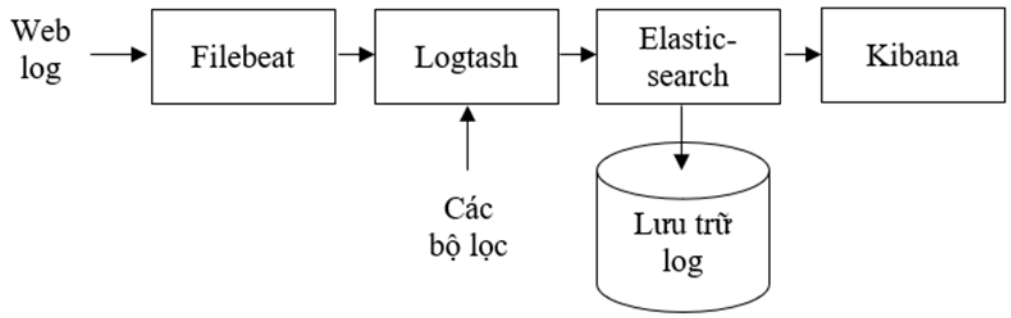
- Logstash là mô đun cho phép tập trung và lọc, chuẩn hóa dữ liệu thu thập từ nguồn thông qua các beat. Logstash hỗ trợ các dạng bộ lọc như grok, chop phép lọc và chuẩn hóa các dạng dữ liệu sử dụng các biểu thứ chính quy.

- Elasticsearch là mô đun cho phép lưu trữ, lập chỉ số và tìm kiếm các dạng dữ liệu log. Elasticsearch hỗ trợ tìm kiếm full-text và lọc dữ liệu sử dụng các bộ lọc tìm kiếm.

- Kibana là mô đun cho phép phân tích, hiển thị dữ liệu log theo nhiều định dạng khác nhau, như hiển thị dưới dạng text, các dạng biểu đồ, đồ thị. Ngoài ra Kibana cũng cấp giao diện web thân thiện, dễ sử dụng cho người dùng.

3.1.2. Quy trình thu thập, xử lý và phân tích web log

Hình 3.2 mô tả mô hình hệ thống phân tích log thử nghiệm thực hiện trong luận văn. Dữ liệu web log gồm các dạng web log mẫu trong EKL Stack, IIS log, Apache log được thu thập bởi filebeat và vận chuyển đến Logstash. Logtask tiếp nhận, lọc và chuẩn hóa log sử dụng các bộ lọc grok (Hình 3.3). Dữ liệu log sau chuẩn hóa được đưa sang Elasticsearch để lưu trữ, lập chỉ số phục vụ tìm kiếm, phân tích. Cuối cũng, dữ liệu log được biểu diễn trên giao diện của Kibana theo các định dạng khác nhau.



Hình 3.2. Mô hình triển khai hệ thống phân tích log thử nghiệm

```
grok {
  match ⇒ { "message" ⇒ ["%{SYSLOGTIMESTAMP:[system][auth][timestamp]
            "%{SYSLOGTIMESTAMP:[system][auth][timestamp]} %{SYSLOGHOST:
            "%{SYSLOGTIMESTAMP:[system][auth][timestamp]} %{SYSLOGHOST:
            "%{SYSLOGTIMESTAMP:[system][auth][timestamp]} %{SYSLOGHOST:
            "%{SYSLOGTIMESTAMP:[system][auth][timestamp]} %{SYSLOGHOST:
            "%{SYSLOGTIMESTAMP:[system][auth][timestamp]} %{SYSLOGHOST:
            "%{SYSLOGTIMESTAMP:[system][auth][timestamp]} %{SYSLOGHOST:
  pattern_definitions ⇒ {
    "GREEDYMULTILINE"⇒ "(.|\n)*"
  }
  remove_field ⇒ "message"
}
date {
  match ⇒ [ "[system][auth][timestamp]", "MMM  d HH:mm:ss", "MMM dd HH
}
geoip {
  source ⇒ "[system][auth][ssh][ip]"
  target ⇒ "[system][auth][ssh][geoip]"
}
```

Hình 3.3. Một phần bộ lọc grok tích hợp trong Logstash

3.1.3. Cài đặt ELK Stack và các công cụ kèm theo

Yêu cầu phần cứng và phần mềm

Hệ thống thử nghiệm được triển khai trên máy ảo chạy hệ điều hành Ubuntu Linux với các yêu cầu phần cứng và phần mềm sau:

- Hệ thống chạy CPU Intel Core i5, 4GB RAM, 100GB HDD
- Ubuntu phiên bản 16.04
- JDK 1.8 trở lên
- Bộ ELK Stack, bao gồm filebeat, logstash, elasticsearch và kibana cùng các tiện ích kèm theo.

Cài đặt

Hệ thống được cài đặt theo các bước sau:

Bước 1: Cài đặt các thành phần nền tảng (nếu chưa có)

- Cài đặt JDK 1.8: sudo apt-get install openjdk-8-jre-headless
- Cài đặt curl (là một công cụ dòng lệnh cho phép kết nối và tải một URL): sudo apt-get install curl

Bước 2: Cài đặt và cấu hình Elasticsearch

- Cài đặt thành phần Elasticsearch: sudo apt-get install elasticsearch
- Chỉnh sửa cấu hình Elasticsearch (tối thiểu 2 tham số network.host: 192.168.112.150 và http.port: 9200):

sudo pico /etc/elasticsearch/elasticsearch.yml

- Thiết lập cho phép chạy tự động và khởi chạy Elasticsearch:

sudo systemctl enable elasticsearch

- sudo systemctl start elasticsearch
- Khi Elasticsearch được cài đặt, cấu hình và chạy thành công, kiểm tra bằng lệnh “curl https://192.168.112.150:9200 --cacert /etc/elasticsearch/certs/http_ca.crt -u elastic”, kết quả cho như trên hình 3.4.

Hình 3.4. Elasticsearch đã được cài đặt và chạy thành công

Bước 3: Cài đặt và cấu hình Kibana

- Cài đặt thành phần Kibana: sudo apt-get install kibana

- Chỉnh sửa cấu hình Kibana:

sudo pico /etc/kibana/kibana.yml

- Thiết lập cho phép chạy tự động và khởi chạy Kibana:

sudo systemctl enable kibana

sudo systemctl start kibana

Bước 4: Cài đặt và cấu hình Logstash

- Cài đặt thành phần Logstash: sudo apt-get install logstash

- Chỉnh sửa cấu hình Logstash: sudo pico /etc/logstash/logstash.yml

- Bổ sung thêm các file cấu hình input, filter và output cho Logstash.

