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

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - NĂM 2022

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS. TS. HOÀNG XUÂN DẬU

(Ghi rõ học hàm, học vị)

Phản biện 1: PGS.TS Đỗ Trung Tuấn

Phản biện 2: PGS.TS Nguyễn Hữu Quỳnh

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 8 giờ 30 ngày 02 tháng 07 năm 2022

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

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.

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***".

2. Tổng quan về vấn đề nghiên cứu:

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.

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. 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.

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

- *Máy chủ Web*
- *Máy chủ proxy web*
- *Trình duyệt máy khách*

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.

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.

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

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

W3C Extended Log Format

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

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.

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.

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. Theo đó, các khâu xử lý cụ thể gồm:

- *Collection of Log Data* : 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.

- *Cleaning of Data*: 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.

- *Convert into Strutured form* : 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.

- *Analysis of Data:* 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.

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

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 và (4) hỗ trợ hiểu hành vi của người dùng trực tuyến.

Có thể thấy, phân tích log truy cập có thể hỗ trợ việc giám sát, kiểm tra việc tuân thủ các chính sách bảo mật, chính sách kiểm toán của cơ quan, tổ chức. Hơn nữa phân tích log truy cập có thể hỗ trợ phản ứng lại các sự cố an toàn thông tin thông qua việc hỗ trợ xác định nguyên nhân và yếu tố gây mất an toàn.

Hỗ trợ khắc sự cố hệ thống cũng là một trong các ứng dụng quan trọng của phân tích log truy cập. Phân tích log truy cập giúp loại bỏ bớt các dữ liệu nhiễu, tổng hợp các thông báo lỗi riêng lẻ, giúp xác định nguyên nhân của sự cố hệ thống rõ ràng và chính xác hơn và trên cơ sở đó người quản trị có thể đưa ra biện pháp khắc phục sự cố phù hợp.

Phân tích log truy cập cũng có thể hỗ trợ điều tra số thông qua việc lần vết, xâu chuỗi các sự kiện log riêng lẻ sử dụng các kỹ thuật khai phá dữ liệu và phân tích tương quan.

Hỗ trợ hiểu được hành vi người dùng trực tuyến là một trong các mục đích chính trong phân tích log truy cập, nhất là phân tích log truy cập các website hay web log. Phân tích web log có thể tạo ra các báo cáo sử dụng các trang web của người dùng, bao gồm lưu lượng truy nhập, các trang tham chiếu, phân bố người dùng theo vị trí địa lý và lượng dữ liệu tải xuống. Đồng thời, phân tích log truy cập cũng giúp trích xuất nhiều thông tin quan trọng về hành vi người dùng trực tuyến và trên cơ sở đó có thể hỗ trợ việc tối ưu hóa website, nhằm nâng cao chất lượng dịch vụ cung cấp và trải nghiệm người dùng.

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

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

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.

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 bao gồm bốn giai đoạn phụ :

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

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
- Các yêu cầu của rô bốt
- Các yêu cầu lỗi

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.

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.

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. 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 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.

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.

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.

2.1.3. Các kỹ thuật phân tích web log

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 cách phân tích tệp phiên người dùng, chúng ta có thể thực hiện các phương pháp thống kê khác nhau như tính trung bình, tần suất ... với các biến số khác nhau như: số trang đã xem, số lượt xem, thời gian xem trên mỗi trang.

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

Luật kết hợp

Phương pháp này được sử dụng để khám phá các luật kết hợp giữa các phần tử dữ liệu trong CSDL. Mẫu đầu ra của thuật toán khai phá dữ liệu là tập luật kết hợp được tìm thấy.

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.

Phân cụm – Clustering

Phân cụm là một kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc về lớp phương pháp Unsupervised Learning trong Học máy. Có nhiều định nghĩa khác nhau về kỹ thuật này nhưng 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.

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 này, một trong những 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.

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.

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

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. 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. 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.

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 .

2.2.3. ELK Stack

ELK Stack là tập hợp 3 phần mềm đi chung với nhau, phục vụ cho công việc logging. 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

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. 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. 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.

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. 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. 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, 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. .

