HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

VƯƠNG MINH VIỆT

NGHIÊN CỨU XÂY DỰNG HỆ THỐNG PHÂN TÍCH LOG TRUY NHẬP CHO PHÁT HIỆN BẤT THƯỜNG VÀ CÁC NGUY CƠ AN TOÀN THÔNG TIN

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – 2019

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

VƯƠNG MINH VIỆT

NGHIÊN CỨU XÂY DỰNG HỆ THỐNG PHÂN TÍCH LOG TRUY NHẬP CHO PHÁT HIỆN BẤT THƯỜNG VÀ CÁC NGUY CƠ AN TOÀN THÔNG TIN

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 – 2019

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi, kết quả đạt được trong luận văn là sản phẩm của riêng cá nhân, không sao chép lại của người khác. Trong toàn bộ nội dung của luận văn, những điều được trình bày hoặc là của cá nhân hoặc là được tổng hợp từ nhiều nguồn tài liệu. Tất cả các tài liệu tham khảo đều có xuất xứ rõ ràng và được trích dẫn hợp pháp. Các số liệu, kết quả nêu trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Hà nội, ngày tháng năm 2019

Học viên

Vương Minh Việt

LỜI CẢM ƠN

Để có thể hoàn thiện được luận văn thạc sĩ của mình, trước tiên, tôi xin bày tỏ lòng biết ơn sâu sắc nhất tới thầy - TS. Hoàng Xuân Dậu (Khoa Công nghệ thông tin, Học viện Công nghệ Bưu chính Viễn thông). Sự gần gũi và nhiệt tình hướng dẫn của thầy là nguồn độnglực rất lớn đối với tôi trong suốt thời gian thực hiện luận văn.

Tôi cũng xin gửi lời cảm ơn chân thành nhất tới tất cả các thầy, cô trong khoa Công nghệ thông tin; khoa Đào tạo sau đại học Học viện Công nghệ Bưu chính Viễn thông đã nhiệt tình giảng dạy, cung cấp, hướng dẫn cho chúng tôi những kiến thức, kinh nghiệm trong suốt quá trình học tập.

Đồng thời tôi cũng xin gửi lời cảm ơn đến người thân trong gia đình, các bạn học viên, đồng nghiệp nơi tôi công tác đã giúp đỡ, động viên, tạo điều kiện tốt nhất cho tôi trong suốt khóa học tại Học viện Công nghệ Bưu chính Viễn thông để tôi có thể hoàn thiện tốt luận văn thạc sĩ của mình.

MỤC LỤC

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH VẼ, ĐỒ THỊ

DANH MỤC CÁC KÝ HIỆU, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
API	Application Programming	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 tải siêu văn bản
ISP	Internet Service Provider	Nhà cung cấp dịch vụ Internet
JSON	JavaScript Object Notation	Một kiểu dữ liệu mở trong JavaScrip
LAN	Local area network	Mạng máy tính cục bộ
LDAP	Lightweight Directory Access Protocol	Một giao thức ứng dụng truy cập các cấu trúc thư mục
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 mang tính cấu
TCP	Transmission Control Protocol	Giao thức điều khiển truyền vận
UDP	User Datagram Protocol	Giao thức dữ liệu người dùng
UI	User Interface	Giao diện người dùng
URI	Uniform Resource Identifier	Mã định danh tài nguyên thống nhất
URL	Uniform Resource Locator	Đường dẫn tham chiếu tới tài nguyên mạng trên Internet
W3C	World Wide Web Consortium	Tên tổ chức quốc tế W3C

MỞ ĐẦU

Log (còn gọi là nhật ký, hay vết) là các mục thông tin do hệ điều hành, hoặc các ứng dụng sinh ra trong quá trình hoạt động. Mỗi bản ghi log thường được sinh ra theo 1 hoạt động, hoặc sự kiện, nên còn được gọi là nhật ký sự kiện (event log). Các nguồn sinh log phổ biến bao gồm các thiết bị mạng (như router, firewall,…), hệ điều hành, các máy chủ dịch vụ (máy chủ web, máy chủ cơ sở dữ liệu, máy chủ DNS, email,…) và các chương trình ứng dụng. Mục đích của việc thu thập, xử lý và phân tích log bao gồm:

- Kiểm tra sự tuân thủ các chính sách an ninh;
- Kiểm tra sự tuân thủ vấn đề kiểm toán và luật pháp;
- Phục vụ điều tra số;
- Phục vụ phản ứng các sự cố mất an toàn thông tin ;
- 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.

Log ghi lại liên tục các thông báo về hoạt động của cả hệ thống hoặc của các dịch vụ được triển khai trên hệ thống vào các log file. Log file thường là các file văn bản thông thường dưới dạng “clear text” tức là bạn có thể dễ dàng đọc được nó, vì thế có thể sử dụng các trình soạn thảo văn bản (vi, vim, nano...) hoặc các trình xem văn bản thông thường (cat, tailf, head...) là có thể xem được các file log.

Việc xử lý và phân tích log có nhiều ứng dụng, đặc biệt trong đảm bảo an toàn thông tin và cải thiện chất lượng hệ thống và các dịch vụ kèm theo, như quảng cáo trực tuyến. Hiện nay, trên thế giới đã có một số nền tảng và công cụ cho thu thập, xử lý và phân tích các dạng log phiên bản thương mại cũng như mã mở, như IBM Qradar SIEM, Splunk, Graylog và Logstash,... Tuy nhiên, việc nghiên cứu sâu các phương pháp xử lý và phân tích log và ứng dụng ở Việt Nam vẫn cần được tiếp tục thực hiện nhằm xây dựng các mô hình, hệ thống xử lý và phân tích log hiệu quả với chi phí hợp lý. Đây cũng là mục đích của đề tài luận văn *“Nghiên cứu xây dựng hệ thống phân tích log truy nhập cho phát hiện bất thường và các nguy cơ an toàn thông tin”*.

Luận văn bao gồm ba chương chính với nội dung như sau:

- Chương 1: Tổng quan về phân tích log truy nhập: khái niệm log truy nhập, các dạng log truy nhập, các phương pháp thu thập, xử lý và phân tích log, ứng dụng của phân tích log và giới thiệu một số nền tảng, công cụ phân tích log.

- Chương 2: Các kỹ thuật và mô hình xử lý, phân tích log truy nhập: Mô hình xử lý log; Thu thập và tiền xử lý; Các kỹ thuật phân tích log: Các kỹ thuật nhận dạng mẫu (Pattern Discovery), phân tích mẫu (Pattern Analysis), phân tích tương quan (Correlation Analysis).

- Chương 3: Cài đặt, thử nghiệm và đánh giá: Giới thiệu môi trường và công cụ thử nghiệm; Cài đặt hệ thống: Cài đặt OSSEC, cài đặt ELK, kết hợp OSSEC và ELK; Nội dung thử nghiệm, kết quả và nhận xét.

CHƯƠNG 1. TỔNG QUAN VỀ PHÂN TÍCH LOG TRUY NHẬP

1.1. Khái quát về log truy nhập

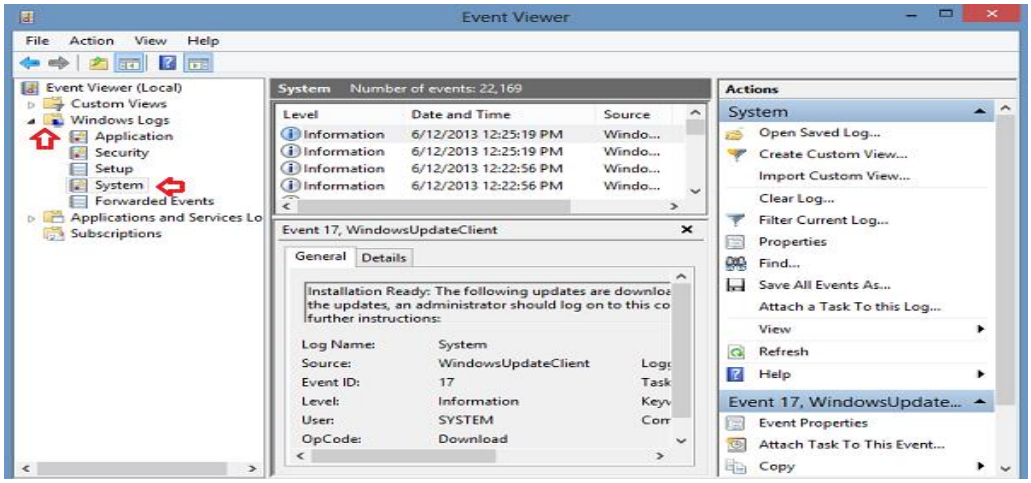
1.1.1. Khái niệm log truy nhập

Log truy nhập hay nhật ký, hoặc vết truy nhập (gọi tắt là log) là một danh sách các bản ghi mà một hệ thống ghi lại khi xuất hiện các yêu cầu truy nhập các tài nguyên của hệ thống [1]. Chẳng hạn, log truy nhập web (gọi tắt là web log) chứa tất cả các yêu cầu truy nhập các tài nguyên của một website. Các tài nguyên của một website, như các file ảnh, các mẫu định dạng và file mã JavaScript. Khi một người dùng thăm một trang web để tìm một sản phẩm, máy chủ web sẽ tải xuống thông tin và ảnh của sản phẩm và log truy nhập sẽ ghi lại các yêu cầu của người dùng đến các tài nguyên thông tin và ảnh của sản phẩm.

Có nhiều nguồn sinh log trong hệ thống, như log sinh bởi hệ điều hành, log sinh bởi các máy chủ dịch vụ mạng, log sinh bởi các ứng dụng và log sinh bởi các thiết bị mạng và thiết bị đảm bảo an toàn thông tin [1]. Log sinh bởi hệ điều hành thường bao gồm các bản ghi các sự kiện khởi động hệ thống, sự kiện đăng nhập, đăng xuất của người dùng, yêu cầu truy nhập các file, các thư mục, các yêu cầu kích hoạt ứng dụng, các yêu cầu truy nhập phần cứng, các yêu cầu truy nhập dịch vụ mạng, các lỗi xuất hiện trong quá trình hoạt động... Hệ điều hành Microsoft Windows sử dụng công cụ Event Viewer (Hình 1.1), còn các hệ điều hành thuộc họ Unix/Linux sử dụng công cụ Syslog để quản lý và lưu trữ log do bản thân hệ điều hành và các module phụ trợ sinh ra.

Nguồn log sinh bởi các máy chủ dịch vụ mạng, máy chủ web, máy chủ DNS, máy chủ email và máy chủ cơ sở dữ liệu là một trong các nguồn log phổ biến nhất. Máy chủ web có thể ghi log truy nhập các trang web cho từng website dưới dạng các file văn bản thuần với mỗi dòng là một bản ghi log. Các thông tin trong mỗi bản ghi web log có thể khác nhau phụ thuộc vào phiên bản máy chủ web sử dụng. Các máy chủ tên miền DNS cũng sinh một lượng lớn log trong quá trình xử lý các yêu cầu phân giải tên miền sang địa chỉ IP và ngược lại từ người dùng. Tương tự, các máy chủ email và cơ sở dữ liệu cũng sinh rất nhiều bản ghi log trong quá trình xử lý các

yêu cầu từ người dùng cũng như từ các ứng dụng. Hình 1.2 biểu diễn các bản ghi log tạo bởi máy chủ e-mail.



Hình 1.1. Xem Windows log sử dụng công cụ Event Viewer

Hình 1.2. Các bản ghi log tạo bởi máy chủ e-mail.

Các thiết bị mạng và các hệ thống đảm bảo an toàn thông tin cũng là một trong các nguồn sinh nhiều log. Các thiết bị mạng phổ biến như các bộ định tuyến (router), các bộ chuyển mạch (switch) và các hệ thống đảm bảo an toàn thông tin, như tường lửa (Firewall), các hệ thống phát hiện và ngăn chặn tấn công, xâm nhập các hệ thống điều khiển truy nhập, cũng sinh nhiều bản ghi log trong quá trình xử lý các yêu cầu truy nhập mạng. Log sinh từ các hệ thống này có thể được lưu tại chỗ, hoặc xuất ra các hệ thống lưu trữ bên ngoài.

Như vậy có thể thấy, có nhiều nguồn sinh dữ liệu log truy nhập với nhiều dạng khác nhau. Tùy vào mục đích sử dụng, 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 mỗi dạng log.

1.1.2. Các dạng log truy nhập

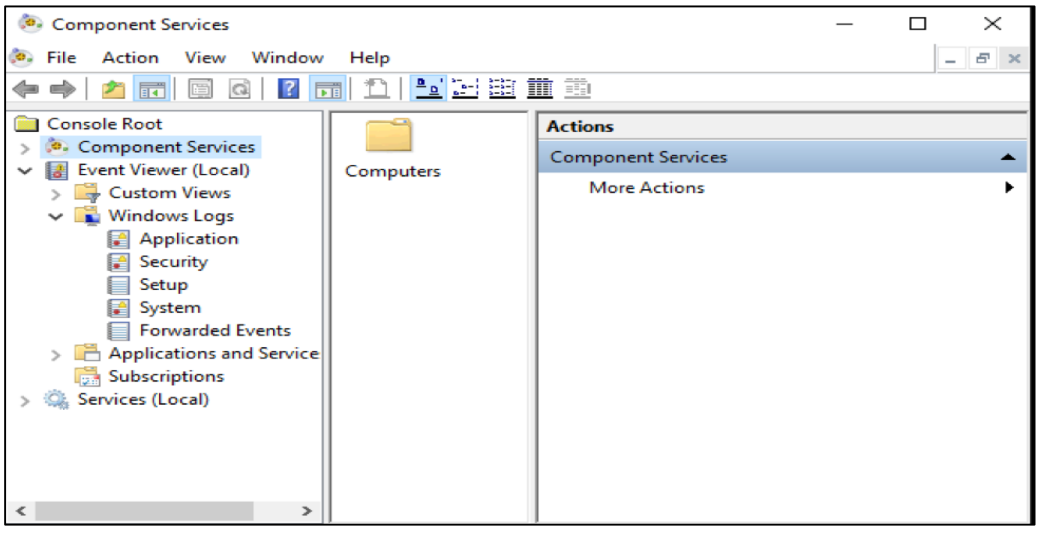
Như đã đề cập, có nhiều nguồn sinh log trong hệ thống, như log sinh bởi hệ điều hành, log sinh bởi các máy chủ dịch vụ mạng và log sinh bởi các thiết bị mạng và thiết bị đảm bảo an toàn thông tin. Mục này trình bày khái quát về các dạng log này.

1.1.2.1. Log sinh bởi hệ điều hành

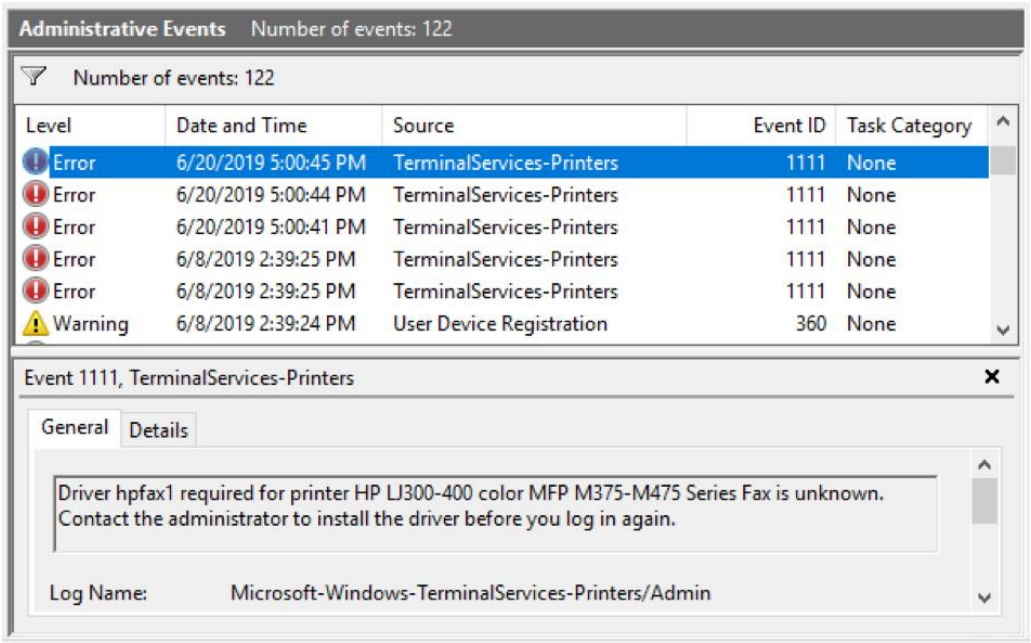
Log sinh bởi hệ điều hành (gọi tắt là log hệ diều hành) gồm các dạng log sinh bởi nhân hệ điều hành và các thành phần thuộc hệ điều hành. Log hệ điều hành được liên tục sinh ra và lưu trong hệ thống trong quá trình khởi động và hoạt động của hệ thống. Mỗi họ hệ điều hành (Windows, Linux, Unix,…) có định dạng và phương pháp quản lý log riêng. Mục này cung cấp mô tả chi tiết hơn về Windows logs và Linux/Unix log.

a. Windows logs

Log của hệ điều hành Microsoft Windows (Windows Logs) gồm 6 thành phần chính có thể truy nhập và duyệt bởi trình Windows Event Viewer như biểu diễn trên Hình 1.3:



Hình 1.3. Các thành phần của Windows Logs [2]



Hình 1.4. Một bản ghi Windows log mô tả lỗi dịch vụ [2]

- Log ứng dụng (Application) là log sinh bởi các ứng dụng chạy trên Windows;
- Log bảo mật (Security) là log sinh bởi các dịch vụ bảo mật của Windows, như xác thực, cấp quyền, quản trị người dùng,…
- Log cài đặt (Setup) là log sinh trong quá trình cài đặt các thành phần của hệ điều hành và các ứng dụng;
- Log hệ thống (System) là log sinh bởi các tính năng và các dịch vụ nền của Windows, như quản lý tiến trình, quản lý hệ thống file, dịch vụ mạng,... Hình 1.4 minh họa một bản ghi Windows log mô tả một lỗi vận hành dịch vụ.
- Các sự kiện chuyển tiếp (Forwarded Events) là các sự kiện log được chuyển tiếp từ các máy khác khi máy Windows là trung tâm quản lý.

b. Linux/Unix logs

Hầu hết các hệ điều hành thuộc họ Linux/Unix đều được trang bị một hệ thống ghi log rất mạnh và mềm dẻo, cho phép ghi lại tất cả các sự kiện xảy ra trong hệ thống. Công cụ quản lý log được sử dụng rộng rãi nhất trong các hệ điều hành thuộc

họ Linux/Unix là *syslog*. Syslog là công cụ quản lý log tập trung, có thể cấu hình được thông qua tập tin cấu hình syslog.conf, như minh họa trên Hình 1.5.

```
*.err;kern.debug;auth.notice /dev/console
daemon,auth.notice           /var/log/messages
lpr.info                     /var/log/lpr.log
mail.*                       /var/log/mail.log
ftp.*                        /var/log/ftp.log
auth.*                       @prep.ai.mit.edu
auth.*                       root,amrood
netinfo.err                  /var/log/netinfo.log
install.*                    /var/log/install.log
*.emerg                      *
*.alert                      |program_name
mark.*                       /dev/console
```