- Thiết lập cho phép chạy tự động và khởi chạy Logstash:

sudo systemctl enable logstash

sudo systemctl start logstash

Bước 5: Cài đặt và cấu hình Filebeat

- Cài đặt thành phần Filebeat: sudo apt-get install filebeat

- Chỉnh sửa cấu hình Filebeat:

sudo pico /etc/filebeat /filebeat.yml

- Thiết lập cho phép chạy tự động và khởi chạy Filebeat:

sudo systemctl enable filebeat

sudo systemctl start filebeat

3.2. Thử nghiệm và kết quả

3.2.1. Giới thiệu tập dữ liệu web log thử nghiệm

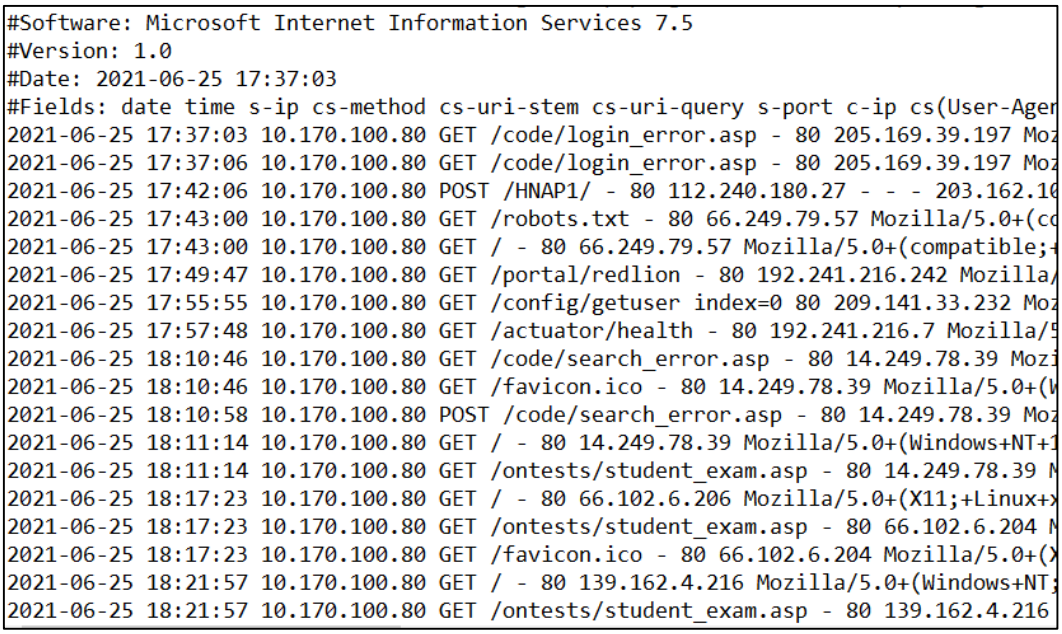
Luận văn sử dụng dữ liệu web log mẫu cung cấp bởi ELK Stack và Microsoft IIS log cho thử nghiệm:

- Web log mẫu gồm hơn 2100 bản ghi thu thập trong tháng 5.2022 (Hình 3.5).

- Microsoft IIS log gồm dữ liệu log vận hành website http://infosecptit.com/ontests/ trong 30 ngày (Hình 3.6).

```
106.77.13.9 - - [2018-07-30T09:54:16.856Z] "GET /beats/metricbeat/metricbeat-6.3.2-amd64.deb HTTP/1.1" 200 1909 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322)"
254.160.84.19 - - [2018-07-30T09:54:46.979Z] "GET /apm HTTP/1.1" 404 9222 "-" "Mozilla/5.0 (X11; Linux i686) AppleWebKit/534.24 (KHTML, like Gecko) Chrome/11.0.696.50 Safari/534.24"
32.208.36.11 - - [2018-07-30T09:56:35.489Z] "GET /styles/ad-blocker.css HTTP/1.1" 200 8017 "-" "Mozilla/5.0 (X11; Linux x86_64; rv:6.0a1) Gecko/20110421 Firefox/6.0a1"
119.73.170.50 - - [2018-07-30T09:59:23.540Z] "GET /elasticsearch/elasticsearch-6.3.2.zip HTTP/1.1" 200 4691 "-" "Mozilla/5.0 (X11; Linux x86_64; rv:6.0a1) Gecko/20110421 Firefox/6.0a1"
215.67.92.140 - - [2018-07-30T10:05:11.690Z] "GET /kibana/kibana-6.3.2-linux-x86_64.tar.gz HTTP/1.1" 200 8458 "-" "Mozilla/5.0 (X11; Linux i686) AppleWebKit/534.24 (KHTML, like Gecko) Chrome/11.0.696.50 Safari/534.24"
52.105.119.80 - - [2018-07-30T10:05:40.315Z] "GET /elasticsearch/elasticsearch-6.3.2.zip HTTP/1.1" 200 12460 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322)"
73.176.17.223 - - [2018-07-30T10:07:30.084Z] "GET /enterprise HTTP/1.1" 200 3183 "-" "Mozilla/5.0 (X11; Linux i686) AppleWebKit/534.24 (KHTML, like Gecko) Chrome/11.0.696.50 Safari/534.24"
155.206.194.40 - - [2018-07-30T10:10:52.414Z] "GET /security-analytics HTTP/1.1" 200 214 "-" "Mozilla/5.0 (X11; Linux i686) AppleWebKit/534.24 (KHTML, like Gecko) Chrome/11.0.696.50 Safari/534.24"
104.32.0.154 - - [2018-07-30T10:13:00.236Z] "GET /kibana/kibana-6.3.2-windows-x86_64.zip HTTP/1.1" 200 6928 "-" "Mozilla/5.0 (X11; Linux x86_64; rv:6.0a1) Gecko/20110421 Firefox/6.0a1"
```