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

Bảng 2.1: So sánh các công cụ và nền tảng phân tích web log

Nền tảng	Ưu điểm	Nhược điểm
IBM QRadar SIEM	- Hỗ trợ thu thập và xử lý nhiều loại log khác nhau với khối lượng lớn và dữ liệu từ luồng mạng - Hỗ trợ phát hiện các bất thường, các nguy cơ ATTT với độ chính xác cao và tỷ lệ cảnh báo sai thấp	- 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
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, cũng như 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 - Thu thập được log từ rất nhiều nguồn khác nhau: log hệ thống, log ứng dụng, log thiết bị mạng, log snmp, 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	- Không có khả năng phân tích chuyên sâu các nguy cơ mất an toàn thông tin, như dấu hiệu xuất hiện các dạng mã độc và các dạng tấn công lên các dịch vụ và tài nguyên mạng. - Hỗ trợ số lượng ít các loại nhật ký

Nền tảng	Ưu điểm	Nhược điểm
	xuất nhật ký trực quan dữ liệu và truy vấn. - 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 có khả năng phân tích chuyên sâu các nguy cơ mất an toàn thông tin, như dấu hiệu xuất hiện các dạng mã độc và các dạng tấn công lên các dịch vụ và tài nguyên mạng. - 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

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 qui.

- 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

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 . 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.

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".

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"
```

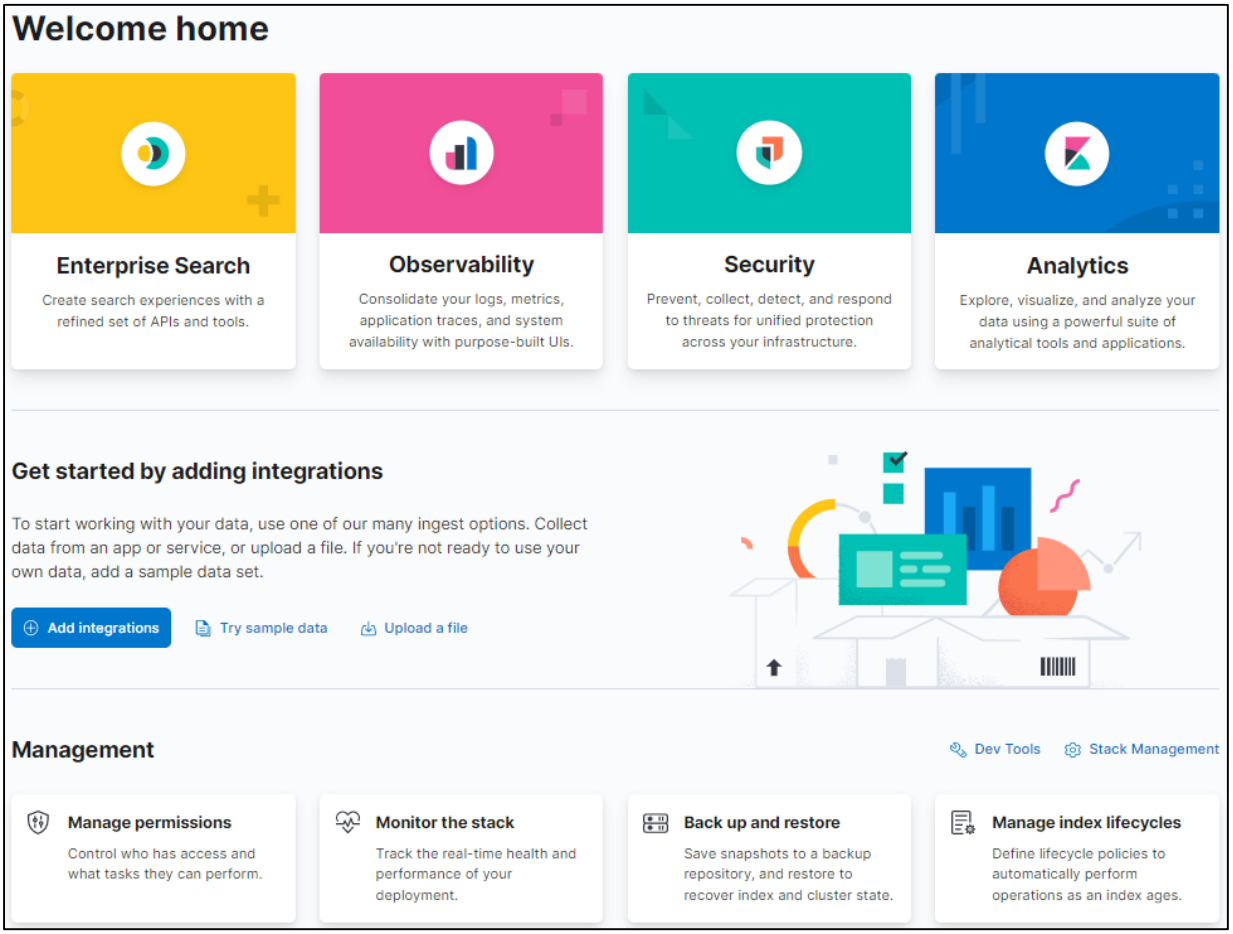
Hình 3.5. Một số bản ghi của web log mẫu cung cấp bởi ELK