Hình 1.5. Một phần tập tin cấu hình syslog - syslog.conf

Syslog hỗ trợ quản lý log từ nhiều nguồn, gồm các thành phần của hệ điều hành, dịch vụ và ứng dụng. Một số dạng log tiêu biểu như auth (log xác thực), console (log gửi tới console), cron (log sinh bởi trình lập lịch cron), daemon (log sinh bởi các tiến trình thường trú trong hệ thống), kern (log sinh bởi nhân hệ điều hành),… Hình 1.6 biểu diễn một số bản ghi kern log của hệ điều hành Linux.

```
Jun  1 22:20:05 secserv kernel: Kernel logging (proc) stopped.
Jun  1 22:20:05 secserv kernel: Kernel log daemon terminating.
Jun  1 22:20:06 secserv exiting on signal 15
Nov 27 08:05:57 galileo kernel: Kernel logging (proc) stopped.
Nov 27 08:05:57 galileo kernel: Kernel log daemon terminating.
Nov 27 08:05:57 galileo exiting on signal 15
```

Hình 1.6. Một số bản ghi kern log của hệ điều hành Linux

1.1.2.2. Log sinh bởi các dịch vụ mạng

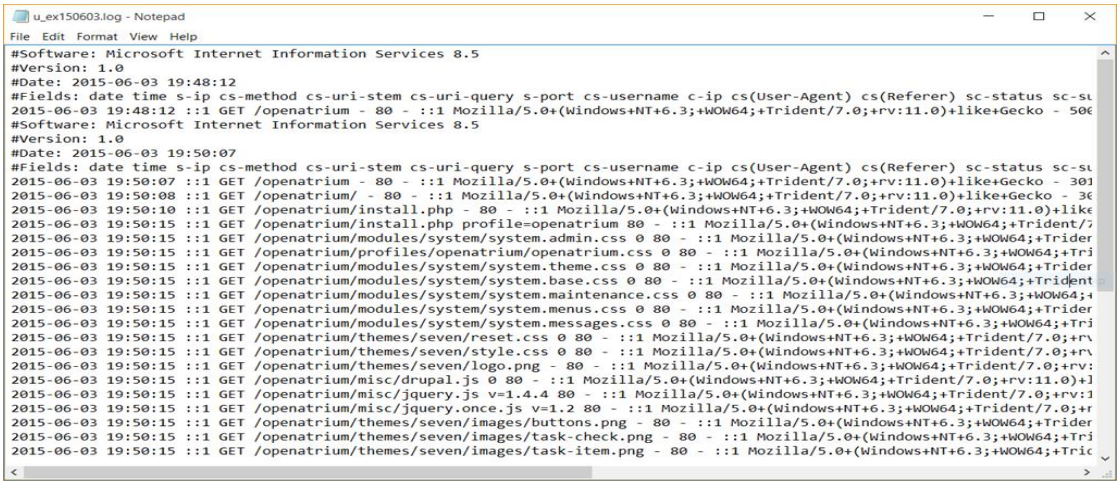
Log sinh bởi các dịch vụ mạng là một trong các nguồn sinh nhiều dữ liệu log nhất và được sử dụng rộng rãi trong xử lý và phân tích log nhất [1]. Các dịch vụ mạng phổ biến như dịch vụ web, dịch vụ DNS, dịch vụ email, dịch vụ máy chủ CSDL đều là các dịch vụ được sử dụng rất rộng rãi và sinh nhiều log. Mỗi dịch vụ mạng có định dạng log riêng và có phương pháp quản lý log riêng. Mục này đề cập đến log sinh bởi

các máy chủ web (Web log), log sinh bởi các máy chủ DNS (DNS log), log sinh bởi các máy chủ email (Mail Log) và log sinh bởi các máy chủ CSDL (Database log).

a. Web log

Các máy chủ web thông dụng như Mozilla Apache và Microsoft IIS thường hỗ trợ nhiều định dạng log, bao gồm W3C Extended log file format, Microsoft IIS log file format, and NCSA log file format, trong đó W3C Extended log file format là định dạng web log hỗ trợ bởi hầu hết máy chủ web và được sử dụng rộng rãi nhất. Hình 1.7 minh hoạt một phần file log theo định dạng W3C Extended log file format. Theo đó, các trường thông tin cơ bản của mỗi bản ghi web log mà định dạng này hỗ trợ bao gồm:

- Date: ngày tháng ghi log
- Time: thời gian ghi log
- S-ip: địa chỉ IP của máy chủ web
- S-port: số hiệu cổng dịch vụ máy chủ web
- Cs-method: phương thức HTTP thực hiện yêu cầu (GET, POST, HEAD,…)
- Cs-uri-stem: địa chỉ URI của trang, hoặc thành phần của trang yêu cầu
- Cs-uri-query: phần truy vấn của yêu cầu
- Cs-username: tên người dùng
- C-ip: địa chỉ IP của máy khách
- Cs(Referrer): trang, hoặc địa chỉ tham chiếu
- Cs(User-Agent): thông tin trình duyệt, hoặc máy khách web
- Sc-bytes: số byte máy chủ gửi trả lời
- Cs-bytes: số byte máy chủ nhận được
- Sc-status: mã thực hiện yêu cầu,…



Hình 1.7. Một phần file log theo định dạng W3C Extended log file format

b. DNS log

Các máy chủ DNS thông dụng như ISC Bind và Microsoft DNS server đều được trang bị khả năng ghi log theo các yêu cầu phân giải địa chỉ từ người dùng. DNS log cũng thường được lưu ở dạng file văn bản thuần và định dạng log do máy chủ DNS ISC Bind được coi như chuẩn thực tế được sử dụng rộng rãi nhất. Hình 1.8 biểu diễn trích xuất một số bản ghi DNS log. Các trường thông tin cơ bản của mỗi bản ghi DNS log mà định dạng này hỗ trợ bao gồm:

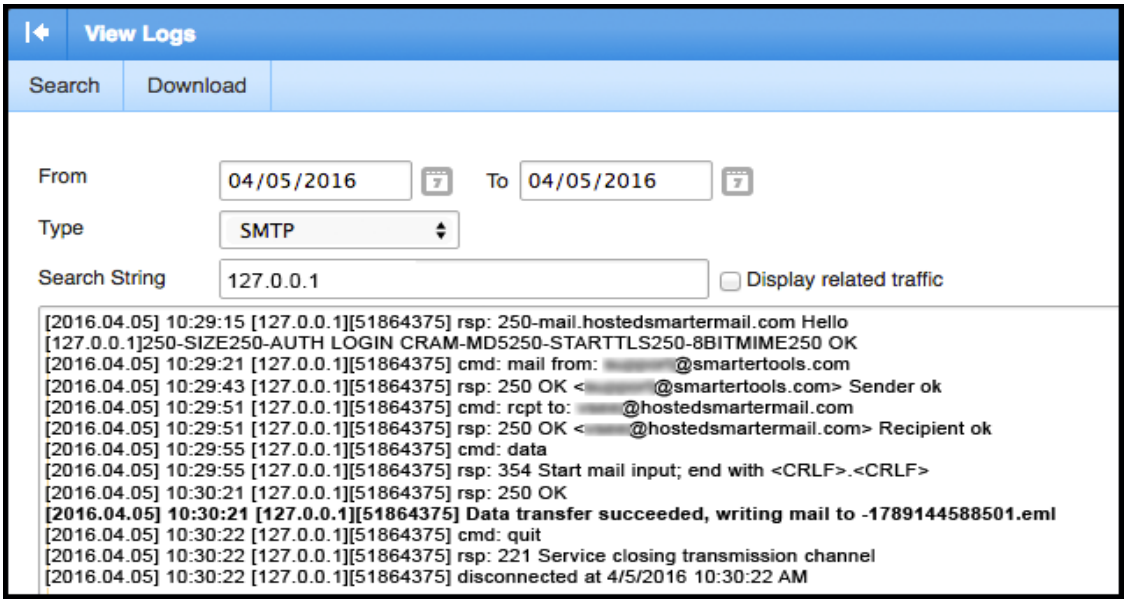
- Date: ngày gửi yêu cầu
- Time: thời gian yêu cầu
- Client IP Address: địa chỉ IP máy khách
- Client Port: cổng máy khách
- Server IP Address: địa chỉ IP máy chủ
- Server Port: cổng dịch vụ máy chủ
- Request Type: loại yêu cầu truy vấn
- Request Name: tên miền truy vấn
- Answer Type: loại trả lời yêu cầu
- Answer Data: dữ liệu trả lời từ máy chủ,…

No.	Time	Source	Destination	Protocol	Length	Info
803	4.416709000	192.168.2.52	192.168.2.4	DNS	85	Standard query 0x8040 A cooking.stackexchange.com
808	4.419638000	192.168.2.52	192.168.2.4	DNS	89	Standard query 0xcc0f A electronics.stackexchange.com
810	4.424106000	192.168.2.52	192.168.2.4	DNS	83	Standard query 0x967f A emacs.stackexchange.com
818	4.471499000	192.168.2.4	192.168.2.52	DNS	101	Standard query response 0x8040 A 198.252.206.16
819	4.472231000	192.168.2.52	192.168.2.4	DNS	85	Standard query 0xbb20 A gamedev.stackexchange.com
820	4.472263000	192.168.2.4	192.168.2.52	DNS	105	Standard query response 0xcc0f A 198.252.206.16
821	4.472725000	192.168.2.52	192.168.2.4	DNS	83	Standard query 0x8b07 A money.stackexchange.com
822	4.473017000	192.168.2.4	192.168.2.52	DNS	99	Standard query response 0x967f A 198.252.206.16
823	4.473380000	192.168.2.52	192.168.2.4	DNS	83	Standard query 0xe44a A music.stackexchange.com
832	4.518543000	192.168.2.4	192.168.2.52	DNS	99	Standard query response 0x8b07 A 198.252.206.16
833	4.519218000	192.168.2.52	192.168.2.4	DNS	86	Standard query 0x893d A outdoors.stackexchange.com
834	4.519302000	192.168.2.4	192.168.2.52	DNS	101	Standard query response 0xbb20 A 198.252.206.16
836	4.520024000	192.168.2.52	192.168.2.4	DNS	89	Standard query 0xefb3 A programmers.stackexchange.com
838	4.523652000	192.168.2.4	192.168.2.52	DNS	99	Standard query response 0xe44a A 198.252.206.16
839	4.524066000	192.168.2.52	192.168.2.4	DNS	86	Standard query 0x422b A puzzling.stackexchange.com
848	4.564736000	192.168.2.4	192.168.2.52	DNS	105	Standard query response 0xefb3 A 198.252.206.16
849	4.565487000	192.168.2.4	192.168.2.52	DNS	102	Standard query response 0x893d A 198.252.206.16
850	4.565516000	192.168.2.52	192.168.2.4	DNS	81	Standard query 0x6350 A rpg.stackexchange.com
851	4.566114000	192.168.2.52	192.168.2.4	DNS	84	Standard query 0x5bcd A travel.stackexchange.com
855	4.567001000	192.168.2.4	192.168.2.52	DNS	102	Standard query response 0x422b A 198.252.206.16
856	4.567692000	192.168.2.52	192.168.2.4	DNS	85	Standard query 0x0261 A tridion.stackexchange.com
867	4.606352000	192.168.2.4	192.168.2.52	DNS	101	Standard query response 0x0261 A 198.252.206.16
868	4.606361000	192.168.2.4	192.168.2.52	DNS	97	Standard query response 0x6350 A 198.252.206.16
869	4.607101000	192.168.2.4	192.168.2.52	DNS	100	Standard query response 0x5bcd A 198.252.206.16
870	4.607158000	192.168.2.52	192.168.2.4	DNS	84	Standard query 0x268d A area51.stackexchange.com
871	4.607175000	192.168.2.52	192.168.2.4	DNS	86	Standard query 0x234b A bicycles.stackexchange.com

Hình 1.8. Trích xuất một số bản ghi DNS log

c. Mail log

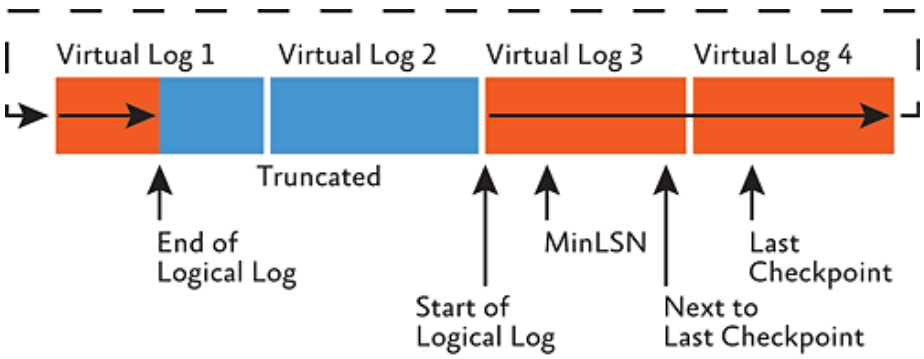
Các máy chủ email hỗ trợ các giao thức gửi nhận email như SMTP và giao thức tải email từ hộp thư về máy khách như POP và IMAP cũng thường sinh các bản ghi log khi thực thi các yêu cầu gửi nhận email. Tương tự máy chủ web, hầu hết các dạng log truy nhập máy chủ email đều ở dạng file văn bản thuần. Tuy nhiên, các dòng log được lưu dưới dạng một phiên đối thoại giữa máy chủ và máy khách email trong quá trình gửi và nhận một email. Hình 1.9 minh họa phần log truy nhập máy chủ email SMTP.



Hình 1.9. Một phần log truy nhập máy chủ email SMTP

d. Database log

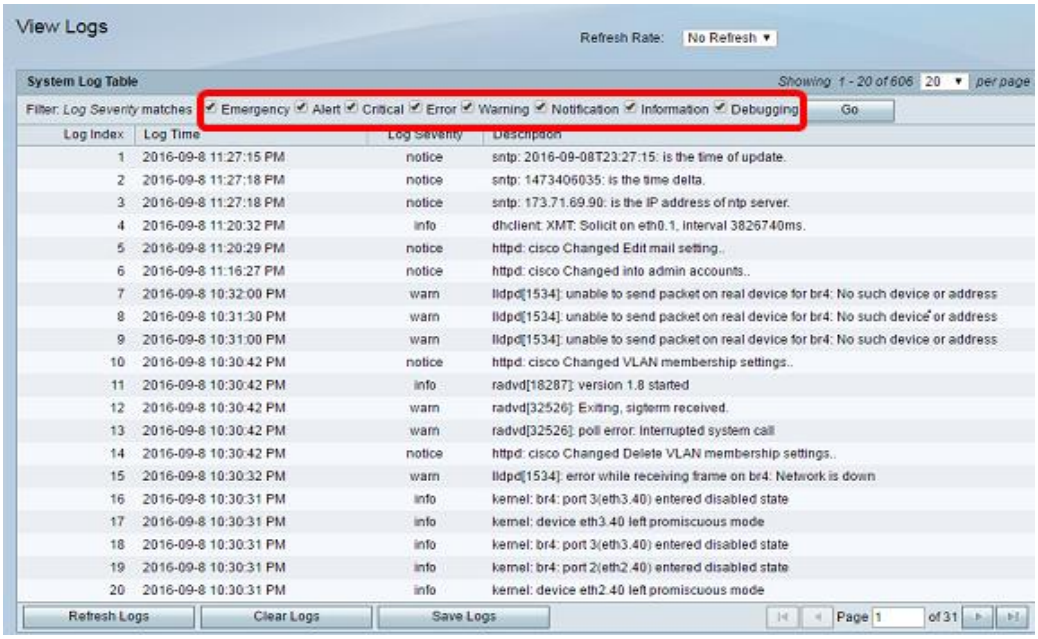
Các máy chủ cơ sở dữ liệu (Database server) thường sinh nhiều bản ghi log trong quá trình hoạt động. Dữ liệu log truy nhập máy chủ cơ sở dữ liệu thường được gọi là log giao dịch (Transaction log) được ghi khi xuất hiện các yêu cầu xử lý các thao tác: đăng nhập, đăng xuất của người dùng; các thao tác quản trị người dùng và cấp quyền truy nhập; các thao tác tạo lập, xóa cơ sở dữ liệu, tạo lập, xóa và thay đổi cấu trúc các bảng cơ sở dữ liệu; tạo lập, xóa và chỉnh sửa mã các thủ tục, hàm và các triger; và các thao tác truy vấn, thêm sửa và xóa các bản ghi trong các bảng cơ sở dữ liệu. Tùy thuộc vào mục đích sử dụng, người quản trị có thể cấu hình máy chủ cơ sở dữ liệu để sinh các bản ghi log theo yêu cầu với khối lượng log sinh ra phù hợp. Mỗi loại máy chủ cơ sở dữ liệu thường có một mô hình quản lý dữ liệu log riêng và Hình 1.10 biểu diễn mô hình quản lý dữ liệu log của máy chủ cơ sở dữ liệu Microsoft SQL Server.



Hình 1.10. Mô hình quản lý dữ liệu log của Microsoft SQL Server

1.1.2.3. Log sinh bởi các thiết bị mạng và thiết bị đảm bảo ATTT

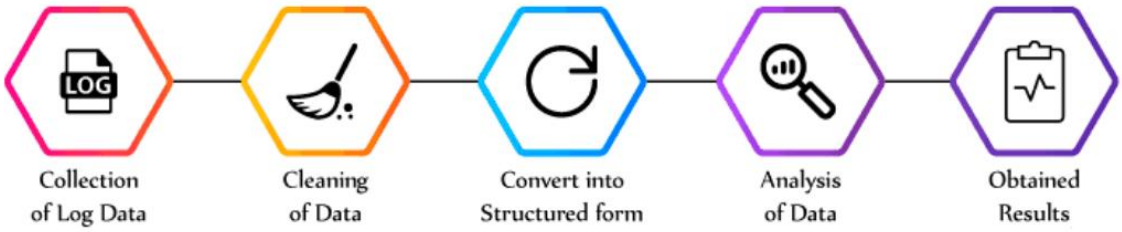
Log sinh bởi các thiết bị mạng và thiết bị đảm bảo an toàn thông tin gồm dữ liệu log sinh bởi các thiết bị mạng (như switch, router, load balancer,…) và các thiết bị/công cụ đảm bảo an toàn thông tin (như tường lửa, các hệ thống IDS/IPS, các hệ thống giám sát,…). Mỗi nhà cung cấp thiết bị mạng (như Cisco Systems, Juniper Networks,…) thường thiết kế định dạng log và phương pháp quản lý riêng cho các thiết bị mạng và đảm bảo ATTT của mình. Hình 1.11 minh hoạt một số bản ghi log của Cisco RV Series Router.