Hình 3.5. Một số bản ghi của web log mẫu cung cấp bởi ELK



```
#Software: Microsoft Internet Information Services 7.5
#Version: 1.0
#Date: 2021-06-25 17:37:03
#Fields: date time s-ip cs-method cs-uri-stem cs-uri-query s-port c-ip cs(User-Ager
2021-06-25 17:37:03 10.170.100.80 GET /code/login_error.asp - 80 205.169.39.197 Moz
2021-06-25 17:37:06 10.170.100.80 GET /code/login_error.asp - 80 205.169.39.197 Moz
2021-06-25 17:42:06 10.170.100.80 POST /HNAP1/ - 80 112.240.180.27 - - - 203.162.10
2021-06-25 17:43:00 10.170.100.80 GET /robots.txt - 80 66.249.79.57 Mozilla/5.0+(co
2021-06-25 17:43:00 10.170.100.80 GET / - 80 66.249.79.57 Mozilla/5.0+(compatible;+
2021-06-25 17:49:47 10.170.100.80 GET /portal/redlion - 80 192.241.216.242 Mozilla/
2021-06-25 17:55:55 10.170.100.80 GET /config/getuser index=0 80 209.141.33.232 Moz
2021-06-25 17:57:48 10.170.100.80 GET /actuator/health - 80 192.241.216.7 Mozilla/5
2021-06-25 18:10:46 10.170.100.80 GET /code/search_error.asp - 80 14.249.78.39 Mozi
2021-06-25 18:10:46 10.170.100.80 GET /favicon.ico - 80 14.249.78.39 Mozilla/5.0+(W
2021-06-25 18:10:58 10.170.100.80 POST /code/search_error.asp - 80 14.249.78.39 Moz
2021-06-25 18:11:14 10.170.100.80 GET / - 80 14.249.78.39 Mozilla/5.0+(Windows+NT+1
2021-06-25 18:11:14 10.170.100.80 GET /ontests/student_exam.asp - 80 14.249.78.39 M
2021-06-25 18:17:23 10.170.100.80 GET / - 80 66.102.6.206 Mozilla/5.0+(X11;+Linux+x
2021-06-25 18:17:23 10.170.100.80 GET /ontests/student_exam.asp - 80 66.102.6.204 M
2021-06-25 18:17:23 10.170.100.80 GET /favicon.ico - 80 66.102.6.204 Mozilla/5.0+(X
2021-06-25 18:21:57 10.170.100.80 GET / - 80 139.162.4.216 Mozilla/5.0+(Windows+NT;
2021-06-25 18:21:57 10.170.100.80 GET /ontests/student_exam.asp - 80 139.162.4.216
```

Hình 3.6. Một số bản ghi của Microsoft IIS log

3.2.2. Một số kết quả

Dữ liệu đầu ra của quá trình thử nghiệm chính là các thống kê dưới đây.

Dưới đây là các giao diện và kết quả thử nghiệm phân tích web log:

- Hình 3.7 mô tả giao diện trang chủ của Kibana;

- Hình 3.8 mô tả trạng thái hoạt động của ELK Stack;

- Hình 3.9 mô tả luồng log thu thập trong 30 ngày gần đây;

- Hình 3.10 mô tả phân bố log thu thập trong 2 ngày gần đây;

- Hình 3.11 mô tả phân bố các loại trình duyệt máy khách truy cập web;

- Hình 3.12 mô tả phân bố các loại trình duyệt kèm nơi máy khách truy cập website;

- Hình 3.13 mô tả phân bố các loại hệ điều hành máy khách truy cập website;

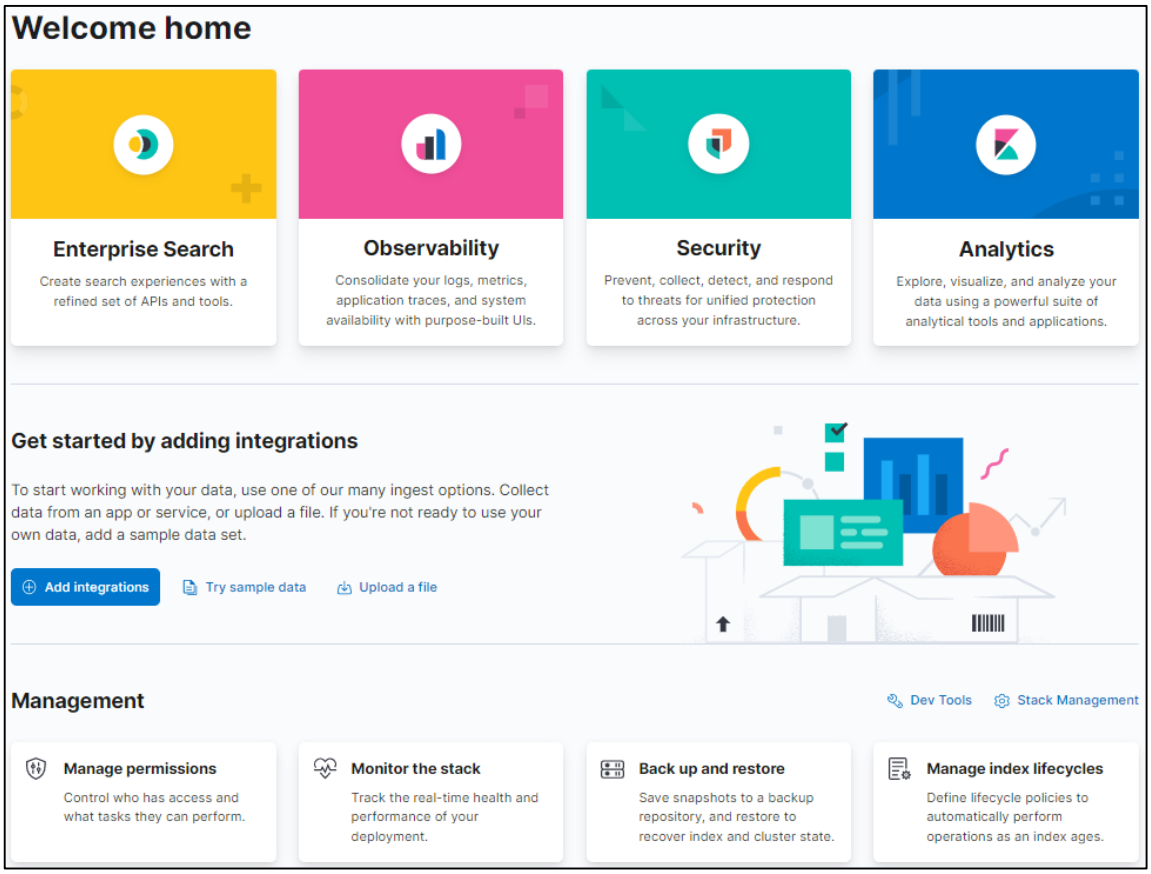
- Hình 3.14 mô tả phân bố truy cập các địa chỉ URL của các website;