```
#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
```

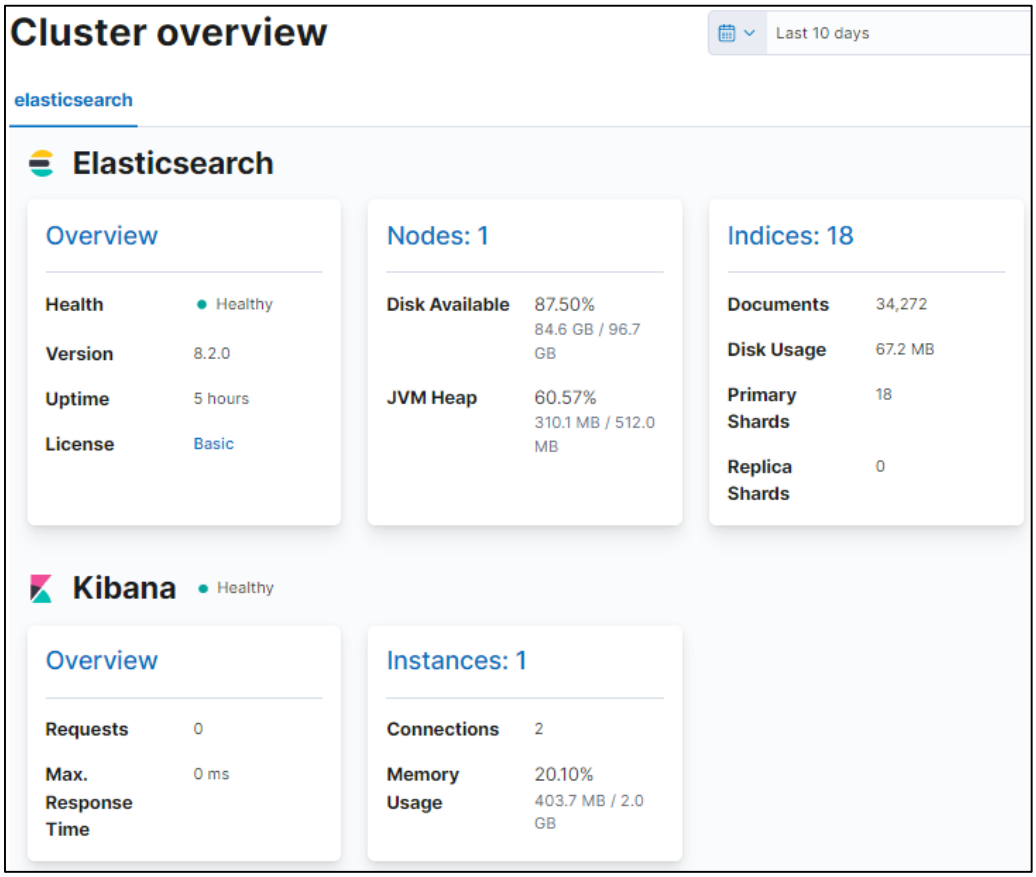
Hình 3.6. Một số bản ghi của Microsoft IIS log

3.2.2. Một số kết quả

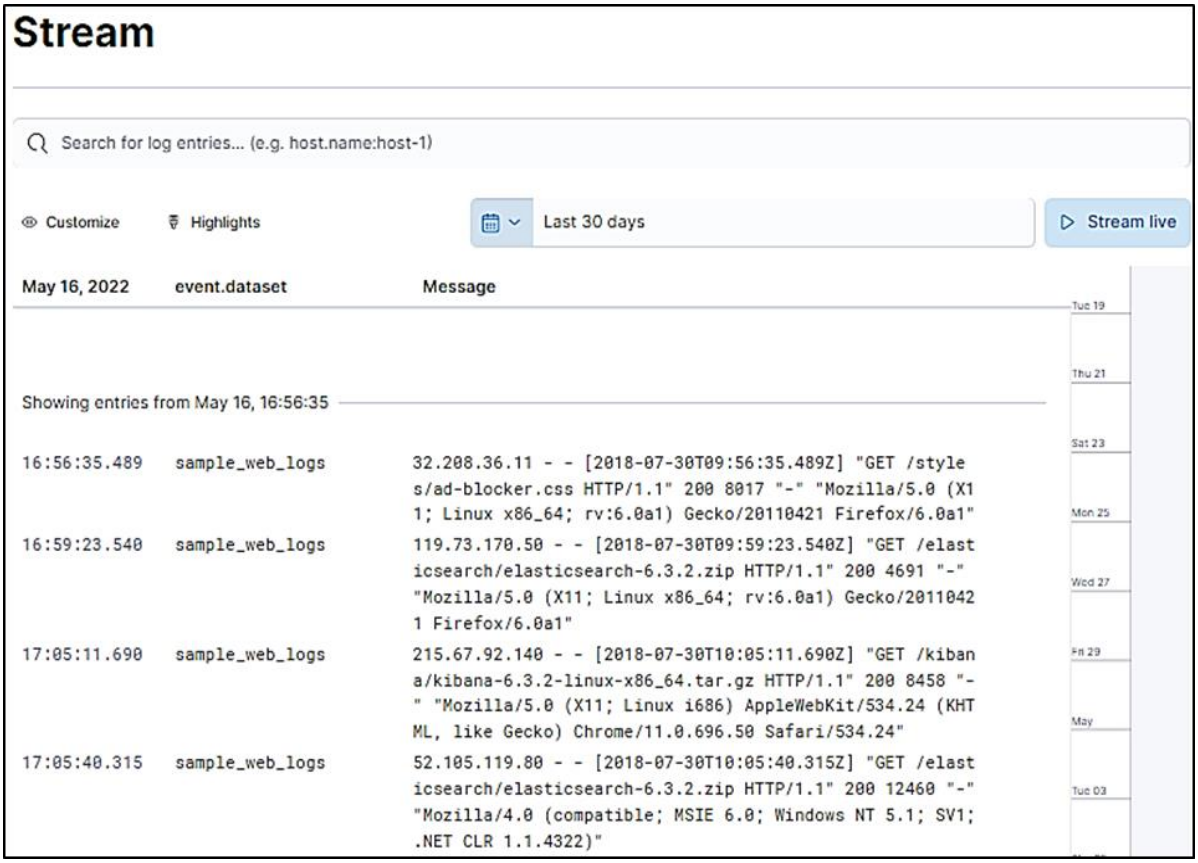