Hình 1.11. Một phần log của Cisco RV Series Router

1.2. Thu thập, xử lý và phân tích log truy nhập

Thu thập, xử lý và phân tích log là các khâu cơ bản của một hệ thống phân tích log. Hình 1.12 biểu diễn các khâu cụ thể của quá trình thu thập, xử lý và phân tích log thường được áp dụng trên thực tế. Theo đó, các khâu xử lý cụ thể gồm:



Hình 1.12. Các khâu của quá trình thu thập, xử lý và phân tích log

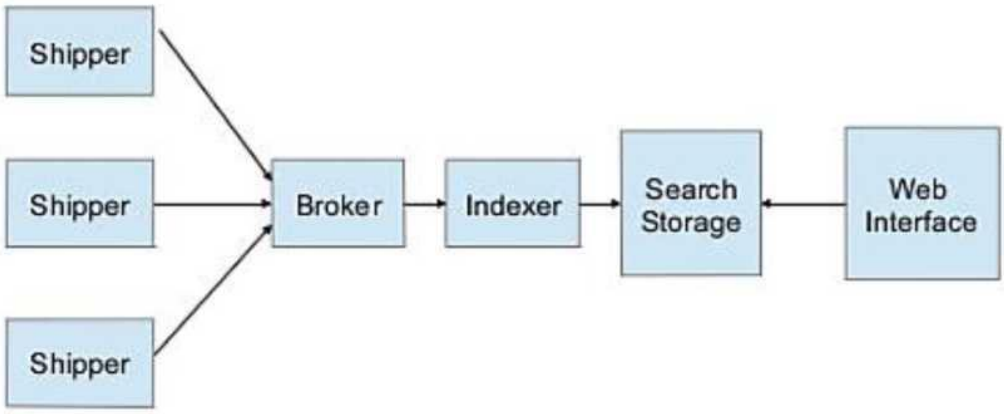
- *Collection of Log Data* là khâu trong đó các bản ghi log thô từ các nguồn sinh log được thu thập và chuyển về trung tâm xử lý.
- *Cleaning of Data* là khâu trong đó các bản ghi log thô được làm sạch để giảm bớt dữ liệu nhiễu.
- *Convert into Structured form* là khâu chuẩn hóa dữ liệu log. Do log có thể được thu thập từ nhiều nguồn với nhiều định dạng khác nhau nên cần thiết

phải được chuẩn hóa và đưa về dạng có cấu trúc, làm đầu vào cho khâu phân tích log.

- *Analysis of Data* là khâu quan trọng nhất trong quá trình phân tích log. Đây là khâu được áp dụng để trích xuất ra các thông tin quan trọng ứng dụng cho đảm bảo an toàn thông tin và các ứng dụng khác.
- *Obtained Results* là khâu kết xuất kết quả ra giao diện người dùng.

Hình 1.13 biểu diễn kiến trúc điển hình của hệ thống thu thập, xử lý và phân tích log. Theo đó, các thành phần chính của hệ thống gồm:

- *Shipper* là mô đun giám sát thu thập log từ các nguồn sinh log khác nhau. Các shipper thường được cài đặt trên các hệ thống được giám sát. Shipper có thể chỉ đơn giản thu thập các bản ghi log thô và gửi về Broker, hoặc nó cũng có thể thực hiện các nhiệm vụ làm sạch và chuẩn hóa dữ liệu log.
- *Broker* là mô đun tiếp nhận dữ liệu log từ nhiều nguồn gửi đến. Sau khi tiếp nhận, dữ liệu log được làm sạch, chuẩn hóa và chuyển tiếp cho khâu tiếp theo.
- *Indexer* là mô đun lập chỉ số cho dữ liệu log. Lập chỉ số là một khâu quan trọng phục vụ tìm kiếm, trích chọn dữ liệu log trong khâu tiếp theo.
- *Search & Strorage* là khâu cung cấp các tính năng tìm kiếm, trích chọn các dữ liệu log quan trọng và quản lý, lưu trữ log.
- *Web Interface* là giao diện người dùng trên nền web cho hệ thống quản lý và phân tích log.



Hình 1.13. Kiến trúc điển hình của hệ thống thu thập, xử lý và phân tích log

1.3. Ứng dụng của phân tích log truy nhập

Việc phân tích log truy cập thường được thực hiện cho các mục đích [5]:

- Đảm bảo an toàn thông tin cho hệ thống;
- Hỗ trợ khắc phục sự cố hệ thống;
- Hỗ trợ điều tra số;
- Hỗ trợ hiểu được hành vi người dùng trực tuyến.

Có thể thấy, phân tích log truy cập phục vụ đảm bảo an toàn thông tin cho hệ thống là một trong các mục đích chính. Cụ thể, 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. Nhiều công cụ đảm bảo an toàn thông tin dựa trên việc 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 [6], VNCS Web Monitoring [7] và hệ thống phát hiện xâm nhập OSSEC [8]. Các công cụ này giám sát, thu thập các dạng log sinh bởi hệ điều hành, các dịch vụ, các ứng dụng trong hệ thống cần giám sát nhằm phát hiện các hành vi bất thường và các dạng tấn công, xâm nhập.

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. Từ đó, kết quả phân tích log có thể được sử dụng để tạo dựng các bằng chứng số cho các sự cố mất an toàn thông tin.

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. Các công cụ phân tích log được phát triển và triển khai trên thực tế cho mục đích này có thể liệt kê bao gồm: Sumo Logic [9], Graylog [10], Webalizer [11] và , ELK Stack [13].

1.4. Một số nền tảng và công cụ xử lý, phân tích log

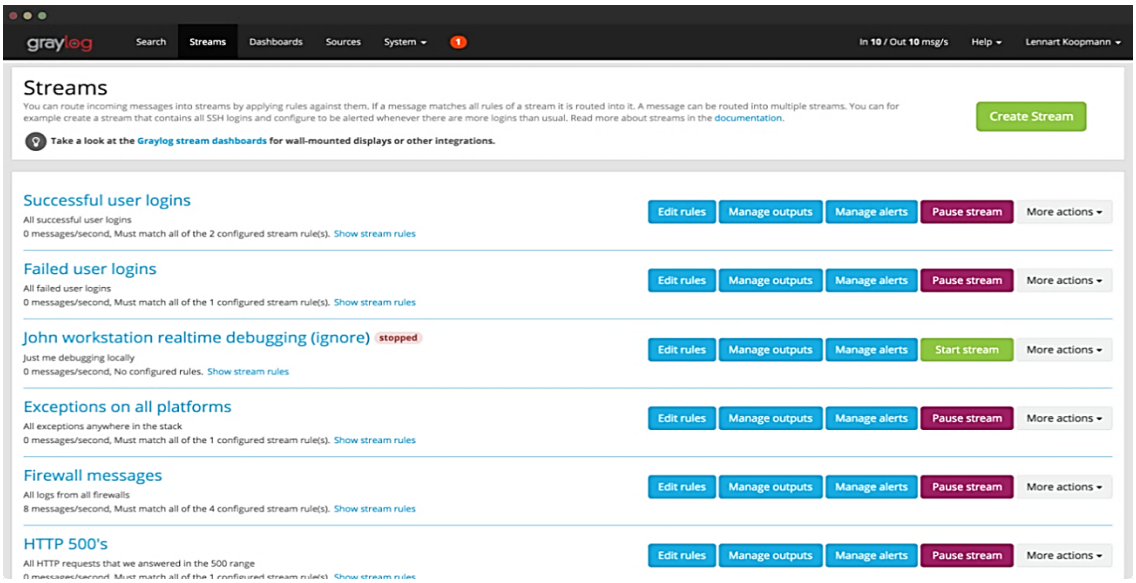
Có nhiều nền tảng và công cụ xử lý, phân tích log truy cập thương mại cũng như mã nguồn mở được cung cấp hiện nay như Splunk [12], Sumo Logic, VNCS Web Monitoring, ELK Stack [13], Graylog, Webzlizer, IBM QRadar SIEM và OSSEC... Mục này giới thiệu khái quát về tính năng và các ưu nhược điểm của một số công cụ phân tích log điển hình, bao gồm Graylog, Webzlizer, và ELK Stack, và một số công cụ thu thập và xử lý log cho đảm bảo ATTT, bao gồm IBM QRadar SIEM và OSSEC.

1.4.1. Các công cụ phân tích log điển hình

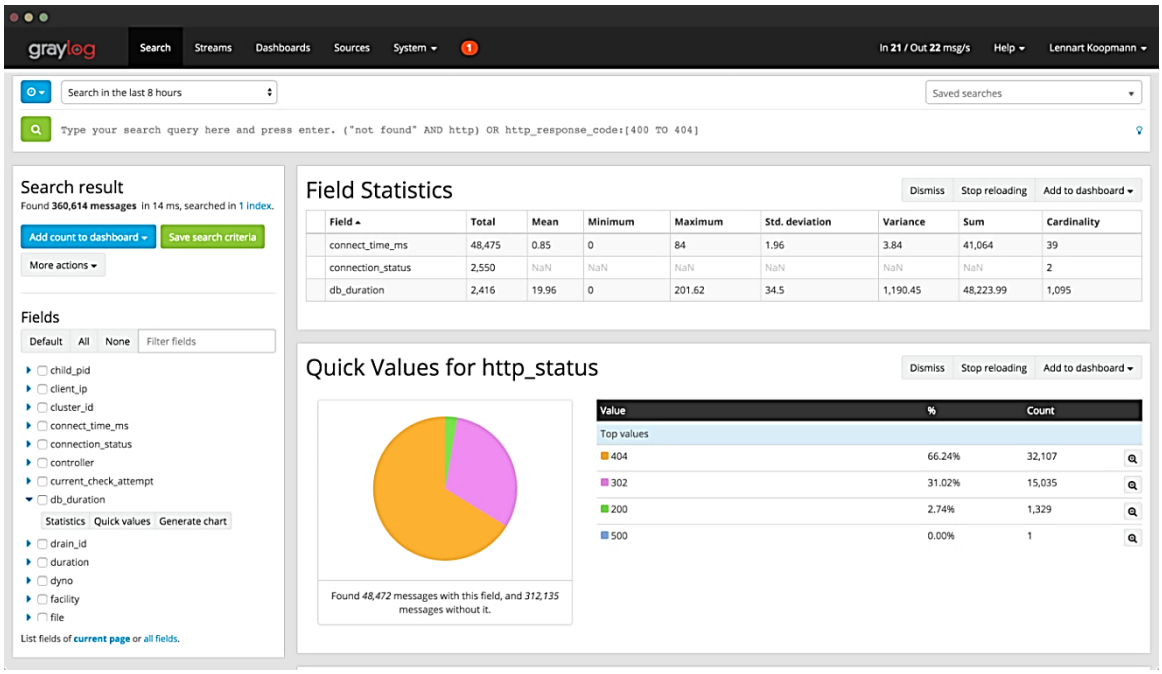
1.4.1.1. Graylog

Graylog [10] là một nền tảng mã mở cho phép xử lý, phân tích log truy cập từ nhiều nguồn theo thời gian thực. Việc thu thập dữ liệu log được thực hiện rất mềm dẻo nhờ khả năng hỗ trợ các công cụ thu thập log của các bên thứ 3, như beats, fluentd và nxlog. Hình 1.14 minh họa màn hình quản lý các nguồn thu thập log của Graylog. Graylog có khả năng phân tích hành vi người dùng, ứng dụng cho phát hiện và cảnh báo các truy cập bất thường cũng như trích xuất các mẫu hành vi truy cập phục vụ cho tối ưu hóa các trang web. Graylog cũng cho phép ánh xạ từ ID sang tên truy nhập của người dùng và ánh xạ từ địa chỉ IP sang vị trí địa lý. Hình 1.15 biểu diễn màn hình báo cáo tổng hợp của Graylog. Mặc dù Graylog có khả năng nhận dạng các hành

vi truy cập bất thường, nhưng nó không cho phép 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.



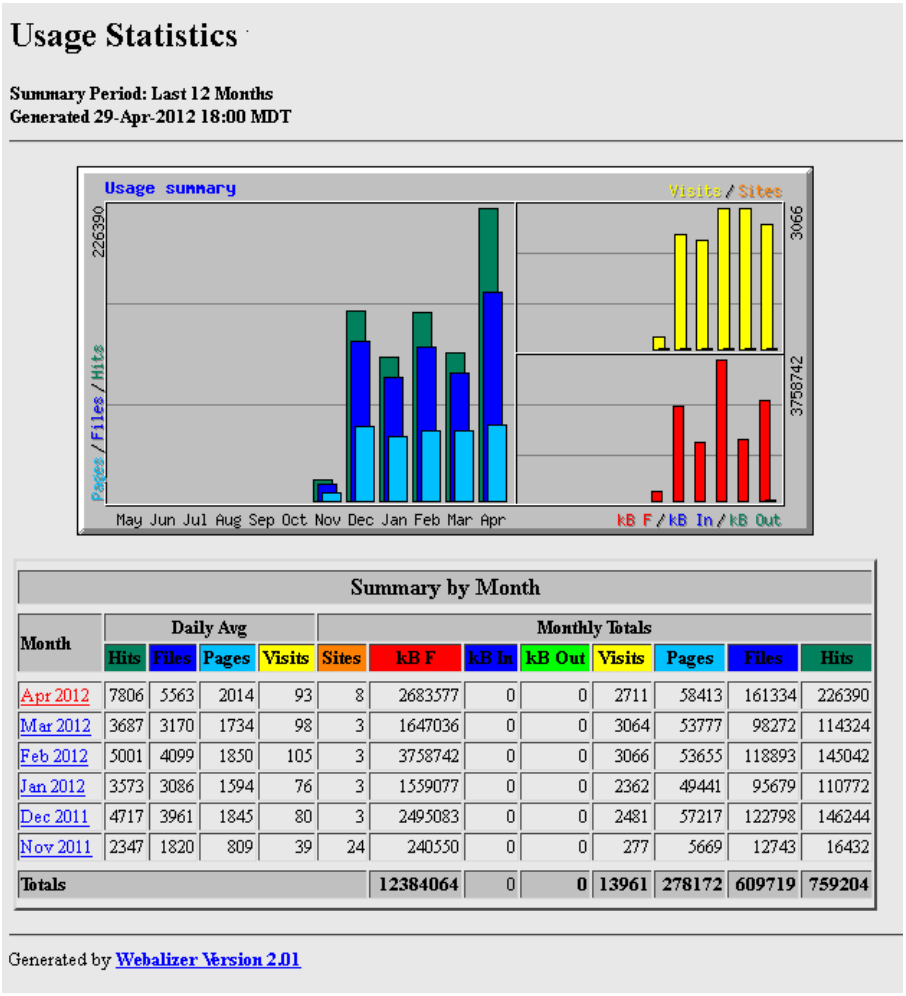
Hình 1.14. Màn hình quản lý các nguồn thu thập log của Graylog [10] https://www.graylog.org



Hình 1.15. Màn hình báo cáo tổng hợp của Graylog [10] https://www.graylog.org

1.4.1.2. Webalizer

Webalizer [11] là công cụ mã mở cho phép xử lý và phân tích log cho các website. Webalizer là một trong các công cụ được sử dụng rộng rãi nhất cho quản trị máy chủ web. Webalizer có khả năng phân tích các dạng log của các trang web và 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. Hình 1.16 minh họa một báo cáo thống kê về truy nhập trang web theo tháng của Webalizer. Ưu điểm của Webalizer là hỗ trợ phân tích nhiều dạng web log và tạo được các báo cáo với các biểu đồ có tính biểu diễn cao. Tuy nhiên, Webalizer chỉ có khả năng phân tích tình hình sử dụng các trang web mà ít có khả năng trích xuất các thông tin cho cảnh báo các nguy cơ mất an toàn thông tin.



Usage Statistics

Summary Period: Last 12 Months
Generated 29-Apr-2012 18:00 MDT

Summary by Month

Month	Daily Avg				Monthly Totals							
	Hits	Files	Pages	Visits	Sites	kB F	kB In	kB Out	Visits	Pages	Files	Hits
Apr 2012	7806	5563	2014	93	8	2683577	0	0	2711	58413	161334	226390
Mar 2012	3687	3170	1734	98	3	1647036	0	0	3064	53777	98272	114324
Feb 2012	5001	4099	1850	105	3	3758742	0	0	3066	53655	118893	145042
Jan 2012	3573	3086	1594	76	3	1559077	0	0	2362	49441	95679	110772
Dec 2011	4717	3961	1845	80	3	2495083	0	0	2481	57217	122798	146244
Nov 2011	2347	1820	809	39	24	240550	0	0	277	5669	12743	16432
Totals						**12384064**	0	**0**	**13961**	**278172**	**609719**	**759204**

Generated by Webalizer Version 2.01

Hình 1.16. Một mẫu báo cáo của Webalizer [11]
http://www.webalizer.org

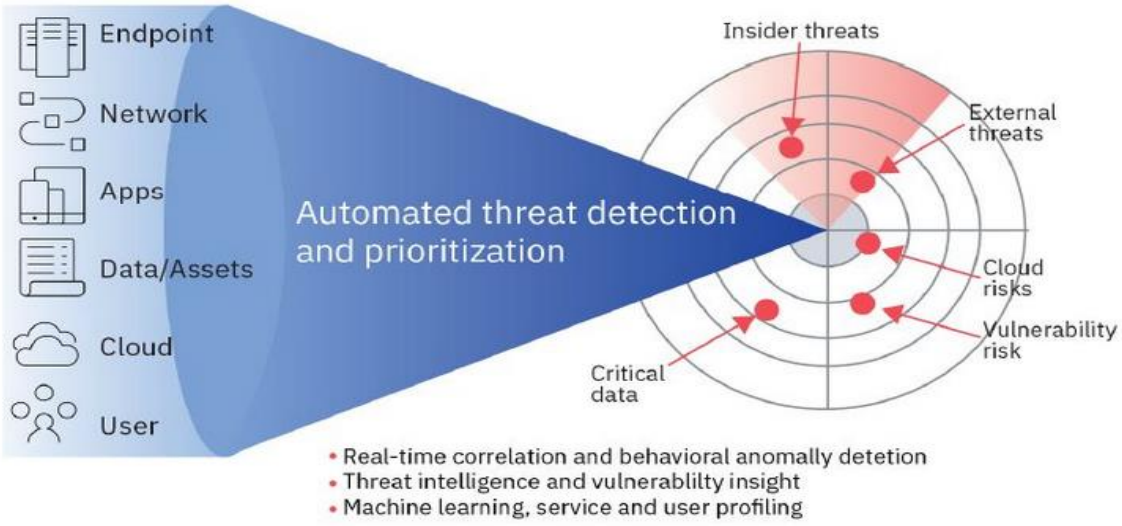
1.4.1.3. ELK Stack

Bộ công cụ xử lý và phân tích log ELK, hay còn gọi là ELK Stack (ElasticSearch, Logstash và Kibana) là bộ công cụ mã mở, miễn phí cho phép thu thập, xử lý và phân tích dữ liệu log thu thập từ nhiều nguồn, đa nền tảng với nhiều định dạng khác nhau. ELK cũng hỗ trợ quản lý, lưu trữ, tìm kiếm dữ liệu log và biểu diễn với nhiều hình thức thể hiện khác nhau. ELK Stack được sử dụng khá rộng rãi để quản lý, xử lý và phân tích dữ liệu log trong các cơ quan, tổ chức, kể các các công ty lớn như Netflix, LinkedIn, Tripware, Medium [17]. Chi tiết về kiến trúc và hoạt động của ELK Stack được trình bày trong chương 2.