- Hình 3.15 mô tả phân bố các cặp đích - nguồn truy cập theo nước;

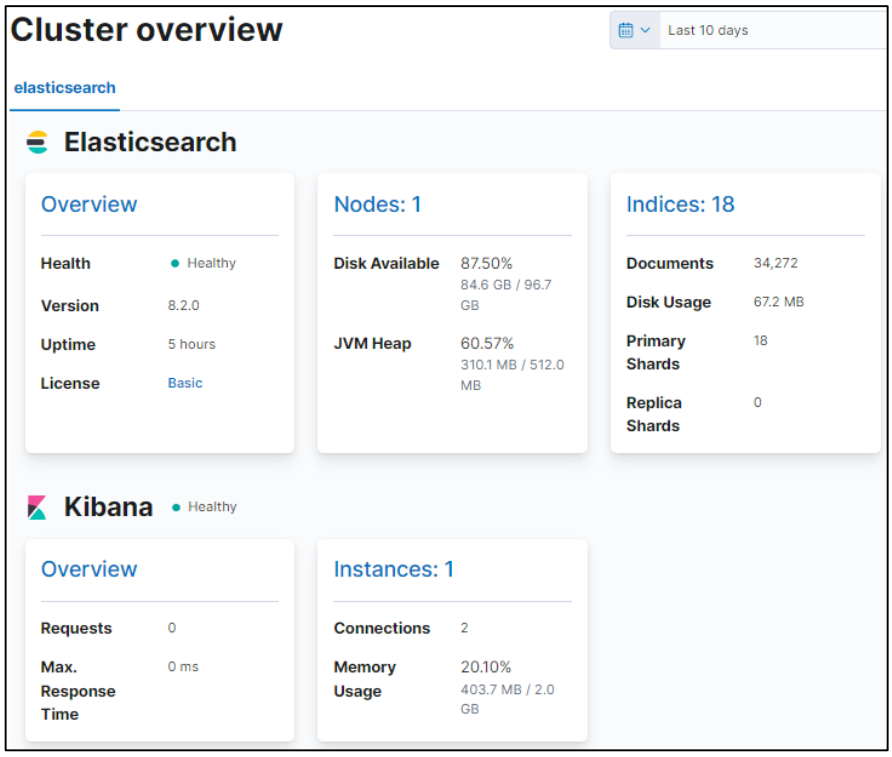
- Hình 3.16 mô tả phân bố nguồn (client) truy cập theo nước;

- Hình 3.17 mô tả một phần màn hình Dashboard phân tích web log;

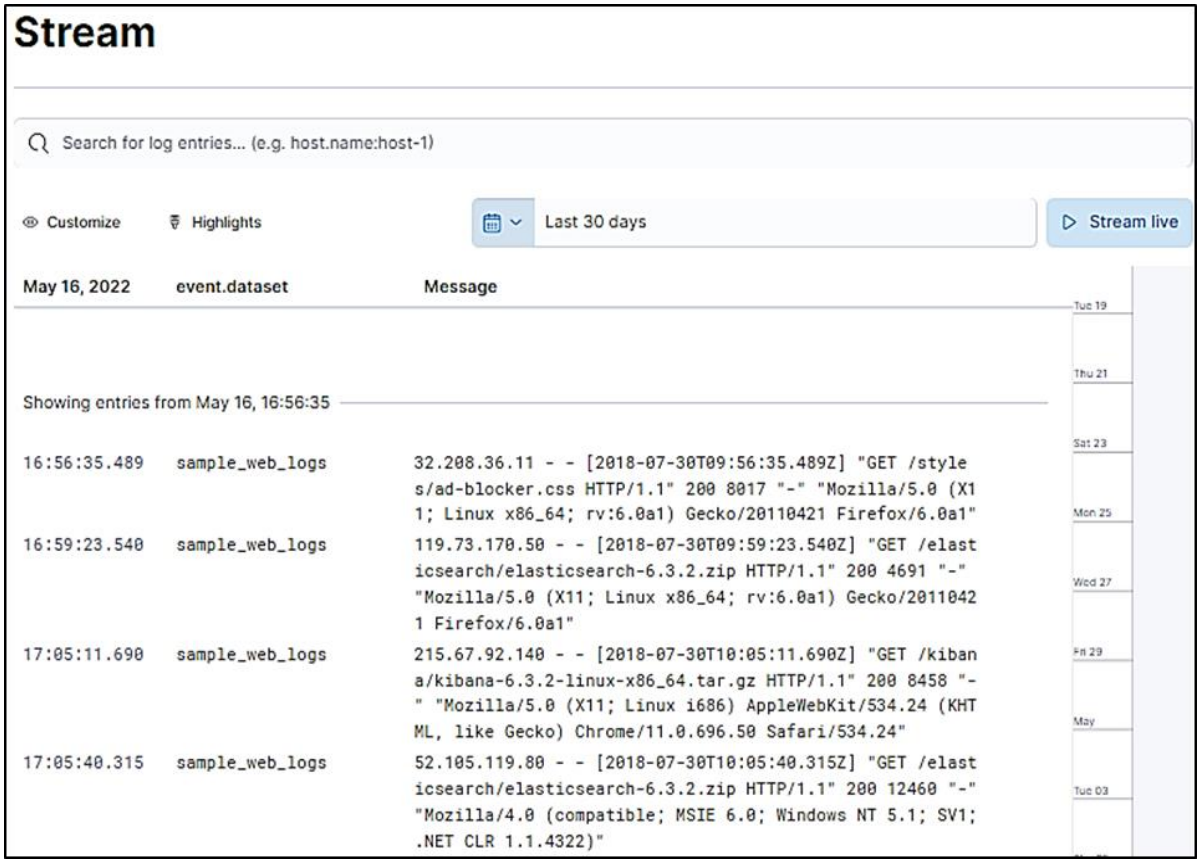
- Hình 3.18 mô tả thống kê lỗi truy cập theo host / URL.