Dưới đây là các giao diện và kết quả thử nghiệm phân tích web log:



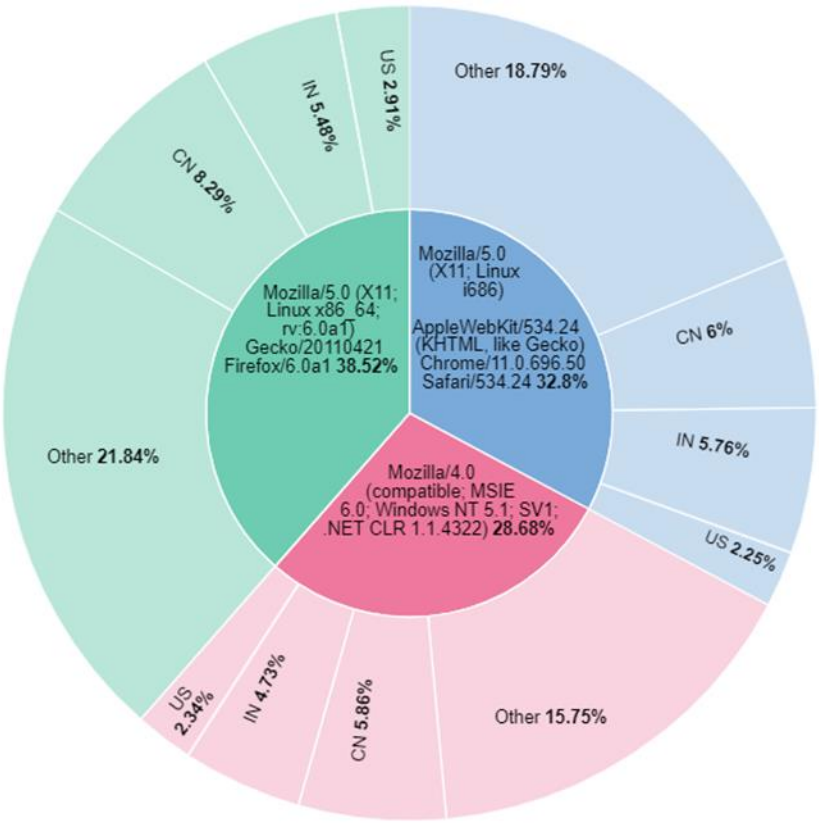
Hình 3.7. Giao diện trang chủ của Kibana



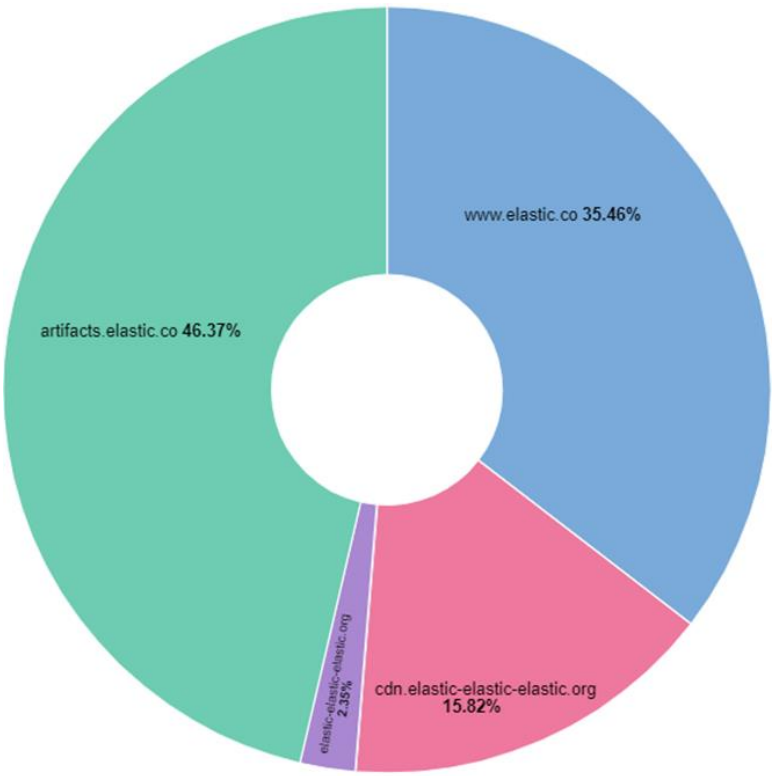
Hình 3.8. Trạng thái hoạt động của ELK Stack