1.4.2. Các công cụ thu thập và xử lý log cho đảm bảo ATTT

1.4.2.1. IBM QRadar SIEM

QRadar SIEM (Security Information and Event Management) [6] là hệ thống quản lý các thông tin và sự cố an ninh được phát triển và cung cấp bởi hãng IBM, Hoa Kỳ. QRadar SIEM cho phép phát hiện các bất thường, các nguy cơ an toàn thông tin với độ chính xác cao và tỷ lệ cảnh báo sai thấp thông qua việc xử lý, phân tích dữ liệu log và luồng mạng từ hàng ngàn thiết bị và ứng dụng phân tán trong mạng, như minh họa trên Hình 1.17.



Hình 1.17. Mô hình thu thập và xử lý dữ liệu của QRadar SIEM [6] https://www.ibm.com

Các tính năng tiêu biểu của QRadar SIEM bao gồm:

- Cảm nhận và phát hiện giả mạo, nội gián và các nguy cơ tiên tiến (advanced threats): QRadar SIEM được triển khai như một nền tảng đơn, có khả năng mở rộng cao để giảm hàng nghìn sự kiện bảo mật thành một danh sách có thể quản lý các hành vi nghi ngờ là xâm nhập. Nền tảng có khả năng thu thập các dữ liệu log từ nhiều nguồn, bao gồm các thiết bị mạng, thiết bị bảo mật, hệ điều hành, ứng dụng, cơ sở dữ liệu và các hệ thống quản lý truy cập và nhận dạng. Ngoài ra, QRadar cũng có khả năng thu thập dữ liệu từ lưu lượng mạng, bao gồm dữ liệu Lớp 7 (lớp ứng dụng) từ các bộ chuyển mạch và các bộ định tuyến.
- Thực hiện việc chuẩn hóa và tương quan các sự kiện tức thời: QRadar tối ưu hóa phát hiện các mối đe dọa và báo cáo tuân thủ bằng cách giảm hàng tỷ sự kiện và chuyển thành một số ít các hành vi xâm nhập và xếp hạng chúng theo mức ảnh hưởng đến hoạt động kinh doanh của tổ chức. QRadar xác định ngưỡng hành vi cơ sở và phát hiện bất thường để xác định những thay đổi trong hành vi liên quan đến ứng dụng, máy chủ, người dùng và các phân đoạn mạng. Đồng thời, nó sử dụng công nghệ IBM X-Force Threat Intelligence để xác định hoạt động liên quan đến các địa chỉ IP đáng ngờ, chẳng hạn như những địa chỉ IP bị nghi ngờ lưu trữ phần mềm độc hại.
- Cảm nhận, theo dõi và liên kết các sự cố và nguy cơ: QRadar cho phép đơn giản hóa và tăng cường điều tra bằng cách thực hiện phân tích sự kiện và luồng mạng bằng cách sử dụng luồng dữ liệu gần thời gian thực hoặc dữ liệu lịch sử. Việc bổ sung thêm IBM QRadar QFlow và IBM QRadar VFlow Collector tăng cường khả năng giám sát, phân tích các ứng dụng, cơ sở dữ liệu, các sản phẩm cộng tác và phương tiện truyền thông xã hội thông qua kiểm tra sâu gói tin của lưu lượng mạng lớp 7.
- Có thể bổ sung dung lượng lưu trữ và năng lực xử lý nhanh chóng và rẻ tiền: Cho phép tăng năng lực lưu trữ bằng việc hỗ trợ bổ sung thêm các nút lưu trữ dữ liệu (QRadar Data Node), nhằm cải thiện hiệu năng tìm kiếm khi thu thập

dữ liệu để điều tra hành vi xâm nhập và loại bỏ các nút thắt cổ chai mà không tăng chi phí cho các điều khoản cấp phép.

- Hỗ trợ vấn đề quản lý và phối hợp phòng chống nguy cơ thông qua việc cho phép truy nhập đến Hệ thống trao đổi bảo mật ứng dụng của IBM (IBM Security App Exchange).
- Hỗ trợ việc thực thi chính sách bảo mật dữ liệu: QRadar bao gồm một công cụ báo cáo trực quan mà không yêu cầu các kỹ năng viết báo cáo và cơ sở dữ liệu nâng cao. Cung cấp tính minh bạch, trách nhiệm giải trình và khả năng đo lường để đáp ứng các yêu cầu về quy định và báo cáo tuân thủ.

IBM QRadar SIEM có thể được tích hợp với hàng trăm sản phẩm bảo mật khác của hãng IBM, hoặc của hãng khác. Trên thực tế, QRadar SIEM đã được triển khai sử dụng và được đánh giá cao ở các cơ quan, tổ chức chính phủ, ngân hàng và doanh nghiệp có quy mô 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 không thực sự thích hợp với các cơ quan, tổ chức có hệ thống mạng có quy mô vừa và nhỏ với nguồn lực hạn chế.

1.4.2.2. Hệ thống phát hiện xâm nhập OSSEC

OSSEC (Open Source HIDS SECurity) là một hệ thống phát hiện xâm nhập cho host mã mở, miễn phí được sử dụng rộng rãi [8]. OSSEC cho phép thu thập và phân tích các dạng log, kiểm tra tính toàn vẹn của các file trong hệ thống, giám sát Windows registry, phát hiện rootkit, cảnh báo dựa trên thời gian và phản hồi chủ động. OSSEC cho phép phát hiện xâm nhập cho hầu hết các hệ điều hành, bao gồm Linux, OpenBSD, FreeBSD, OS X, Solaris và Windows. OSSEC có kiến trúc đa nền tảng, tập trung cho phép nhiều hệ thống dễ dàng được giám sát và quản lý. OSSEC có một công cụ phân tích nhật ký có khả năng phân tích tương quan các dữ liệu log thu thập từ nhiều thiết bị với các định dạng khác nhau. OSSEC tuân thủ các yêu cầu Tiêu chuẩn bảo mật dữ liệu công nghiệp thẻ thanh toán (PCI DSS) [8]. Chi tiết về kiến trúc và hoạt động của OSSEC được trình bày trong chương 2.

1.4.3. Nhận xét

Bảng 1.1 cung cấp thông tin so sánh các ưu điểm và nhược điểm của các nền tảng, công cụ xử lý, phân tích log truy nhập đã đề cập ở trên.

Bảng 1.1. So sánh các công cụ xử lý log truy cập

Công cụ	**Ưu điểm**	**Nhược điểm**
Graylog	- Mã mở, miễn phí - Hỗ trợ phân tích log truy cập từ nhiều nguồn và phân tích hành vi người dùng, dùng cho phát hiện và cảnh báo các truy cập bất thường cũng như trích xuất các mẫu hành vi truy cập phục vụ cho tối ưu hóa các trang web	- 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.
Webalizer	- Mã mở, miễn phí - Có khả năng phân tích nhiều dạng web log - Các báo cáo dưới dạng biểu đồ có tính biểu biễn cao.	- Chỉ có khả năng phân tích tình hình sử dụng các trang web - Ít có khả năng trích xuất các thông tin cho cảnh báo các nguy cơ mất an toàn thông tin.
ELK Stack	- Mã mở, miễn phí - Chi phí cài đặt, vận hành thấp - Hỗ trợ trích xuất các mẫu hành vi truy cập phục vụ cho tối ưu hóa các trang web - Giao diện hiển thị đa dạng, phong phú	- 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.

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 thu thập dữ liệu từ hàng ngàn thiết bị 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.
OSSEC	- Mã mở, miễn phí - Hỗ trợ thu thập và xử lý nhiều loại log khác nhau - Hỗ trợ phát hiện các bất thường, các nguy cơ ATTT - Hỗ trợ giám sát tính toàn vẹn của các file và tham số hệ thống	- Giao diện hiển thị và cảnh báo hạn chế - Khó quản trị, giám sát - Việc kết nối giám sát nhiều phân đoạn mạng gặp khó khăn.

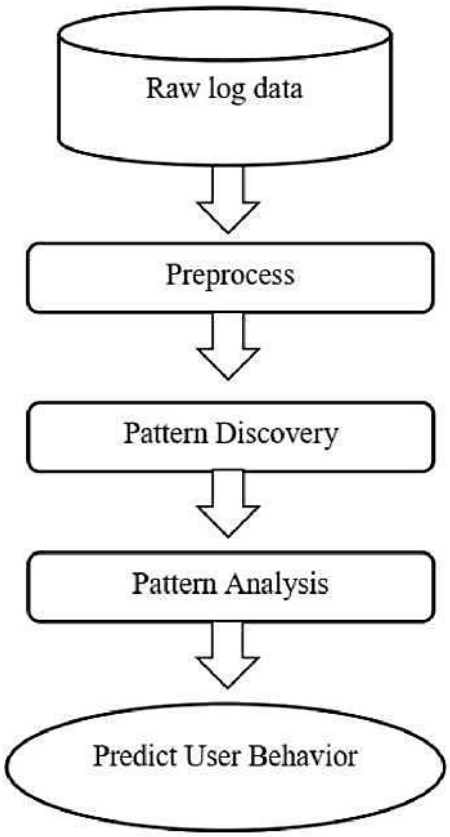
1.5. Kết luận chương

Chương này đã trình bày khái quát về log truy nhập, các nguồn sinh log, vấn đề thu thập, xử lý và phân tích log. Chương cũng giới thiệu chi tiết các dạng log truy nhập phổ biến, các khâu xử lý, phân tích log cũng như ứng dụng của phân tích log. Đồng thời, chương cũng khảo sát một số nền tảng và công cụ xử lý, phân tích log phổ biến hiện nay và rút ra nhận xét.

CHƯƠNG 2. CÁC KỸ THUẬT VÀ MÔ HÌNH XỬ LÝ, PHÂN TÍCH LOG TRUY NHẬP

2.1. Mô hình xử lý log

Hình 2.1 mô tả mô hình xử lý log truy nhập khái quát, mô hình gồm các pha chính: Pha tiền xử lý và chuẩn hóa - Preprocess; Pha nhận dạng mẫu - Pattern Discovery; Pha phân tích mẫu - Pattern Analysis; Pha dự đoán hành vi người dùng - Predict User Behavior.



Hình 2.1. Mô hình xử lý log truy nhập khái quát

- Tiền xử lý và chuẩn hóa - Preprocess:

Trong pha này, hệ thống nhận dữ liệu log từ các nguồn khác nhau, trích xuất các thông tin cần thiết và đưa về một định dạng thống nhất. Ngoài ra, pha này còn chịu trách nhiệm tiền xử lý một số thông tin như: người dùng, phiên làm việc... Pha này gồm các bước sau: Làm sạch và hợp nhất dữ liệu, nhận dạng người dùng, nhận

dạng phiên làm việc. Trong xử lý web log, còn bổ sung thêm bước nhận dạng pageview, hoàn tất đường dẫn (path completion).

- Nhận dạng mẫu - Pattern Discovery:

Pha này sử dụng các phương pháp và thuật toán như: thống kê, học máy, khai phá dữ liệu, nhận dạng mẫu để xác định các pattern của người dùng. Trong phân tích web log, các pattem cơ bản cần xác định bao gồm: Các trang web ưa thích, thời gian xem trung bình mỗi trang web, các lĩnh vực quan tâm... Pha này có thể sử dụng các kỹ thuật phân tích dữ liệu như: phân tích thống kê, phân cụm, phân lớp, luật kết hợp, các mẫu tuần tự, hay mô hình hóa phụ thuộc.

- Phân tích mẫu - Pattern Analysis:

Pha này có nhiệm vụ phân tích các pattern đã tìm được ở pha trước, chỉ ra các pattern không có nhiều giá trị và loại bỏ chúng khỏi quá trình phân tích log. Pha này được thực hiện nhờ các câu truy vấn SQL, hoặc sử dụng phân tích xử lý trực tuyến hay cũng có thể nhờ các kỹ thuật hiển thị hóa dữ liệu để lọc và phân tích pattern.

- Dự đoán hành vi người dùng - Predict User Behavior:

Sau khi đã phân tích và lọc các pattern, những pattern còn lại sẽ được dùng để đưa ra các kết luận về hành vi người dùng. Với phân tích web log, các hành vi người dùng điển hình gồm: Các trang web thường xuyên truy cập, các lĩnh vực quan tâm, thời gian trung bình xem mỗi trang web.

Việc thu thập dữ liệu ở đây chính là việc lấy các thông tin liên quan đến tình trạng hoạt động của các thiết bị trong hệ thống mạng. Tuy nhiên, trong những hệ thống mạng lớn thì các dịch vụ hay các thiết bị không đặt tại trên máy, một địa điểm mà nằm trên các máy chủ, các hệ thống con riêng biệt nhau. Các thành phần hệ thống cũng hoạt động trên những nền tảng hoàn toàn khác nhau. Mô hình Log tập trung được đưa ra để giải quyết vấn đề này. Cụ thể, là tất cả Log sẽ được chuyển về một trung tâm để phân tích và xử lý.

2.2. Thu thập và tiền xử lý

2.2.1. Thu thập log

Log truy nhập có thể được sinh ra ở nhiều vị trí khác nhau trong mạng, do đó có nhiều cách để thu thập log. Log có thể được nhận từ nhiều nguồn khác nhau như: từ file, từ mạng internet hay từ đầu ra của các ứng dụng khác. Một số nguồn thu thập log cụ thể có thể kể ra như:

- 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 giao thức UDP.
- Nhận các sự kiện từ framework Elastic Beats.
- Đọc các kết quả truy vấn từ một cụm Elasticsearch.

Luận văn này sử dụng phương pháp lấy các sự kiện để xử lý từ file log (giám sát Windows logs, đầu ra từ OSSEC), đọc các bản tin syslog (từ hệ điều hành Ubuntu) và thu thập log giám sát tính toàn vẹn hệ thống file sử dụng các OSSEC Agent.

2.2.2. Tiền xử lý và chuẩn hóa log

Quá trình tiền xử lý và chuẩn hóa thực hiện việc làm sạch, hợp nhất dữ liệu từ nhiều nguồn khác nhau và chuẩn hóa dữ liệu theo một định dạng thống nhất. Quá trình này cung cấp các dữ liệu tối ưu và thống nhất cho quá trình phân tích log.

2.2.2.1. Làm sạch và hợp nhất dữ liệu

Làm sạch dữ liệu là quá trình loại bỏ các dữ liệu thừa trong dữ liệu log thô thu thập được mà không mất mát thông tin, đồng thời giúp làm giảm lượng dữ liệu log. Chẳng hạn với web log, làm sạch dữ liệu nhằm xóa bỏ các tham chiếu không liên quan hoặc không quan trọng cho mục đích phân tích log như: các file CSS của trang web, các file icon, âm thanh của trang web. Quá trình này còn xóa bỏ các trường dữ liệu của file log không cung cấp nhiều thông tin quan trong cho quá trình phân tích log như: phiên bản giao thức HTTP. Ngoài ra việc làm sạch dữ liệu còn xóa bỏ các tham chiếu là kết quả do các crawler hoặc các công cụ tìm kiếm thực hiện. Có thể

duy trì một danh sách các crawler của các công cụ tìm kiếm phổ biến để có thể phát hiện và xóa bỏ kết quả log của chúng.

Trong khi đó, hợp nhất dữ liệu cho phép tổng hợp dữ liệu từ các file log có dạng khác nhau. Với web log, ở những trang web lớn, các nội dung log được lưu ở nhiều nguồn khác nhau. Trong trường hợp các nguồn dữ liệu này không có cơ chế dùng chung định danh phiên để hợp nhất dữ liệu thì có thể dùng các phương pháp dựa trên kinh nghiệm như dựa trên trường "referrer" trong server log, kết hợp với các phương pháp nhận dạng người dùng và nhận dạng phiên làm việc để có thể thực hiện hợp nhất dữ liệu.

2.2.2.2. Chuẩn hóa log

Chuẩn hóa log là khâu chuyển dữ liệu log sau làm sạch và hợp nhất về dạng chuẩn nhằm thuận tiện cho quá trình xử lý, phân tích chuyên sâu. Nhìn chung, không có một dạng log chuẩn cho mọi dạng log truy nhập, nên người ta thường sử dụng một dạng log chuẩn cho một nhóm nguồn log có định dạng gần tương tự nhau. Chẳng hạn, có thể sử dụng W3C Extended log file format làm định dạng web log chuẩn để chuyển đổi tất cả các dạng web log khác, và ISC Bind log làm định dạng DNS log chuẩn cho tất cả các dạng DNS log khác.

2.3. Các kỹ thuật phân tích log

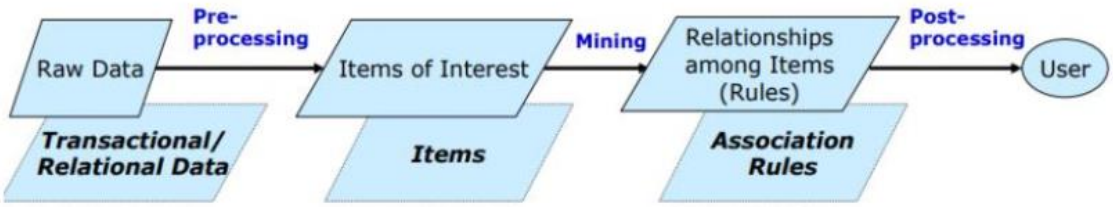
2.3.1. Các kỹ thuật nhận dạng và phân tích mẫu

2.3.1.1. Phân tích thống kê

Thống kê là một kỹ thuật phổ biến nhất trong phân tích log. Bằng cách phân tích các file phiên làm việc của người dùng trong web log, ta có thể thực hiện các phương pháp thống kê khác nhau như: lấy trung bình, tần suất... với các biến khác nhau như: các trang đã xem, số lượt xem, thời gian xem mỗi trang web. Nhiều công cụ phân tích hiện nay cho kết quả là các báo cáo định kỳ về các thống kê của trang web như: các trang web được truy cập nhiều nhất, thời gian trung bình xem một trang web, số lượt truy cập trung bình một trang web. Phương pháp phân tích thống kê có khả năng cung cấp nhiều thông tin hữu ích cho cải thiện hiệu năng của hệ thống hay cho việc marketing.

2.3.1.2. Luật kết hợp

Phương pháp này nhằm phát hiện ra các luật kết hợp giữa các thành phần dữ liệu trong CSDL. Mẫu đầu ra của giải thuật khai phá dữ liệu là tập luật kết hợp tìm được.



Hình 2.2. Quá trình sử dụng luật kết hợp

Hình 2.2 mô tả cách ta có thể sử dụng luật kết hợp, có thể lấy một ví dụ đơn giản về luật kết hợp như sau: Phân tích CSDL bán hàng nhận được thông tin về những khách hàng mua card màn hình cũng có khuynh hướng mua quạt tản nhiệt trong cùng lần mua được miêu tả trong luật kết hợp sau:

"Mua card màn hình ^ Mua quạt tản nhiệt"

[Độ hỗ trợ: 4%, độ tin cậy: 70%]

Độ hỗ trợ và độ tin cậy là hai độ đo của sự đáng quan tâm của luật. Chúng tương ứng phản ánh sự hữu ích và sự chắc chắn của luật đã khám phá. Độ hỗ trợ 4% có nghĩa là 4% của tất cả các tác vụ đã phân tích chỉ ra rằng card màn hình và quạt tản nhiệt là đã được mua cùng nhau. Còn độ tin cậy 70% có nghĩa là 70% các khách hàng mua card màn hình cũng mua quạt tản nhiệt.