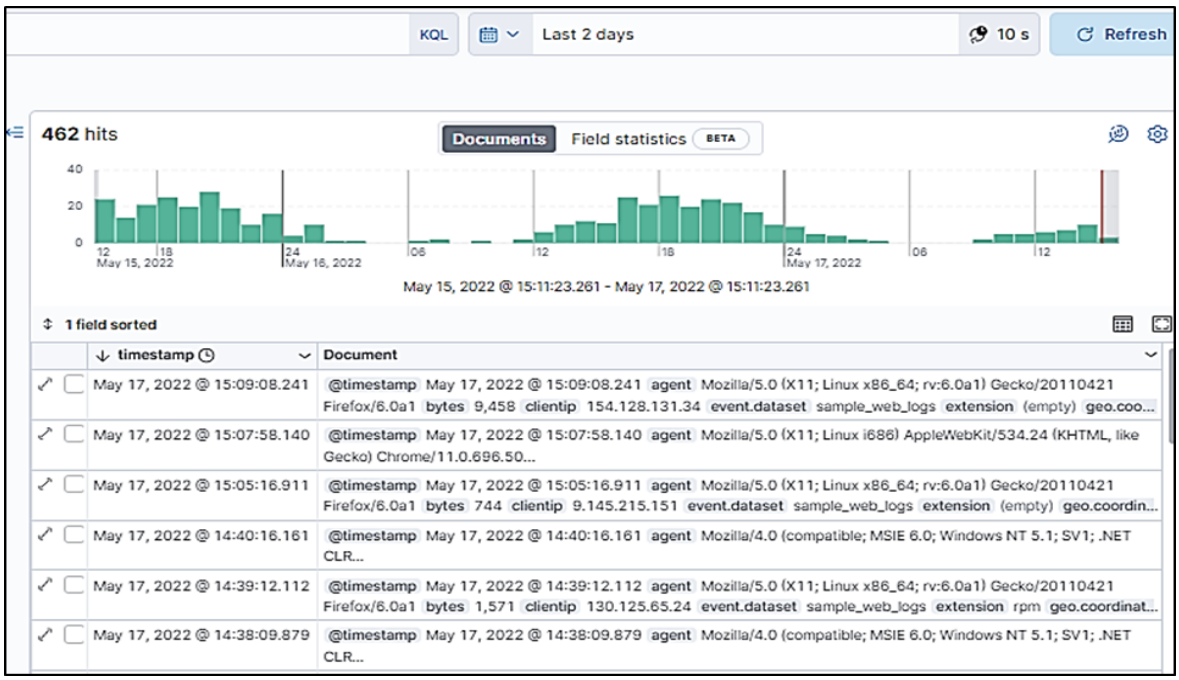
Hình 3.7. Giao diện trang chủ của Kibana



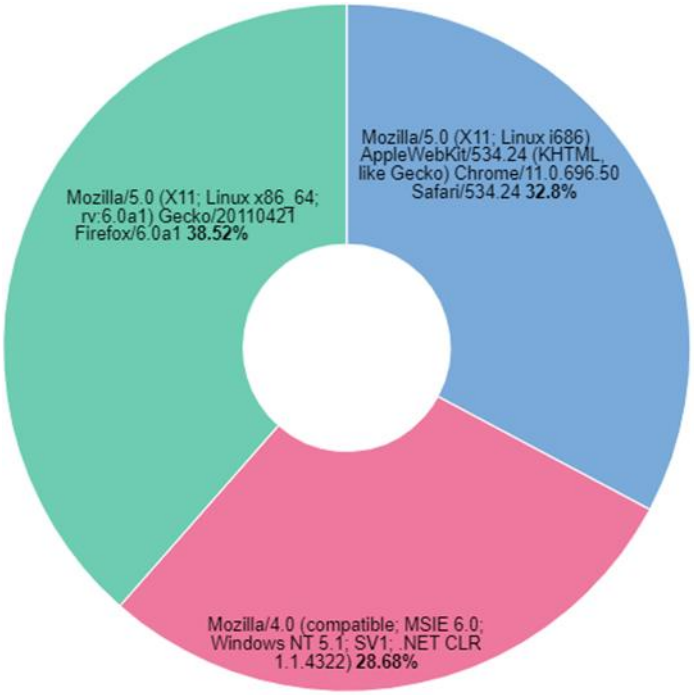
Hình 3.8. Trạng thái hoạt động của ELK Stack



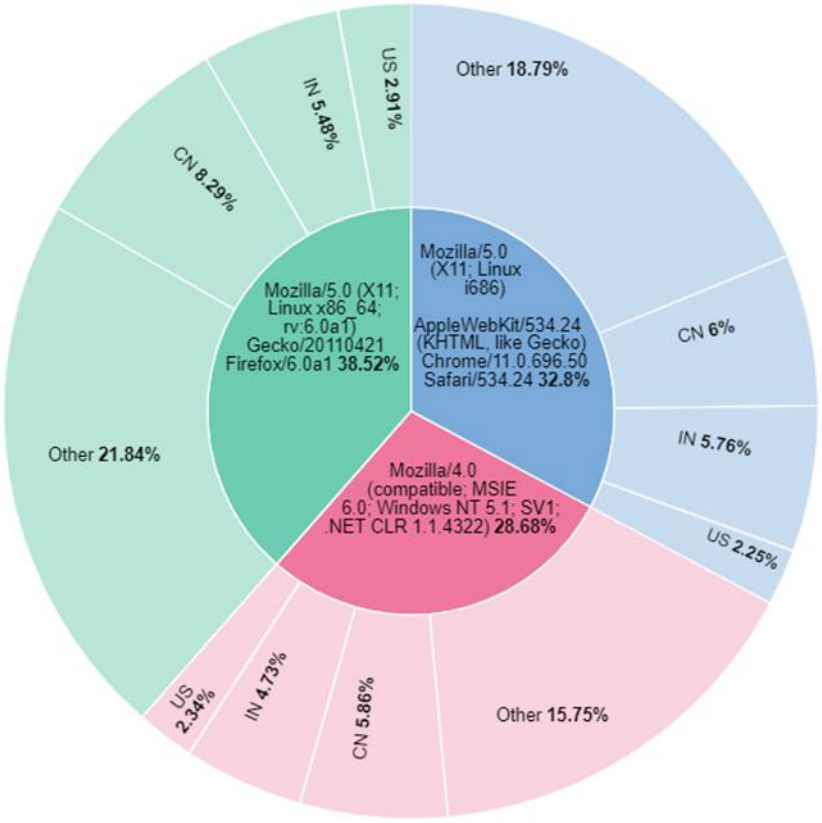
Hình 3.9. Luồng log thu thập trong 30 ngày gần đây