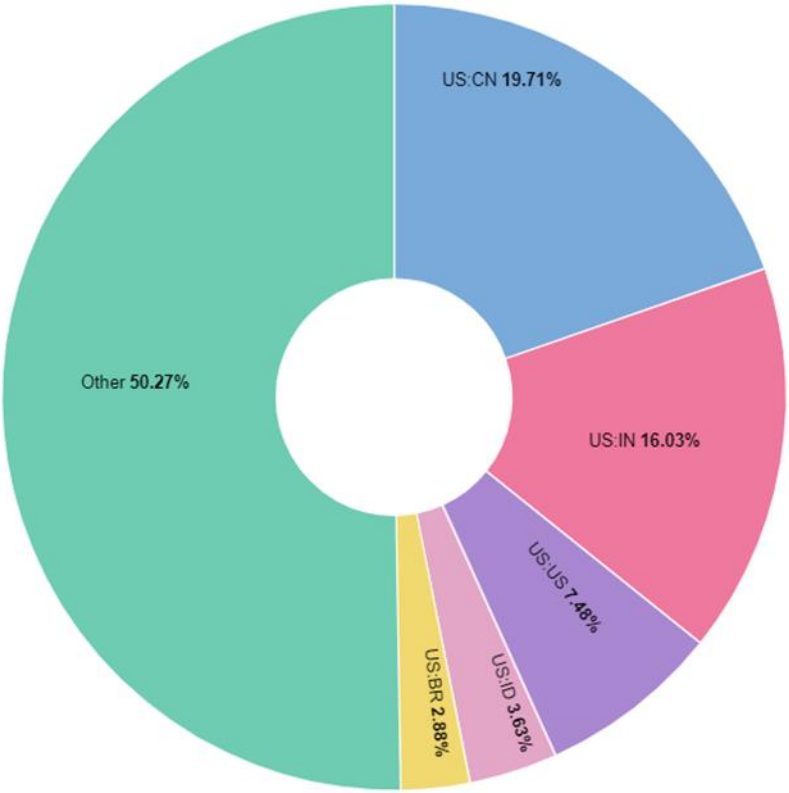
Hình 3.9. Luồng log thu thập trong 30 ngày gần đây



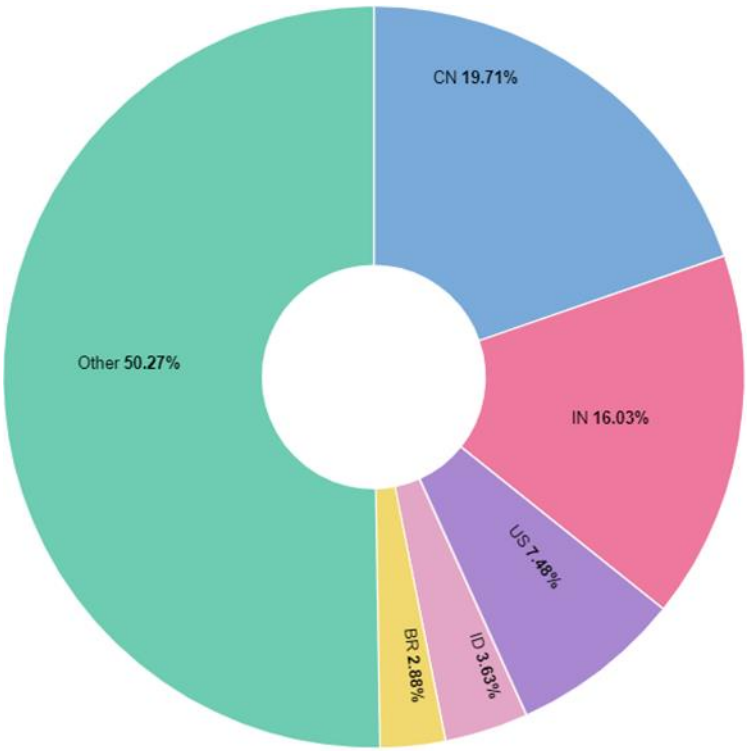
Hình 3.12. Phân bố các loại trình duyệt kèm nơi máy khách truy cập website



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

[Logs] Errors by host

URL	Visits	Unique	HTTP 4xx	↓ HTTF	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

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.