2.3.1.3. Phân lớp

Bài toán phân lớp là quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp đã cho trước nhờ một mô hình phân lớp (model). Mô hình này được xây dựng dựa trên một tập dữ liệu được xây dựng trước đó có gán nhãn (hay còn gọi là tập huấn luyện). Quá trình phân lớp là *quá trình gán nhãn* cho đối tượng dữ liệu.

Nhiệm vụ của bài toán phân lớp là cần tìm một mô hình phân lớp để khi có dữ liệu mới thì có thể xác định được dữ liệu đó thuộc vào phân lớp nào. Có nhiều bài

toán phân lớp dữ liệu như phân lớp nhị phân (binary), phân lớp đa lớp (multiclass), phân lớp đa trị.

Trong phân tích log truy nhập, phân lớp thường dùng để ánh xạ một người dùng vào một lớp hay một loại cụ thể. Việc phân lớp trong phân tích web log có thể được thực hiện nhờ 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, việc phân lớp server log có thể giúp phân loại được 46% người dùng đặt hàng các sản phẩm ở trang 'Dell Laptop' có độ tuổi từ 18-23 và sống ở miền Bắc là chủ yếu.

2.3.1.4. Phân cụm

Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp Unsupervised Learning trong Machine Learning. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các quy trình tìm các nhóm đối tượng đã cho vào các cụm - clusters, sao cho các đối tượng trong cùng một cụm tương tự nhau và các đối tượng khác cụm thì không tương tự nhau.

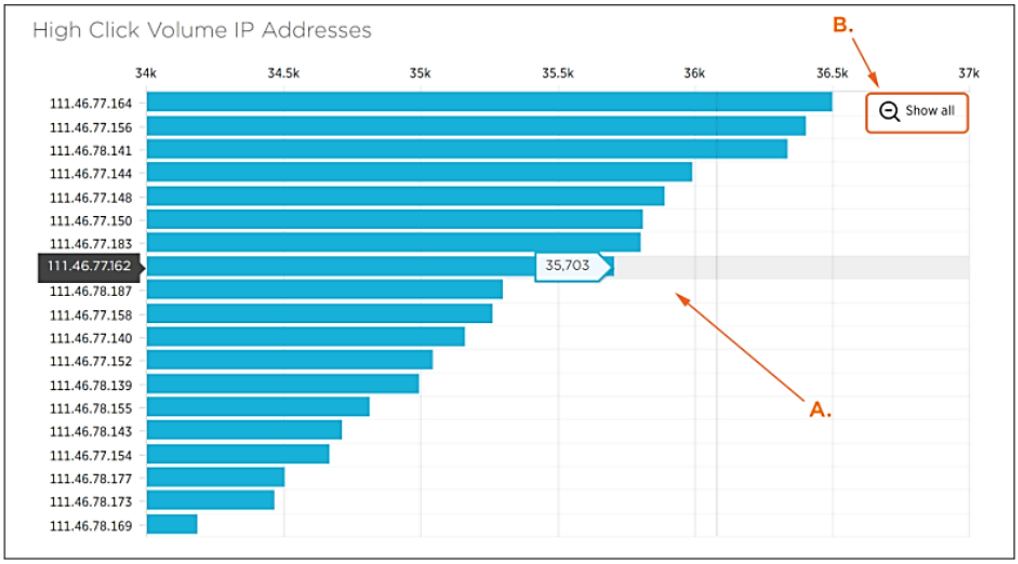
Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Các thuật toán phân cụm đều sinh ra các cụm. Tuy nhiên, không có tiêu chí nào được xem là tốt nhất để đánh giá hiệu quả của phân tích phân cụm, điều này phụ thuộc vào mục đích của phân cụm như: data redution, "natural clustes", "useful" clusters, outlier detection.

Trong phân tích web log, có hai kiểu phân cụm có thể được thực hiện: usage cluster và page cluster. Việc phân cụm những người dùng có pattern giống nhau có nhiều thông tin giá trị cho marketing và thương mại điện tử. Ví dụ, với những nhóm người nhất định thì có thể đưa ra những gợi ý mua hàng phù hợp với sở thích của nhóm người dùng đó mà thôi. Mặt khác, phân cụm các trang web giúp nhận biết được các nhóm trang web có nội dung liên quan đến nhau. Thông tin này đặc biệt hữu ích cho các công cụ tìm kiếm, nhờ những thông tin này chúng có thể đưa ra các trang gợi ý phù hợp với truy vấn của người dùng.

2.3.1.5. Phân tích mẫu

Đây là bước cuối cùng của quá trình phân tích log truy nhập. Quá trình này nhằm lọc ra những luật hay những pattern không có nhiều giá trị đã được tạo ra ở

bước Pattern Discovery. Có nhiều phương pháp để thực hiện việc này, một trong các phương pháp phổ biến và được sử dụng nhiều nhất là nhờ các câu truy vấn SQL hoặc cũng có thể sử dụng phân tích xử lý trực tuyến - OLAP.



Hình 2.3. Phân tích mẫu sử dụng data visualization

Ngoài ra, ở bước này ta cũng áp dụng các kỹ thuật hiển thị hóa dữ liệu - data visualization như các sơ đồ, biểu đồ thống kê để phục vụ phân tích các pattern. Hình 2.3 mô tả một ví dụ sử dụng hiển thị hóa dữ liệu. Ta thấy rằng biểu diễn dữ liệu bằng biểu đồ, đồ thị thống kê giúp dễ dàng nhận ra được sự tương quan dữ liệu cũng như nhận ra xu hướng phát triển của dữ liệu.

2.3.2. Phân tích tương quan

Phân tích tương quan (Correlation Analysis) là một phương pháp thống kê được sử dụng để đánh giá tính chắc chắn của mối quan hệ giữa hai biến định lượng [14]. Một mối tương quan cao có nghĩa là hai hoặc nhiều biến có mối quan hệ chặt chẽ với nhau, trong khi tương quan yếu có nghĩa là các biến đó hầu như không liên quan. Nói cách khác, đó là quá trình nghiên cứu tính chắc chắn của mối quan hệ đó với dữ liệu thống kê có sẵn. Kỹ thuật này có mối liên kết chặt chẽ với kỹ thuật phân tích hồi quy tuyến tính, là một phương pháp thống kê để mô hình hóa mối liên hệ giữa một biến phụ thuộc.

Trong xử lý và phân tích log, kỹ thuật phân tích tương quan có thể được sử dụng để kết nối 2 hoặc nhiều sự kiện log nhằm cho ra một kết luận đúng đắn nhất. Chẳng hạn, khi có 2 sự kiện sau xảy ra trong 1 hệ thống:

- Một ứng dụng định kỳ thực hiện nhiều truy vấn hệ thống DNS, trong đó đa số là truy vấn đến tên miền không có thực (Failed DNS Query);
- Ứng dụng sau đó cố gắng mở kết nối đến một máy chủ bên ngoài mạng.

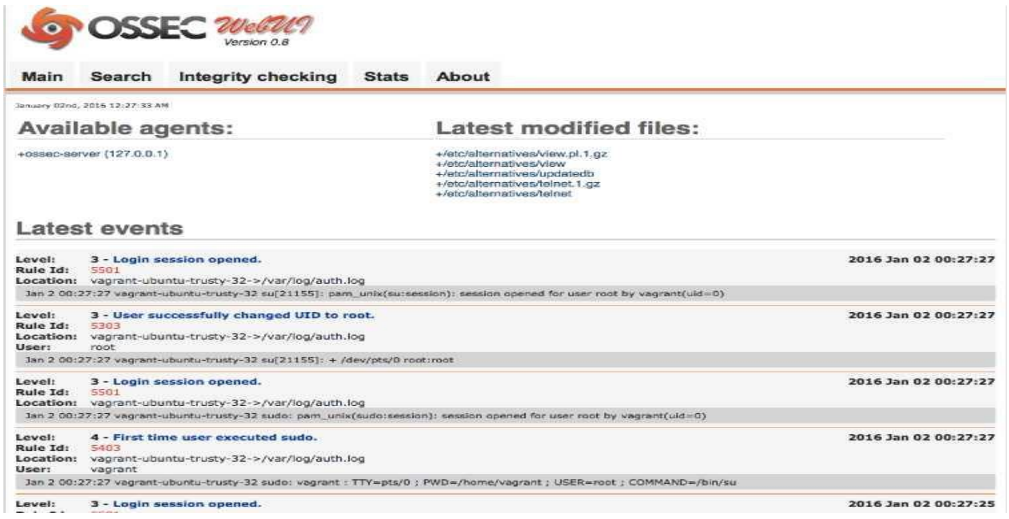
Với kỹ thuật tương quan, có thể đi đến kết luận, hệ thống đó đã bị nhiễm 1 bot và bot đang trong quá trình sinh tên miền tự động, tìm địa chỉ IP và khi tìm được IP, nó cố gắng kết nối đến máy chủ cấp lệnh và điều khiển của một botnet.

2.4. Xây dựng mô hình phân tích log dựa trên OSSEC kết hợp ELK Stack cho phát hiện bất thường và các nguy cơ ATTT

2.4.1. Hệ thống phát hiện xâm nhập OSSEC

2.4.1.1. Giới thiệu

OSSEC là hệ thống phát hiện xâm nhập dựa trên host (HIDS - Host-based Intrusion Detection) dựa trên log mã nguồn mở, miễn phí, đa nền tảng có thể mở rộng và có nhiều cơ chế bảo mật khác nhau. OSSEC có thể phát hiện xâm nhập bằng cả chữ ký hoặc dấu hiệu bất thường. OSSEC cung cấp kiến trúc đa nền tảng tập trung, cho phép quản lý bảo mật máy tính từ một vị trí trung tâm.



Hình 2.4. Giao diện người dùng của OSSEC
https://www.ossec.net/

Các dấu hiệu bình thường và bất thường được mô tả trong bộ luật của OSSEC. OSSEC có một công cụ phân tích và tương quan mạnh mẽ, tích hợp giám sát và phân tích log, kiểm tra tính toàn vẹn của file, kiểm tra registry của Windows, thực thi chính sách tập trung, giám sát chính sách, phát hiện rootkit, cảnh báo thời gian thực và phản ứng một cách chủ động cuộc tấn công đang diễn ra. Các hành động này cũng có thể được định nghĩa trước bằng luật trong OSSEC để OSSEC hoạt động theo ý muốn của người quản trị. Ngoài việc được triển khai như một HIDS, nó thường được sử dụng như một công cụ phân tích log, cho phép theo dõi và phân tích các bản ghi log của IDS, các máy chủ Web và các bản ghi xác thực. OSSEC có thể sử dụng để giám sát các máy chạy hầu hết các hệ điều hành, bao gồm Linux, OpenBSD, FreeBSD, Mac OS X, Sun Solaris và Microsoft Windows. OSSEC còn có thể được tích hợp trong trong các hệ thống bảo mật lớn hơn là SIEM (Security information and event management). Tuy nhiên, OSSEC server chỉ có thể cài đặt trên các hệ điều hành họ Linux/Unix và OSSEC chỉ có thể cài đặt trên Windows với tư cách là một agent.

2.4.1.2. Các tính năng nổi bật của OSSEC

Các tính năng nổi bật của OSSEC bao gồm:

- *Theo dõi và phân tích các log:* OSSEC thu thập log theo thời gian thực từ nhiều nguồn khác nhau để phân tích (giải mã, lọc và phân loại) và đưa ra cảnh báo dựa trên bộ luật được xây dựng trước. OSSEC phát hiện các cuộc tấn công trên mạng, hệ thống hoặc ứng dụng cụ thể bằng cách sử dụng log làm nguồn thông tin chính. Log cũng rất hữu ích để phát hiện việc khai thác lổ hổng phần mềm, vi phạm chính sách và các hình thức hoạt động không phù hợp khác. Một số loại log mà OSSEC có thể phân tích là log proxy, log web, log ghi lại xác thực, system log.

- *Kiểm tra tính toàn vẹn của file:* Sử dụng hàm băm mật mã, có thể tính toán giá trị băm của mỗi file trong hệ điều hành dựa trên tên file, nội dung file và giá trị băm này là duy nhất. OSSEC có thể giám sát các ổ đĩa để phát hiện các thay đổi của giá trị băm này khi có ai đó, hoặc điều gì đó, sửa đổi nội dung của file hoặc thay thế phiên bản file này bằng một phiên bản file khác.

- *Giám sát Registry:* Hệ thống Registry là danh sách thư mục tất cả các cài đặt phần cứng và phần mềm, các cấu hình hệ điều hành, người dùng, nhóm người dùng, và các preference trên một hệ thống Microsoft Windows. Các thay đổi được thực hiện bởi người dùng và quản trị viên đối với hệ thống được ghi lại trong các khóa registry để các thay đổi được lưu khi người dùng đăng xuất hoặc hệ thống được khởi động lại. Registry cũng cho thấy kernel của hệ điều hành tương tác với phần cứng và phần mềm máy tính như thế nào. HIDS có thể giám sát những thay đổi này đối với các khóa registry quan trọng để đảm bảo rằng người dùng hoặc ứng dụng không cài đặt một chương trình mới hoặc sửa đổi chương trình hiện có với mục đích xấu.

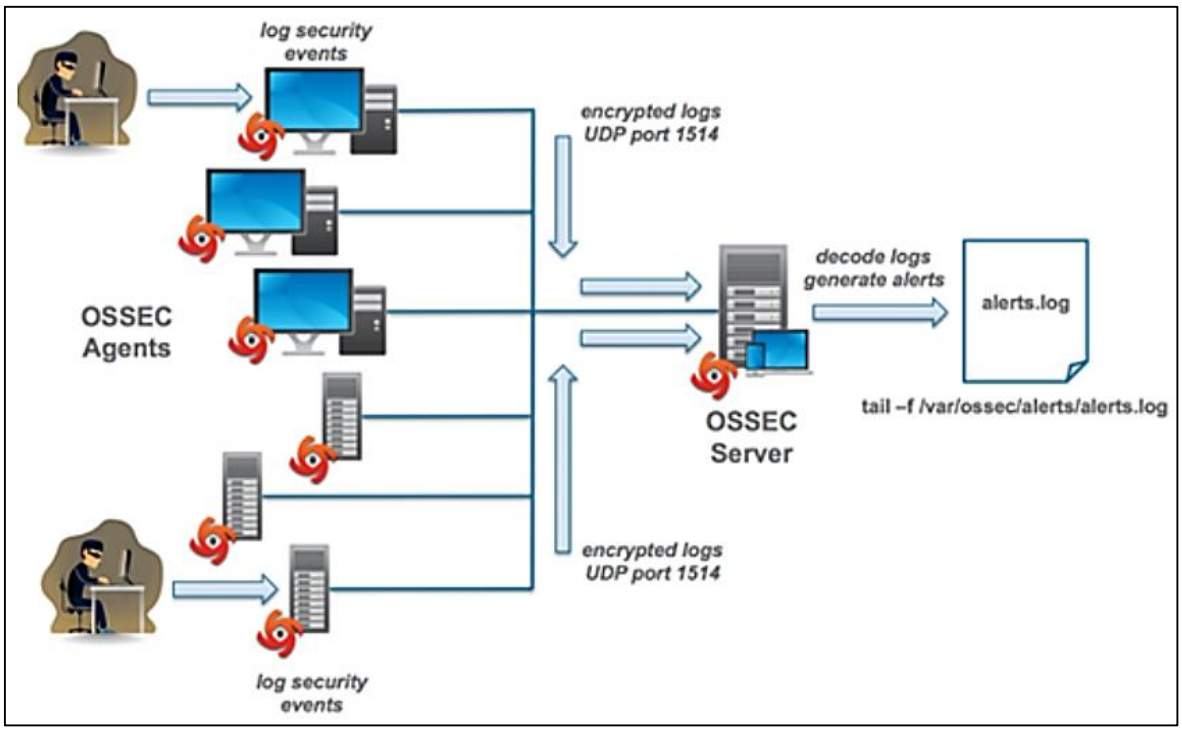
- *Phát hiện Rootkit:* OSSEC phát hiên Rootkit dựa trên chữ ký, rootkit là công cụ cho phép kẻ đột nhập khả năng xâm nhập trở máy tính bị cài rootkit và xóa dấu vết về sự tồn tại của nó. Kẻ xâm nhập có thể sử dụng rootkit để ăn cắp thông tin và tài nguyên từ máy tính nạn nhân. OSSEC có khả năng phát hiện rootkit bằng cách đọc file cơ sở dữ liệu về rootkit và tiến hành quét hệ thống định kỳ, thực hiện các lời gọi hệ thống để phát hiện các file không bình thường, các tiến trình ẩn, các dấu hiệu vượt quyền, các cổng ẩn và so sánh chúng với cơ sở dữ liệu để phát hiện rootkit.

- *Phản ứng chủ động:* Phản ứng chủ động cho phép các IDS nói chung và OSSEC nói riêng tự động thực thi các lệnh hoặc phản ứng khi một sự kiện hoặc tập hợp sự kiện cụ thể được kích hoạt. Phản ứng chủ động có thể được xác định bằng luật. Các lợi ích của phản ứng chủ động là rất lớn, nhưng cũng rất nguy hiểm, có thể ngăn chặn kết nối hợp pháp hoặc là lổ hổng để kẻ tấn công khai thác. Ví dụ: quản trị viên hợp pháp có thể tạo ra báo động sai và chặn người dùng/máy chủ hợp pháp truy cập nếu các luật được thiết kế kém.

- *Giám sát toàn vẹn tập tin:* Còn được gọi là syscheck, là một xác nhận hợp lệ định kỳ về tính toàn vẹn của hệ điều hành hoặc các ứng dụng file bằng cách so sánh trạng thái hiện tại và giá trị được lưu trữ đã biết. Nó là một phần rất quan trọng trong việc phát hiện xâm nhập, và nó thường sử dụng các hàm băm để kiểm tra, phát hiện các thay đổi. OSSEC sử dụng mã MD5/SHA1 để giám sát các file cấu hình quan trọng trong một hệ thống.

2.4.1.3. Kiến trúc và hoạt động của OSSEC

OSSEC được thiết kế theo mô hình client – server, gồm 2 thành phần chính là OSSEC server và OSSEC agents. Hình 2.5 biểu diễn luồng hoạt động của hệ thống phát hiện xâm nhập OSSEC [8][15]. Theo đó, OSSEC gồm 2 thành phần chính: (1) OSSEC Server và (2) các OSSEC agents. Các OSSEC agents có nhiệm vụ giám sát, thu thập dữ liệu tại các hệ thống cần bảo vệ và chuyển về cho OSSEC Server xử lý, phân tích. Các OSSEC agents cũng có khả năng thực thi các lệnh/phản hồi gửi từ OSSEC Server. Các OSSEC agents hỗ trợ nhiều nền tảng hệ điều hành, cho phép giám sát các hệ thống chạy hệ điều hành Linux, OpenBSD, FreeBSD, OS X, Solaris và Windows.