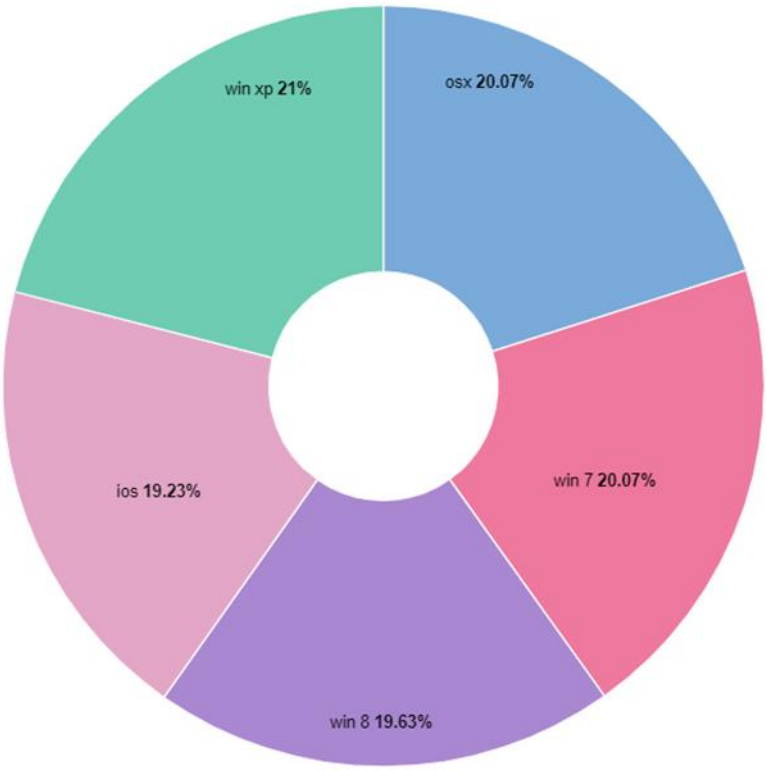
Hình 3.10. Phân bố log thu thập trong 2 ngày gần đây



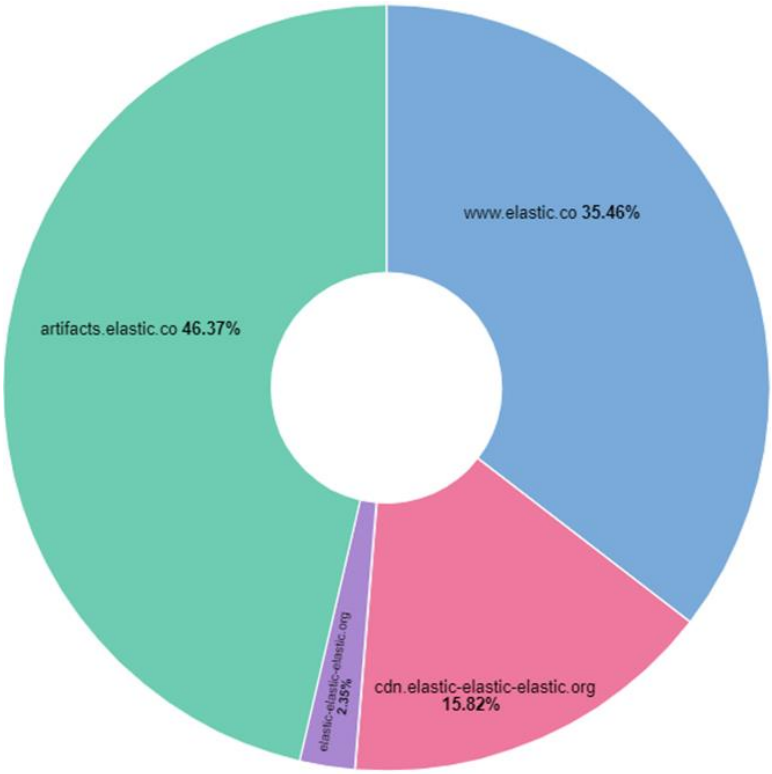
Hình 3.11. Phân bố các loại trình duyệt máy khách truy cập website



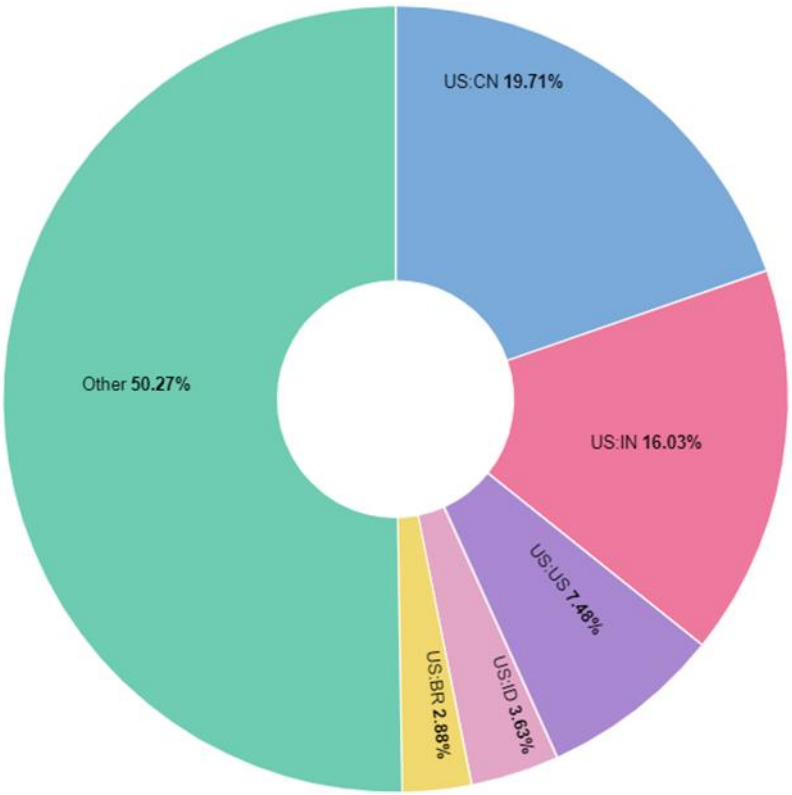
Hình 3.12. Phân bố các loại trình duyệt kèm nơi máy khách truy cập website