Hình 2.5. Luồng hoạt động của hệ thống phát hiện xâm nhập OSSEC [8][15] https://www.ossec.net/

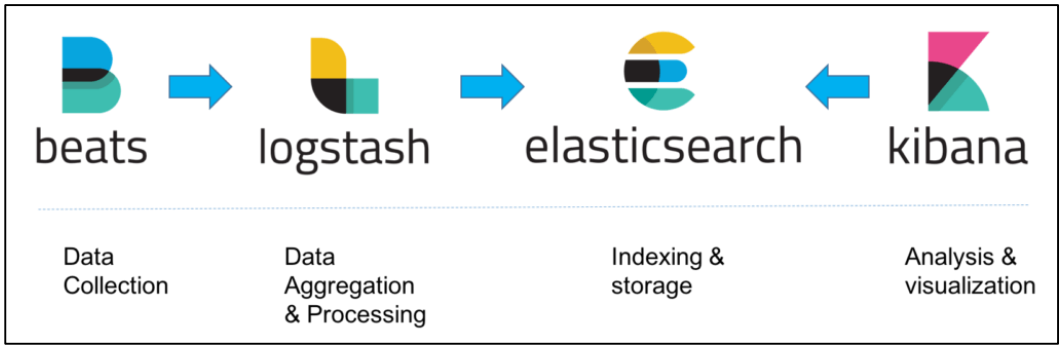
OSSEC Server là trung tâm quản lý hệ thống, có nhiệm vụ tiếp nhận và xử lý dữ liệu thu thập từ nhiều OSSEC agents. OSSEC Server phát hiện các xâm nhập, bất thường và các vi phạm chính sách bảo mật hệ thống dựa trên một tập các luật. Với phiên bản ổn định hiện tại 3.30, OSSEC hỗ trợ sẵn hơn 2800 luật cho phép phát hiện

hầu hết các dạng xâm nhập, bất thường. Đầu ra chuẩn của OSSEC Server là các cảnh báo (alert) được lưu vào file log. Hiện nay OSSEC Server chỉ có thể hoạt động trên các hệ điều hành Linux và Unix. Hệ thống OSSEC cũng hỗ trợ một giao diện web đơn giản cho phép quan sát dữ liệu thu thập và các kết quả phân tích ở dạng thô.

2.4.2. Bộ công cụ xử lý và phân tích log ELK Stack

2.4.2.1. Giới thiệu

Hình 2.6 biểu diễn các thành phần chính của ELK Stack và tương tác giữa chúng. Theo đó, các thành phần ELK Stack gồm:



Hình 2.6. Các thành phần của bộ công cụ xử lý và phân tích log ELK [12] https://www.elastic.co

- Beats gồm các công cụ thu thập dữ liệu log. Beats được cài đặt trên các hệ thống đích, cho phép giám sát thu thập log theo thời gian thực. Beats hỗ trợ nhiều nền tảng và khả năng thu thập log với nhiều định dạng khác nhau.
- Logstash là bộ tập trung và tổng hợp dữ liệu log. Logstash tiếp nhận dữ liệu log từ các beats, thực hiện tiền xử lý/làm sạch và chuẩn hóa dữ liệu log, sau đó chuyển cho ElasticSearch.
- ElasticSearch là một cơ sở dữ liệu kiểu NoSQL cho phép quản lý, lập chỉ số và tìm kiếm dữ liệu log kiểu full-text.
- Kibana là giao diện web của hệ thống, hỗ trợ phân tích và hiển thị dữ liệu log cũng như kết quả phân tích dưới dạng bảng hoặc biểu đồ.

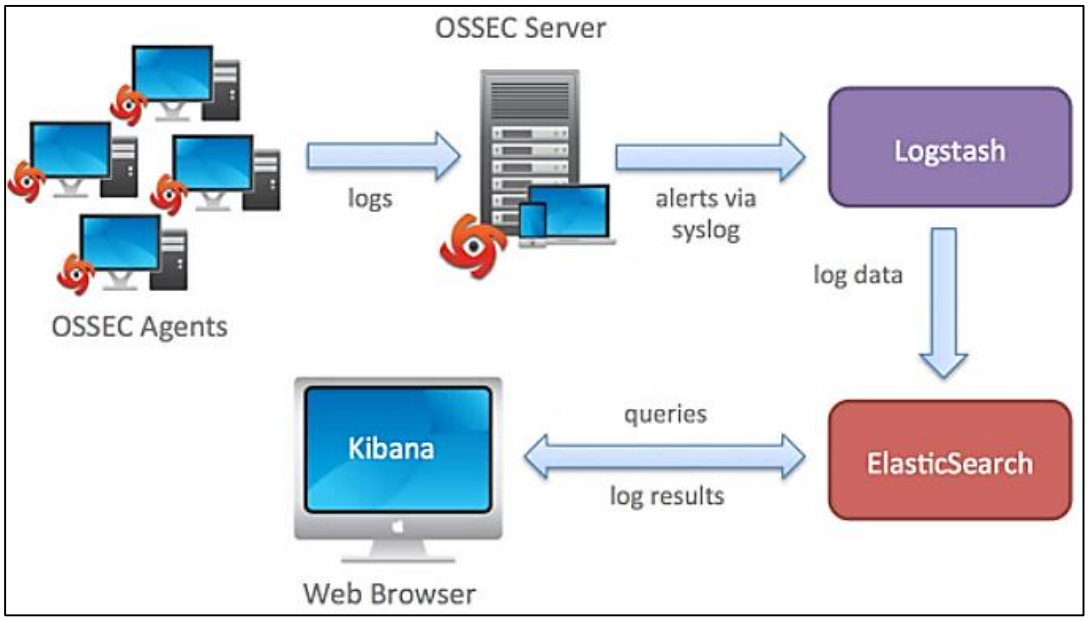
2.4.2.2. Các ưu điểm khi sử dụng ELK Stack

ELK Stack hỗ trợ thu thập, quản lý log với nhiều tính năng phong phú. Có thể liệt kê các ưu điểm khi sử dụng ELK Stack như sau:

- Hỗ trợ thu thập log từ nhiều nguồn: Logstash có thể đọc được log từ rất nhiều nguồn, từ log file cho đến log CSDL cho đến UDP hay REST request.
- Dễ tích hợp: Dù bạn có dùng Nginx hay Apache, dùng MSSQL, MongoDB hay Redis, Logstash đều có thể đọc hiểu và xử lý log của bạn nên việc tích hợp rất dễ dàng.
- Hoàn toàn miễn phí: Chỉ cần tải về, cài đặt và sử dụng, không tốn một đồng nào cả.
- Khả năng mở rộng tốt: Logstash và Elasticsearch có thể chạy trên nhiều nút nên hệ thống ELK có khả năng mở rộng rất tốt. Khi có thêm dịch vụ, thêm người dùng, muốn log nhiều hơn, bạn chỉ việc thêm nút cho Logstash và Elasticsearch là xong.
- Search và filter mạnh mẽ: Elasticsearch cho phép lưu trữ thông tin kiểu NoSQL, hỗ trợ Full-Text Search nên việc query rất dễ dàng và mạnh mẽ.
- Cộng đồng sử dụng lớn, có nhiều tài liệu hướng dẫn và hỗ trợ, nên dễ dàng tiếp cận.

2.4.3. Mô hình triển khai tích hợp OSSEC và ELK Stack

Đầu ra tiêu chuẩn của hệ thống phát hiện xâm nhập OSSEC là các cảnh báo dưới dạng các dòng log lưu vào file, như biểu diễn trên Hình 2.5. Gói phần mềm OSSEC cũng có thành phần hỗ trợ giao diện web, nhưng có tính năng khá hạn chế và không hỗ trợ phân tích chuyên sâu log kết quả [8][15]. Trong khi đó, ELK Stack là bộ công cụ cho phép thu thập, xử lý và phân tích log chuyên sâu với nhiều tính năng mạnh và khả năng hiển thị, trình bày phong phú. Do vậy, việc tích hợp ELK Stack với hệ thống phát hiện xâm nhập OSSEC cho phép khai thác hiệu quả các điểm mạnh của ELK Stack, bổ trợ hiệu quả cho OSSEC. Điều này giúp tạo thành một hệ thống xử lý và phân tích log cho phát hiện bất thường và nguy cơ an toàn thông tin với khả năng quản lý log với khối lượng lớn và các tính năng phân tích log chuyên sâu và khả năng hiển thị log cũng như kết quả xử lý đa dạng dưới nhiều hình thức khác nhau.



Hình 2.7. Mô hình tích hợp OSSEC và ELK [15]

Hình 2.7 biểu diễn mô hình tích hợp OSSEC và ELK [15]. Theo đó, dữ liệu log và các cảnh báo (alert) xuất ra từ OSSEC được xử lý tiếp như sau:

- Dữ liệu log và các cảnh báo (gọi chung là log) được thu thập và xử lý bởi thành phần Logstash. Tại đây, log được làm sạch, chuẩn hóa và chuyển sang khâu tiếp theo.
- Dữ liệu log sau chuẩn hóa được chuyển đến ElasticSearch quản lý và lập chỉ số phục vụ phân tích, tìm kiếm.
- Kibana là thành phần cuối cùng trong hệ thống cho phép phân tích log chuyên sâu và biểu diễn log và kết quả xử lý dưới nhiều dạng khác nhau (báo cáo, đồ thị, biểu đồ,…).

2.5. Kết luận chương

Chương 2 đã trình bày về các kỹ thuật xử lý và phân tích log, bao gồm mô hình khái quát cho xử lý và phân tích log, vấn đề tiền xử lý, chuẩn hóa log, cũng như các kỹ thuật phân tích log. Phần cuối chương mô tả việc xây dựng mô hình phân tích log dựa trên OSSEC kết hợp ELK Stack cho phát hiện bất thường và các nguy cơ ATTT làm cơ sở cho thử nghiệm tại chương 3.

CHƯƠNG 3. CÀI ĐẶT, THỬ NGHIỆM VÀ ĐÁNH GIÁ

3.1. Môi trường thử nghiệm và mô hình triển khai cài đặt

3.1.1. Môi trường và công cụ thử nghiệm

Môi trường thử nghiệm sử dụng trong luận văn là hệ thống mạng mô phỏng dựa trên phần mềm ảo hóa VMWare Professional 15. Các phần mềm và công cụ thử nghiệm bao gồm:

- Phần mềm ảo hóa VMWare Professional 15: Tạo môi trường ảo hóa để cài đặt các máy chủ, máy cần giám sát trong hệ thống.
- Hệ điều hành Ubuntu 16.04: Cài đặt máy chủ của hệ thống chứa OSSEC server, ELK Stack và một số công cụ cho tích hợp.
- Hệ điều hành Windows 7: Máy tramk chạy Windows 7 là máy được giám sát 01.
- Hệ điều hành Windows Server 2008: Máy chủ chạy Windows là máy được giám sát 02.
- Bộ phần mềm Wazuh [16]: Là gói công cụ mã mở tích hợp OSSEC và Wazuh API, Wazuh application cho phép kết nối với ELK Stack.
- Filebeat là công cụ cho phép giám sát thu thập log.
- Bộ công cụ ELK Stack: Là gói phần mềm xử lý và phân tích log, gồm ElasticSearch, Logstash và Kibana [12].

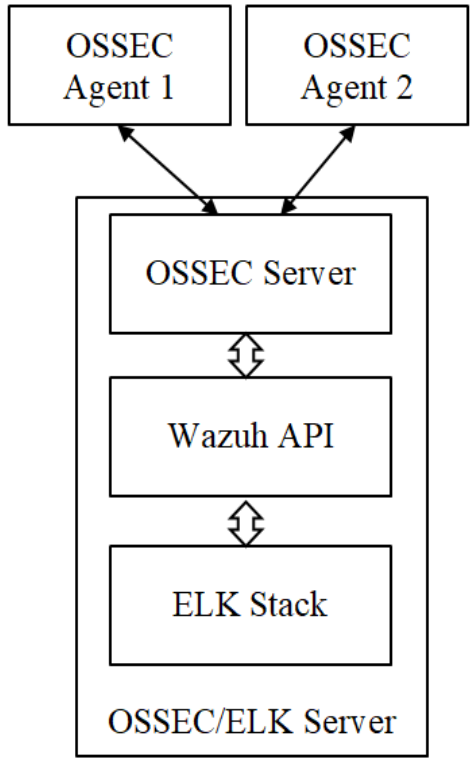
3.1.2. Mô hình cài đặt hệ thống thử nghiệm

Hình 3.1 biểu diễn mô hình cài đặt hệ thống thử nghiệm. Theo đó, hệ thống thử nghiệm được triển khai cài đặt gồm 3 máy như sau:

- Các máy chạy OSSEC agent là các máy được giám sát phát hiện xâm nhập, bất thường. Trên mỗi máy được cài đặt một OSSEC/Wazuh agent để thực hiện giám sát và thu thập dữ liệu log trong hệ thống và chuyển về OSSEC server. Hệ thống thử nghiệm gồm 2 máy được giám sát, một máy chạy Windows 7 và một máy chạy Windows server 2008.

- OSSEC/ELK Server: Là máy chủ chạy hệ điều hành Ubuntu 16.04. Đây là máy chủ chính của hệ thống cho cài đặt OSSEC server, ELK Stack và Wazuh API. Các thành phần được cài đặt trên máy chủ này gồm:

+ OSSEC server thu thập log và các dữ liệu khác từ các OSSEC agents, thực hiện xử lý và phát hiện.
+ Wazuh API là giao diện giao tiếp giữa OSSEC server và ELK Stack.
+ ELK Stack là bộ công cụ xử lý và phân tích log kết xuất từ OSSEC server.



Hình 3.1. Mô hình cài đặt hệ thống thử nghiệm

3.2. Triển khai cài đặt hệ thống thử nghiệm

Do bộ công cụ Wazuh đã tích hợp OSSEC server và các công cụ quản lý vào gói phần mềm Wazuh Manager, nên các thành phần thực tế cần cài đặt trên máy chủ OSSEC/ELK Server bao gồm: Wazuh Manager, Wazuh API, Filebeat và ELK Stack. Filebeat là công cụ cho phép thu thập log trên bản thân máy chủ OSSEC/ELK Server. Thành phần phải cài đặt trên các máy trạm/máy được giám sát là Wazuh agent. Bản chất của Wazuh agent là OSSEC agent được đóng gói trong gói phần mềm Wazuh.

3.2.1. Cài đặt Wazuh Manager, Wazuh API và Filebeat

3.2.1.1. Thêm thông tin gói phần mềm Wazuh vào thư viện quản lý của Ubuntu

- Cài đặt các công cụ hỗ trợ curl, apt-transport-https và lsb-release:

```
# apt-get update
# apt-get install curl apt-transport-https lsb-release gnupg2
```

- Cài đặt khóa GPG:

```
# curl -s https://packages.wazuh.com/key/GPG-KEY-WAZUH | apt-key add -
```

- Thêm gói phần mềm Wazuh vào thư viện:

```
# echo "deb https://packages.wazuh.com/3.x/apt/ stable main" | tee -a /etc/apt/sources.list.d/wazuh.list
```

- Cập nhật gói phần mềm:

```
# apt-get update
```

3.2.1.2. Cài đặt Wazuh Manager

```
# apt-get install wazuh-manager
```

3.2.1.3. Cài đặt Wazuh API

- Cài đặt NodeJS

```
# curl -sL https://deb.nodesource.com/setup_8.x | bash -
# apt-get install nodejs
```

- Cài đặt Wazuh API

```
# apt-get install wazuh-api
```

3.2.1.4. Cài đặt Filebeat

- Cài đặt thông tin Elastic vào thư viện và khóa GPG của nó:

```
# apt-get install curl apt-transport-https
# curl -s https://artifacts.elastic.co/GPG-KEY-elasticsearch | apt-key add -
```

```
# echo "deb
https://artifacts.elastic.co/packages/7.x/apt stable
main" | tee /etc/apt/sources.list.d/elastic-7.x.list
# apt-get update
```

- Cài đặt Filebeat

```
# apt-get install filebeat=7.4.2
```

- Tải file cấu hình của Filebeat từ kho lưu trữ của Wazuh. File cấu hình này cho phép chuyển các cảnh báo từ OSSEC sáng ElasticSearch:

```
# curl -so /etc/filebeat/filebeat.yml
https://raw.githubusercontent.com/wazuh/wazuh/v3.10.
2/extensions/filebeat/7.x/filebeat.yml
```

- Tải mẫu cảnh báo cho ElasticSearch:

```
# curl -so /etc/filebeat/wazuh-template.json
https://raw.githubusercontent.com/wazuh/wazuh/v3.10.
2/extensions/elasticsearch/7.x/wazuh-template.json
```

- Tải mô đun Wazuh cho Filebeat:

```
# curl -s
https://packages.wazuh.com/3.x/filebeat/wazuh-
filebeat-0.1.tar.gz | sudo tar -xvz -C
/usr/share/filebeat/module
```

- Sửa tham số output.elasticsearch.hosts trong file cấu hình của Filebeat /etc/filebeat/filebeat.yml (192.168.186.130 là IP của máy chạy ElasticSearch):

```
output.elasticsearch.hosts:
['http://192.168.186.130:9200']
```

- Cài đặt và chạy Filebeat tự động như một dịch vụ:

```
# systemctl daemon-reload
# systemctl enable filebeat.service
```

```
# systemctl start filebeat.service
```

3.2.2. Cài đặt ELK Stack

Bộ công cụ ELK Stack gồm 3 thành phần là ElasticSearch, Logstash và Kibana. Tuy nhiên, do ELK Stack trong hệ thống thử nghiệm chỉ tiếp nhận dữ liệu log từ 1 hệ thống OSSEC do Filebeat thu thập, nên dữ liệu log từ OSSEC chuyển qua Filebeat sang thẳng ElasticSearch, mà không cần có Logstash. Trong trường hợp ELK Stack tiếp nhận log từ nhiều nguồn, hoặc từ nhiều hệ thống OSSEC, Logstash được sử dụng để tập trung dữ liệu. Do vậy, phần này chỉ mô tả việc cài đặt ElasticSearch và Kibana trong hệ thống thử nghiệm. Do ELK Stack được viết bằng Java, nên cần có máy ảo Java 8 được cài đặt, nếu hệ thống chưa có Java. Giả thiết hệ thống đã có Java 8 được cài đặt sẵn.