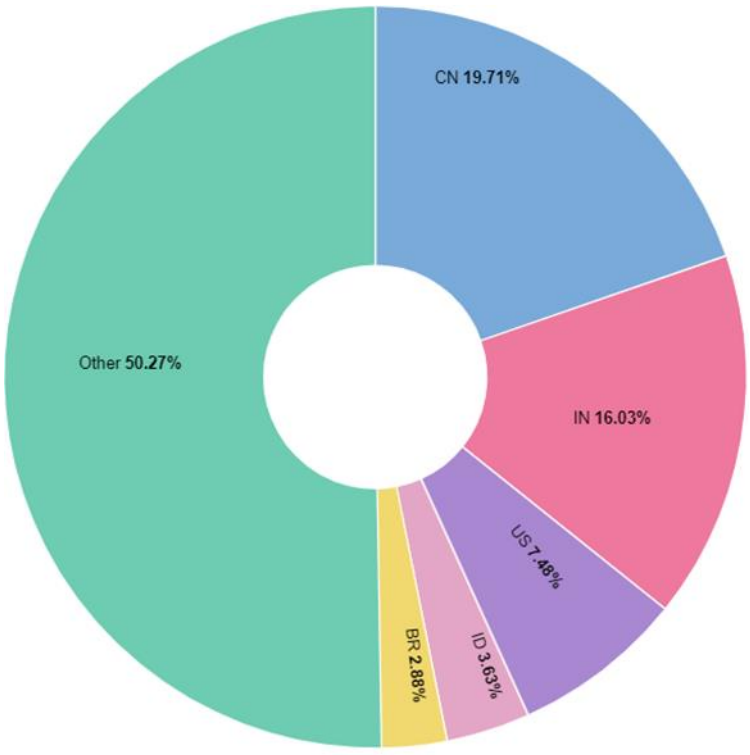
Hình 3.13. Phân bố các loại hệ điều hành máy khách truy cập website



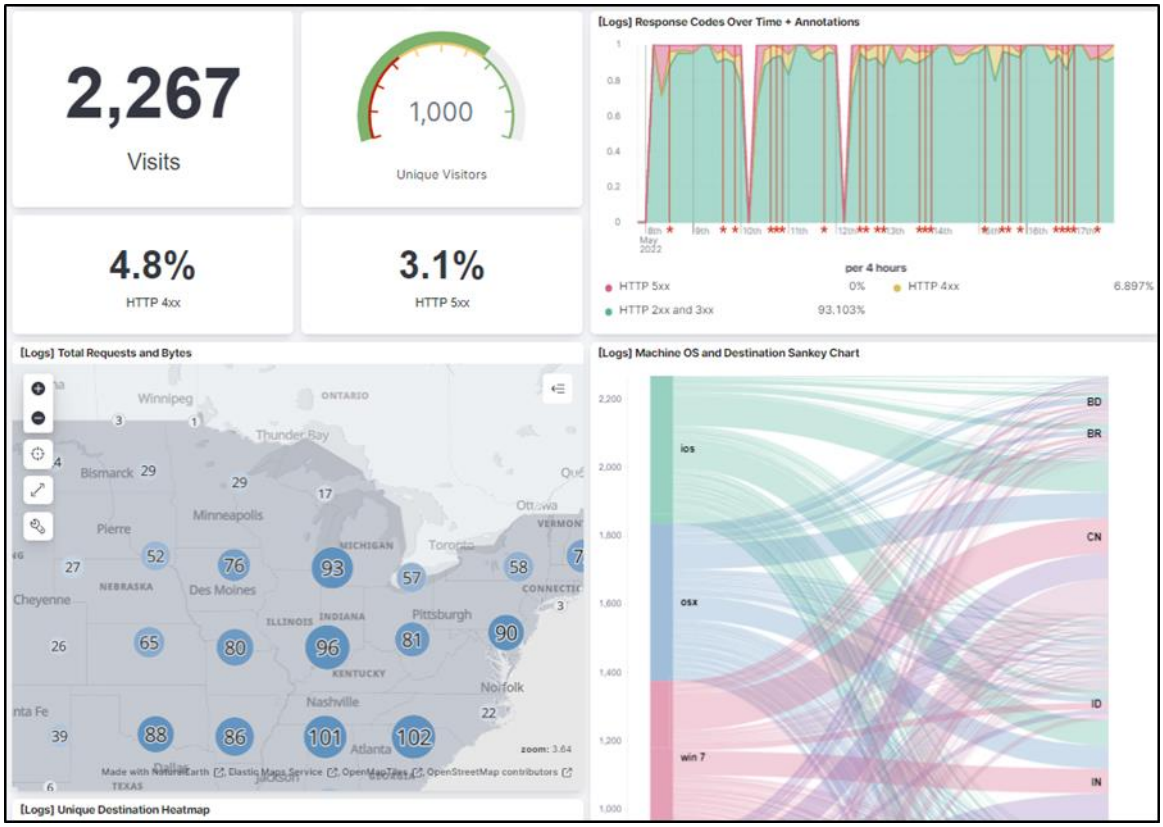
Hình 3.14. Phân bố truy cập các địa chỉ URL của các website



Hình 3.15. Phân bố các cặp đích - nguồn truy cập theo nước



Hình 3.16. Phân bố nguồn (client) truy cập theo nước



Hình 3.17. Một phần màn hình Dashboard phân tích web log

[Logs] Errors by host

URL	Visits	Unique	HTTP 4xx	HTTP	95th perce	Median of
https://elastic-elastic-elastic.org/people/type:astronauts/name:klaus-dietrich-flade/profile	1	1	0.0%	100.0%	0B	0B
https://www.elastic.co/products	11	11	0.0%	9.1%	916B	664B
https://www.elastic.co/solutions/enterprise-search	11	11	0.0%	9.1%	884B	397B
https://artifacts.elastic.co/downloads/beats/metricbeat/metricbeat-6.3.2-i686.rpm	77	73	3.9%	6.5%	15KB	5KB
https://artifacts.elastic.co/downloads/apm-server/apm-server-6.3.2-amd64.deb	63	62	9.5%	6.3%	10KB	6KB
https://artifacts.elastic.co/downloads/kibana/kibana-6.3.2-linux-x86_64.tar.gz	54	53	9.3%	5.6%	9KB	5KB
https://artifacts.elastic.co/downloads/beats/metricbeat/metricbeat-6.3.2-amd64.deb	58	57	5.2%	5.2%	14KB	6KB
https://www.elastic.co/downloads	61	59	4.9%	4.9%	10KB	6KB
https://cdn.elastic-elastic-elastic.org/styles/main.css	42	41	7.1%	4.8%	10KB	6KB
https://artifacts.elastic.co/downloads/apm-server/apm-server-6.3.2-windows-x86.zip	70	68	5.7%	4.3%	10KB	6KB
https://artifacts.elastic.co/downloads/kibana/kibana-6.3.2-darwin-x86_64.tar.gz	52	51	5.8%	3.8%	10KB	5KB
https://www.elastic.co/downloads/enterprise	56	55	8.9%	3.6%	14KB	6KB

Hình 3.18. Thống kê lỗi truy cập theo host / URL

Sau khi thực hiện phân tích log, kết quả đầu ra có thể kể đến như :dựa vào sơ đồ hình 3.11 ta có thể biết phân bố các loại trình duyệt máy khách đã truy cập website. Ta có thể biết được trình duyệt máy khách nào truy cập nhiều, trình duyệt máy khách nào truy cập ít, từ đó có hướng nâng cấp website. Hình 3.13 mô tả phân bố các loại hệ điều hành máy khách truy cập website. Hình 3.14 mô tả phân bố truy

cập các địa chỉ URL của các website. Từ đó ta biết được nhu cầu truy cập của khách hàng, khách hàng có nhu cầu tìm hiểu về vấn đề gì của website. Hình 3.16 mô tả phân bố nguồn (client) truy cập theo nước. Từ đó sẽ xác định được đối tượng truy cập vào website, khách hàng phân bổ ở nước nào trên thế giới. Hình 3.18 mô tả thống kê lỗi truy cập theo host / URL. Từ đó xác định được các lỗi truy cập của từng địa chỉ URL, sau đó sẽ có hướng nâng cấp, sửa chữa website kịp thời.

3.2.3. Nhận xét, đánh giá

Mô hình hệ thống xử lý và phân tích log thử nghiệm sử dụng ELK Stack đã được cài đặt và chạy thử thành công. Hệ thống cung cấp các tính năng:

- Thu thập dữ liệu web log từ các máy chủ web sử dụng filebeat và vận chuyển log về máy chủ ELK.

- Logstash được tích hợp các bộ lọc grok cho phép tiền xử lý và chuẩn hóa các loại dữ liệu web log, như IIS log, hoặc Apache log.

- Cung cấp các chức năng quản lý, lập chỉ số, lưu trữ và tìm kiếm dữ liệu web log.

- Phân tích dữ liệu log và biểu diễn kết quả ở nhiều dạng biểu đồ, đồ thị khác nhau có tính minh họa cao.

- Hỗ trợ các tính năng phân lớp và phát hiện bất thường trong dữ liệu log.

3.3. Kết luận chương

Chương 3 đã mô tả việc triển khai thử nghiệm hệ thống xử lý và phân tích web log, bao gồm giới thiệu mô hình tổng quát của hệ thống, mô hình triển khai thử nghiệm hệ thống, vấn đề cài đặt hệ thống xử lý log dựa trên ELK, việc thử nghiệm và các kết quả.

KẾT LUẬN

Các kết quả đạt được

Luận văn tập trung nghiên cứu, khảo sát các kỹ thuật và công cụ phân tích web log, đồng thời triển khai thử nghiệm một hệ thống quản lý và phân tích log thương mại cũng như mã mở. Cụ thể luận văn đã thực hiện các nội dung sau:

- Giới thiệu khái quát về web log, các định dạng web log, vấn đề xử lý và phân tích web log và ứng dụng của phân tích web log.

- Trình bày mô hình và các kỹ thuật xử lý và phân tích web log.

- Khảo sát một số công cụ xử lý và phân tích web log thương mại và mã mở tiêu biểu.

- Xây dựng và triển khai thử nghiệm một mô hình hệ thống thu thập, xử lý và phân tích log sử dụng ELK Stack và đánh giá kết quả.

Hướng phát triển của luận văn

Luận văn này có thể được phát triển tiếp theo các hướng sau:

- Tích hợp thêm các thành phần thu thập và tiền xử lý log, cho phép xử lý và phân tích các dạng web log khác, cũng nhưng các dạng log của hệ thống và các dịch vụ, ứng dụng.

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

Tiếng Việt:

[1] Phạm Duy Lộc, Hoàng Xuân Dậu (2018), Khảo sát các nền tảng và kỹ thuật xử lý log truy cập dịch vụ mạng cho phát hiện nguy cơ mất an toàn thông tin, Tạp chí khoa học Đại học Đà lạt, Tập 8, Số 2, 2018, trang 89–108.

[2] VNCS (2018) - Giải pháp giám sát website tập trung, http://vncs.vn/portfolio/giai-phap-giam-sat-websites-tap-trung, truy cập tháng 11.2018.

Tiếng Anh:

[3] Roger Meyer (2008), Detecting Attacks on Web Applications from Log Files, SANS Institute.

[4] Shaimaa Ezzat Salama, Mohamed I. Marie, Laila M. El-Fangary, Yehia K. Helmy (2011), Web Server Logs Preprocessing for Web Intrusion Detection, journal of Computer and Information Science Vol. 4, No. 4, July 2011, Canadian Center of Science and Education.

[5] Faradzhullaev, R. (2008). Analysis of Web server log files and attack detection. Journal of Automatic Control and Computer Sciences, 42(1), 50-54.

Trang web:

[6] OSSEC, https://www.ossec.net/, truy cập tháng 10.2021.

[7] Splunk, https://www.splunk.com, truy cập tháng 10.2021.

[8] IBM QRadar SIEM, https://www.ibm.com/products/qradar-siem, truy cập tháng 10.2021.

[9] Graylog, https://www.graylog.org, truy cập tháng 10.2021.

[10] ELK Stack, https://www.elastic.co/what-is/elk-stack, truy cập tháng 10.2021.

[11] LOGalyze, https://sourceforge.net/software/product/LOGalyze/, truy cập tháng 10.2021.

[12] Rsyslog (2018), https://www.rsyslog.com, truy cập tháng 11.2018

[13] NXLog (2018), https://nxlog.co, truy cập tháng 11.2018.

[14] Elastichsearch (2018), https://www.elastic.co, truy cập tháng 11.2018.