3.2.2.1. Nạp thông tin gói ELK và khóa GPG vào thư viện

```
# apt-get install curl apt-transport-https
# curl -s https://artifacts.elastic.co/GPG-KEY-
elasticsearch | apt-key add -
# echo "deb
https://artifacts.elastic.co/packages/7.x/apt stable
main" | tee /etc/apt/sources.list.d/elastic-7.x.list
# apt-get update
```

3.2.2.2. Cài đặt ElasticSearch

- Cài đặt ElasticSearch

```
# apt-get install elasticsearch=7.4.2
```

- Chỉnh sửa tham số địa chỉ IP lắng nghe trong file cấu hình /etc/elasticsearch/elasticsearch.yml của ElasticSearch:

```
network.host: 192.168.186.130
```

- Chỉnh sửa tham số tên nút và cluster khởi tạo trong file cấu hình /etc/elasticsearch/elasticsearch.yml của ElasticSearch:

```
node.name: Ossec_Server
cluster.initial_master_nodes: ["Ossec_Server"]
```

- Cài đặt và chạy ElasticSearch tự động như một dịch vụ:

```
# systemctl daemon-reload
# systemctl enable elasticsearch.service
# systemctl start elasticsearch.service
```

- Khi ElasticSearch đã chạy, nạp mẫu định dạng của Filebeat:

```
# filebeat setup --index-management -E
setup.template.json.enabled=false
```

- Kiểm tra trạng thái hoạt động của ElasticSearch

```
# curl http://192.168.186.130:9200
```

3.2.2.3. Cài đặt Kibana

- Cài đặt gói Kibana

```
# apt-get install kibana=7.4.2
```

- Cài đặt trình mở rộng Wazuh cho Kibana:

```
# sudo -u kibana /usr/share/kibana/bin/kibana-plugin
install
https://packages.wazuh.com/wazuhapp/wazuhapp-
3.10.2_7.4.2.zip
```

- Sửa tham số server.host trong file cấu hình /etc/kibana/kibana.yml của Kibana:

```
server.host: 192.168.186.130
```

- Sửa tham số elasticsearch.hosts trong file cấu hình /etc/kibana/kibana.yml của Kibana:

```
elasticsearch.hosts: ["http://192.168.186.130:9200"]
```

- Cài đặt và chạy Kibana tự động như một dịch vụ:

```
# systemctl daemon-reload
# systemctl enable kibana.service
# systemctl start kibana.service
```

- Kiểm tra trạng thái hoạt động của Kibana

```
# curl http://192.168.186.130:5601
```

3.2.3. Cài đặt Wazuh agent trên các máy được giám sát

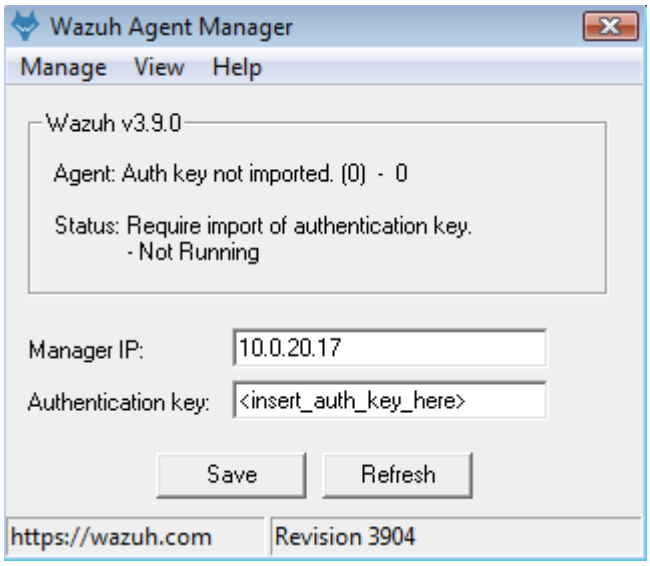
Trên mỗi máy được giám sát, cần tải, cài đặt và đăng ký wazuh agent với hệ thống.

- Tải wazuh agent: tải agent cho Windows từ địa chỉ URL https://packages.wazuh.com/3.x/windows/wazuh-agent-3.10.2-1.msi
- Cài đặt:

+ Cài đặt thông qua giao diện sử dụng Windows Explorer
+ Cặt đặt qua cửa sổ lệnh CMD: msiexec.exe wazuh-agent-3.10.2-1.msi

- Sử dụng lệnh và các chức năng của /var/ossec/bin/manage_agents để thêm và sinh chuỗi xác thực cho mỗi agent:

+ Thêm một agent (tên, địa chỉ IP)
+ Tạo chuỗi xác thực cho agent đó
+ Nạp địa chỉ IP của Wazuh Manager (192.168.186.130) và chuỗi xác thực vào các ô Manager IP và Authentication key, bấm Save và menu Manage / Restart để khởi động lại agent trong giao diện quản lý của agent, như biểu diễn trên Hình 3.2.



Hình 3.2. Giao diện quản lý, đăng ký Wazuh agent với Wazuh Manager

3.3. Thử nghiệm và kết quả

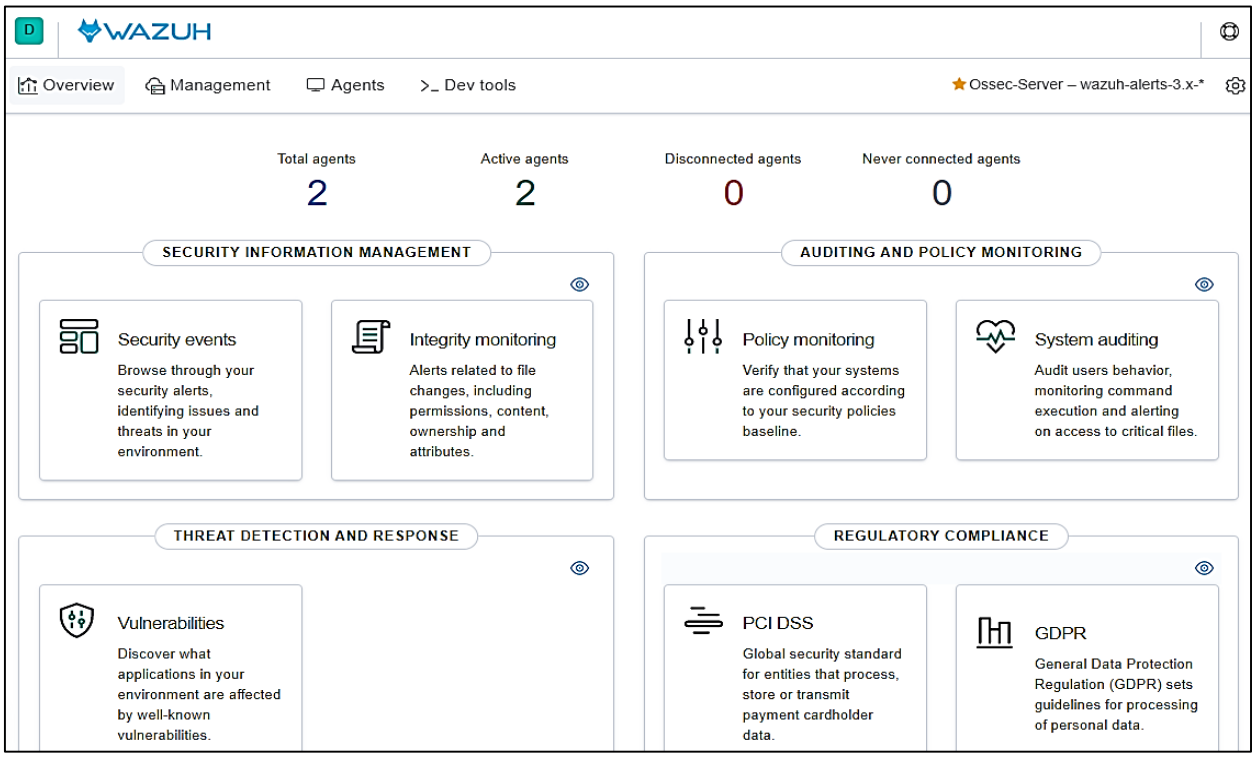
3.3.1. Nội dung thử nghiệm

Sau khi hoàn thành cài đặt và cấu hình xong các thành phần của hệ thống như mô tả trong Mục 3.2, mục này thử nghiệm một số tính năng thu thập, xử lý và phát hiện các bất thường trong hệ thống thử nghiệm. Cụ thể, các tính năng đã thử nghiệm bao gồm:

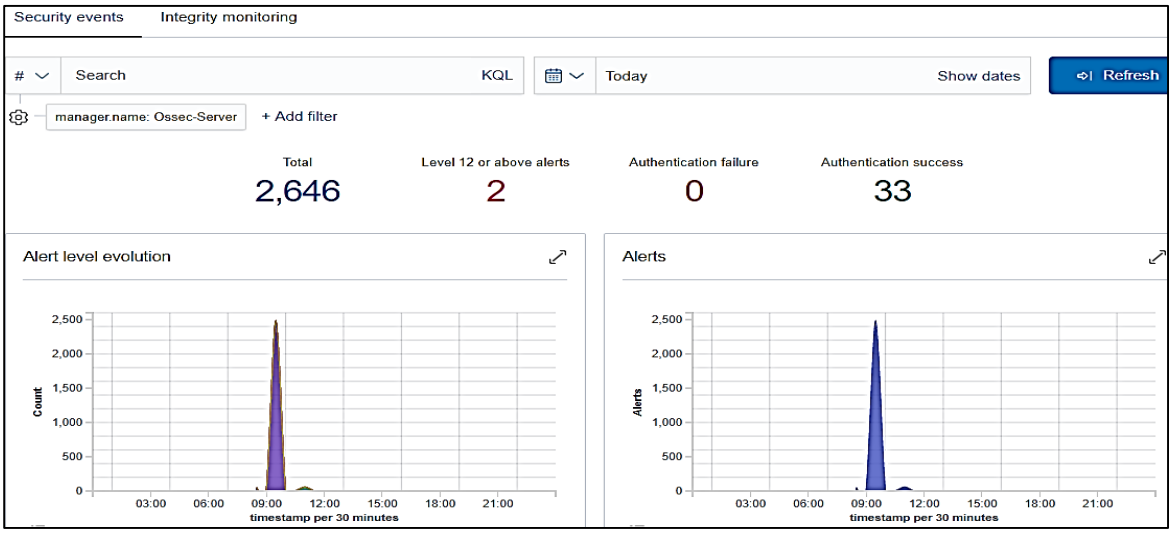
- Hiển thị màn hình tổng hợp các sự kiện an ninh và giám sát toàn vẹn file
- Quản lý hệ thống
- Quản lý và hiển thị thông tin từ các agent
- Công cụ cho nhà phát triển

3.3.2. Kết quả

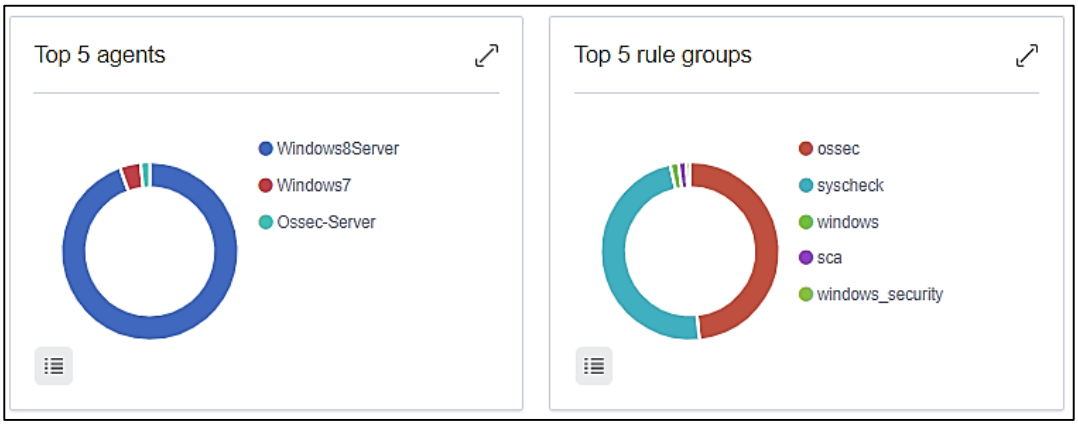
Mục này trình bày một số giao diện hệ thống là kết quả các thử nghiệm các nội dung đã trình bày ở Mục 3.3.1.



Hình 3.3. Giao diện tổng hợp của Wazuh OSSEC-ELK



Hình 3.4. Tổng hợp các sự kiện an ninh

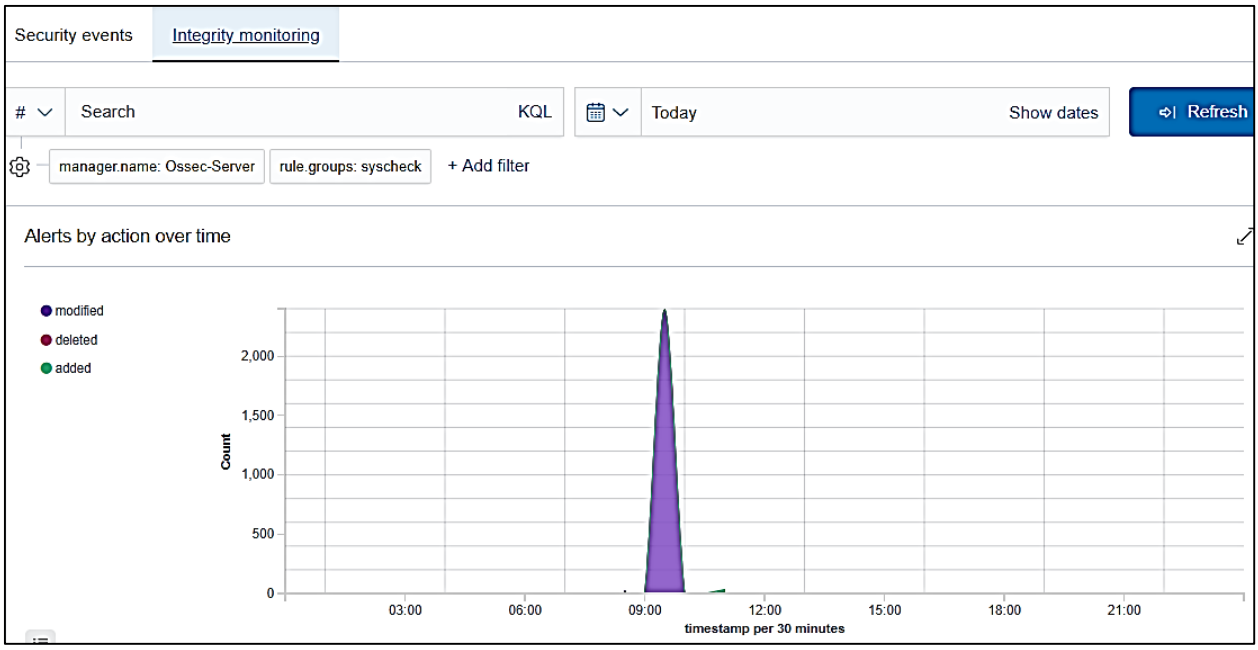


Hình 3.5. Các sự kiện an ninh thu thập từ top 5 agent và top 5 nhóm luật được kích hoạt

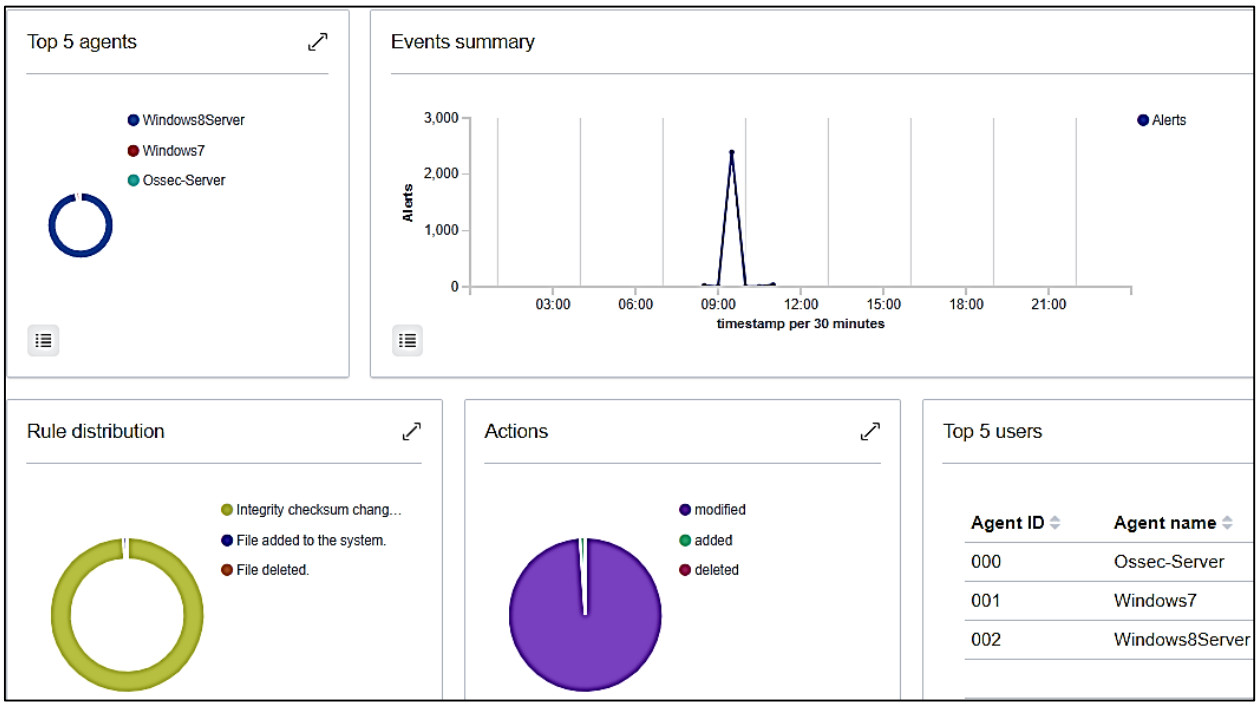
Alerts summary

Rule ID	Description	Level	Count
550	Integrity checksum changed.	7	2,428
554	File added to the system.	5	25
61104	Service startup type was changed	3	18
60106	Windows Logon Success	3	16
5501	PAM: Login session opened.	3	10
60137	Windows User Logoff	3	6
5502	PAM: Login session closed.	3	6
60775	SessionEnv was unavailable to handle a notification event	5	5
60118	Windows Workstation Logon Success	3	5
533	Listened ports status (netstat) changed (new port opened or closed).	7	5

Hình 3.6. Tổng hợp các cảnh báo an ninh



Hình 3.7. Tổng hợp giám sát tính toàn vẹn của file

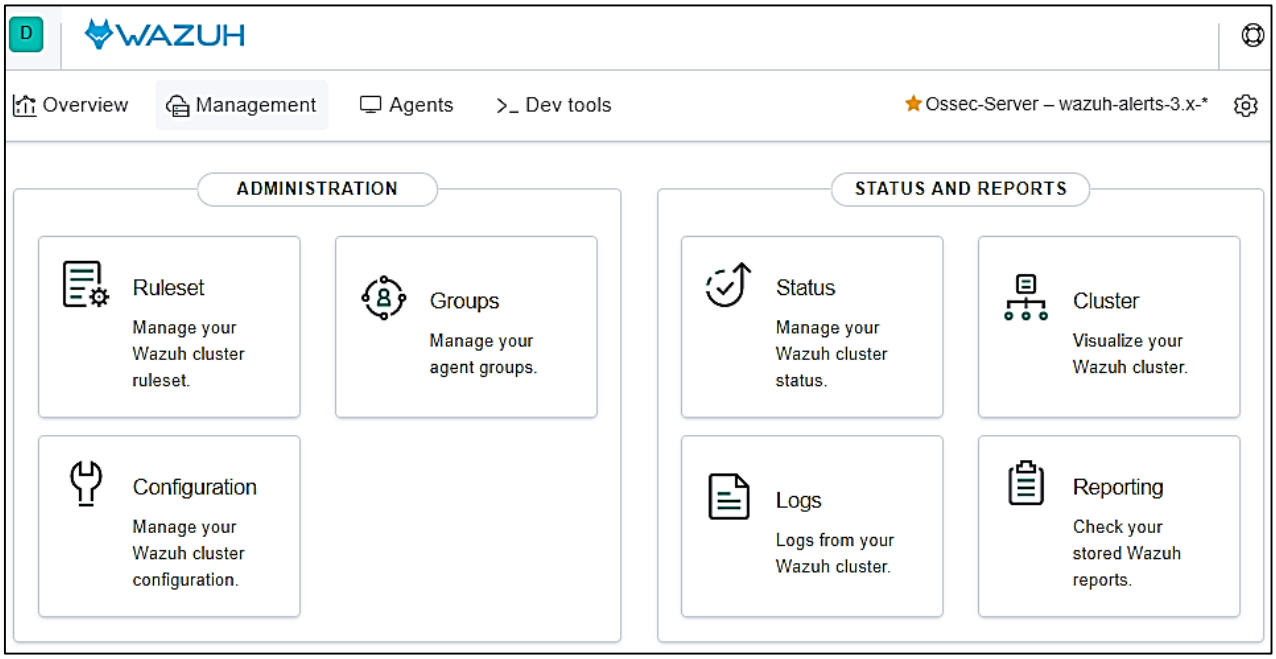


Hình 3.8. Giám sát tính toàn vẹn của file chia theo agent

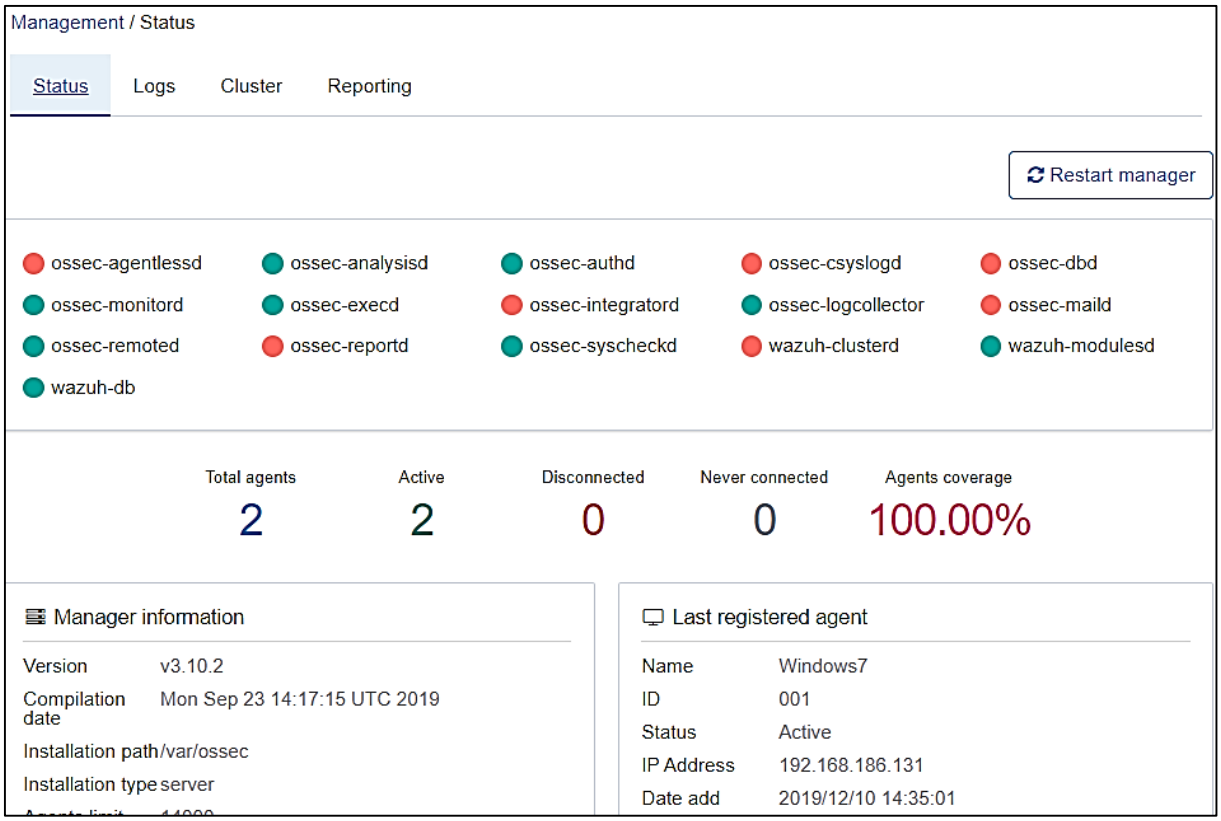
Alerts summary

Agent	Path	Action	Count
Ossec-Server	/etc/cups/subscriptions.conf.O	modified	3
Ossec-Server	/etc/cups/subscriptions.conf	modified	3
Windows7	HKEY_LOCAL_MACHINE\System\CurrentControlSet\Services\VSS\Diag\VolSnap	modified	3
Windows7	HKEY_LOCAL_MACHINE\System\CurrentControlSet\Services\LanmanServer\Parameters	modified	3
Ossec-Server	/boot/grub/grubenv	modified	2
Windows7	HKEY_LOCAL_MACHINE\System\CurrentControlSet\Services\mssmbios\Data	modified	2
Windows7	HKEY_LOCAL_MACHINE\System\CurrentControlSet\Services\VSS\Diag\SPP	modified	2
Windows7	HKEY_LOCAL_MACHINE\Security\SAM\Domains\Account	modified	2
Windows8Server	HKEY_LOCAL_MACHINE\Security\SAM\Domains\Account	modified	2
Windows7	HKEY_LOCAL_MACHINE\System\CurrentControlSet\Services\VSS\Diag\Lovelace	added	1

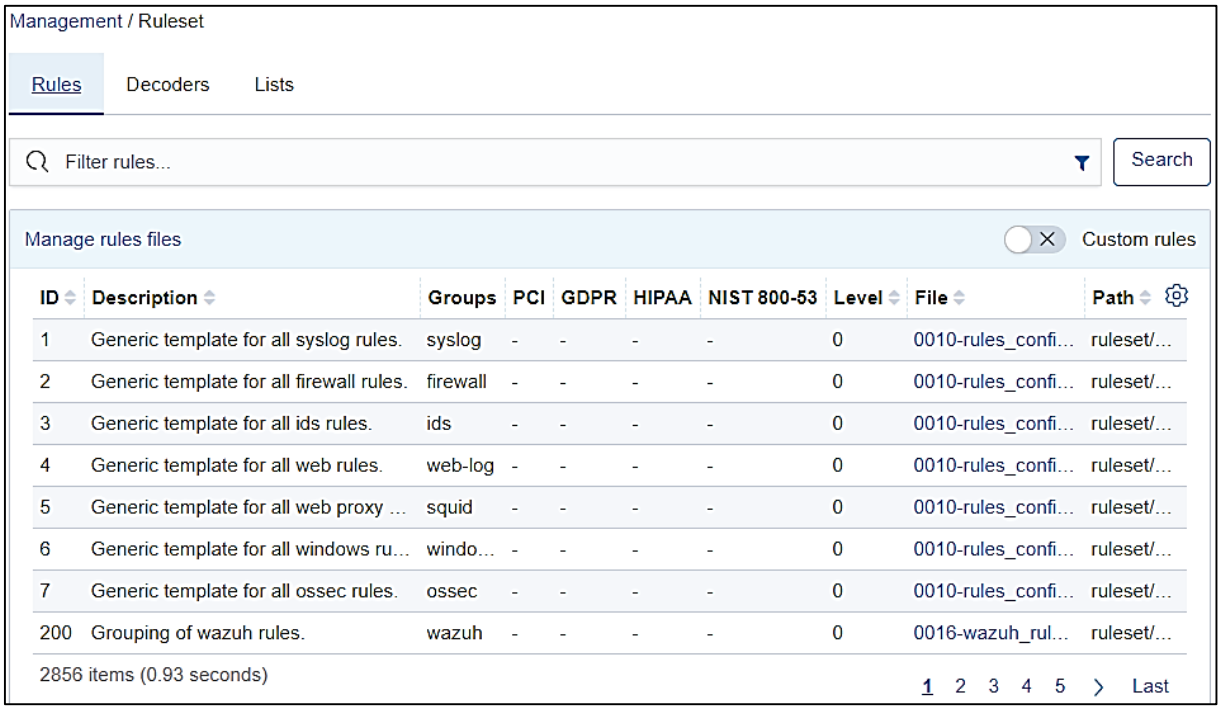
Hình 3.9. Tổng hợp các cảnh báo giám sát toàn vẹn file



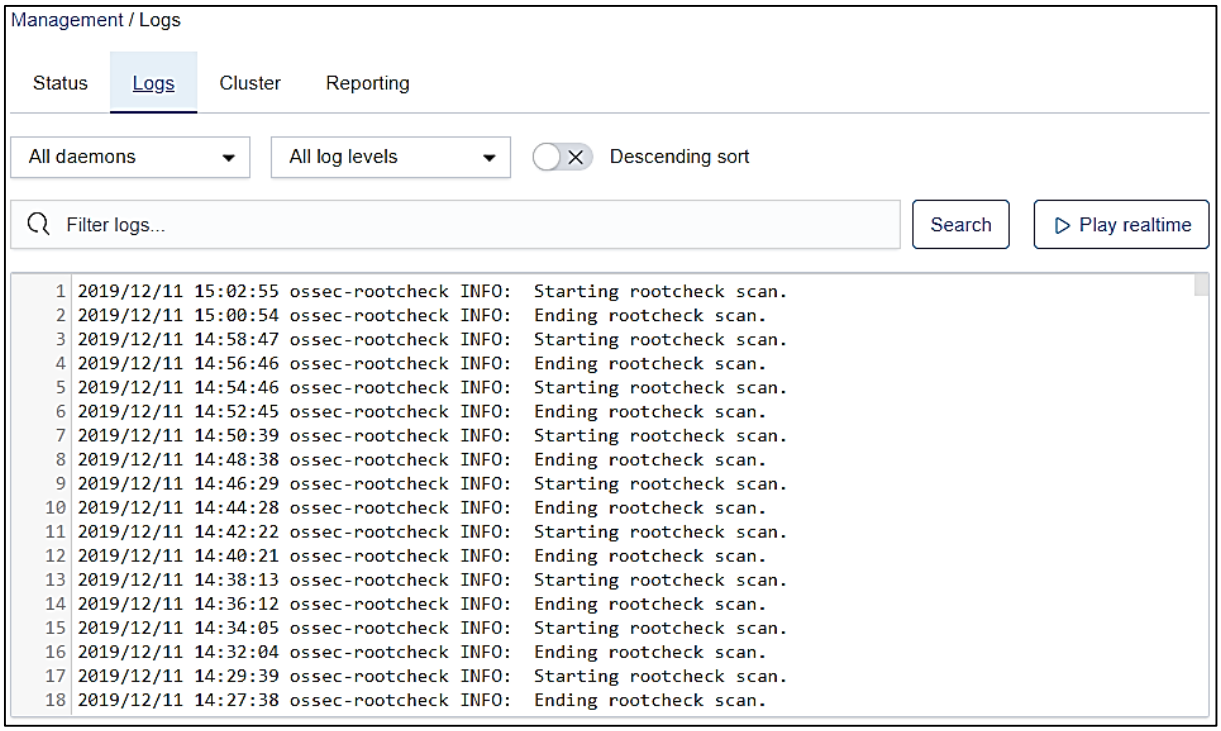
Hình 3.10. Màn hình quản lý hệ thống Management



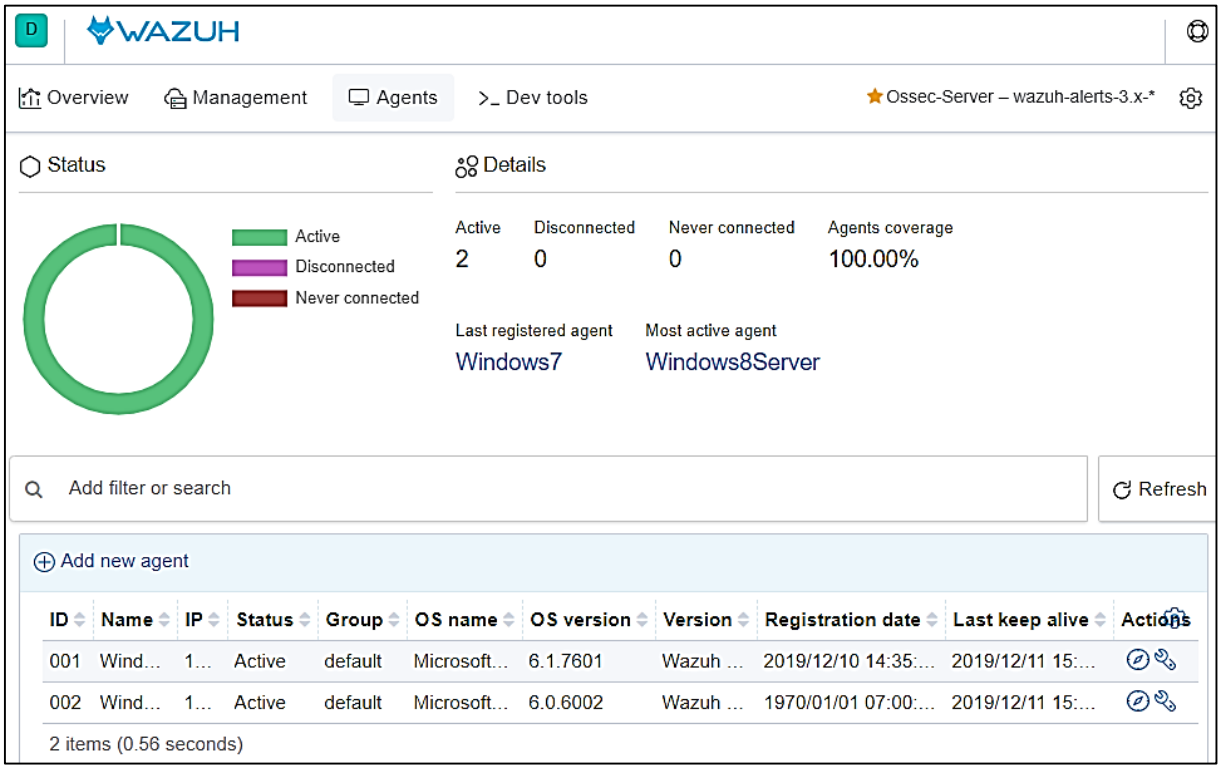
Hình 3.11. Trạng thái hệ thống



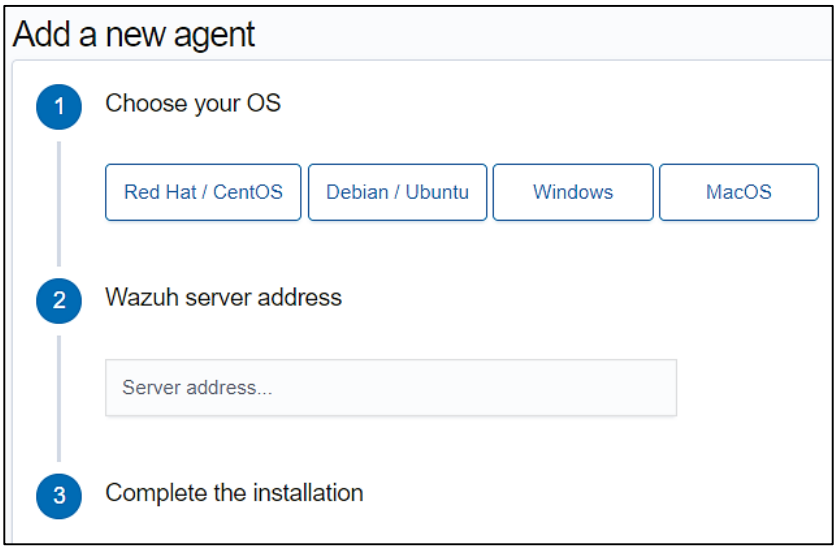
Hình 3.12. Tập luật dựng sẵn của OSSEC



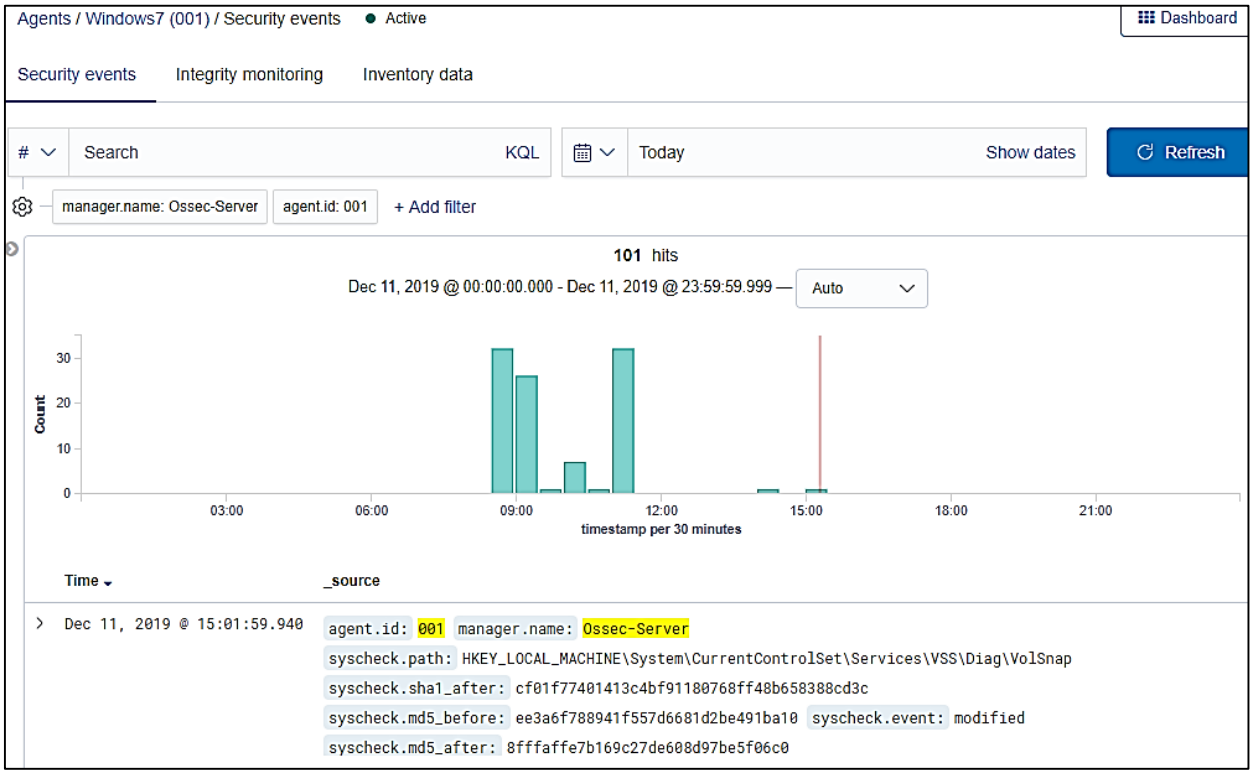
Hình 3.13. Hiển thị log thu thập hỗ trợ hiển thị theo thời gian thực



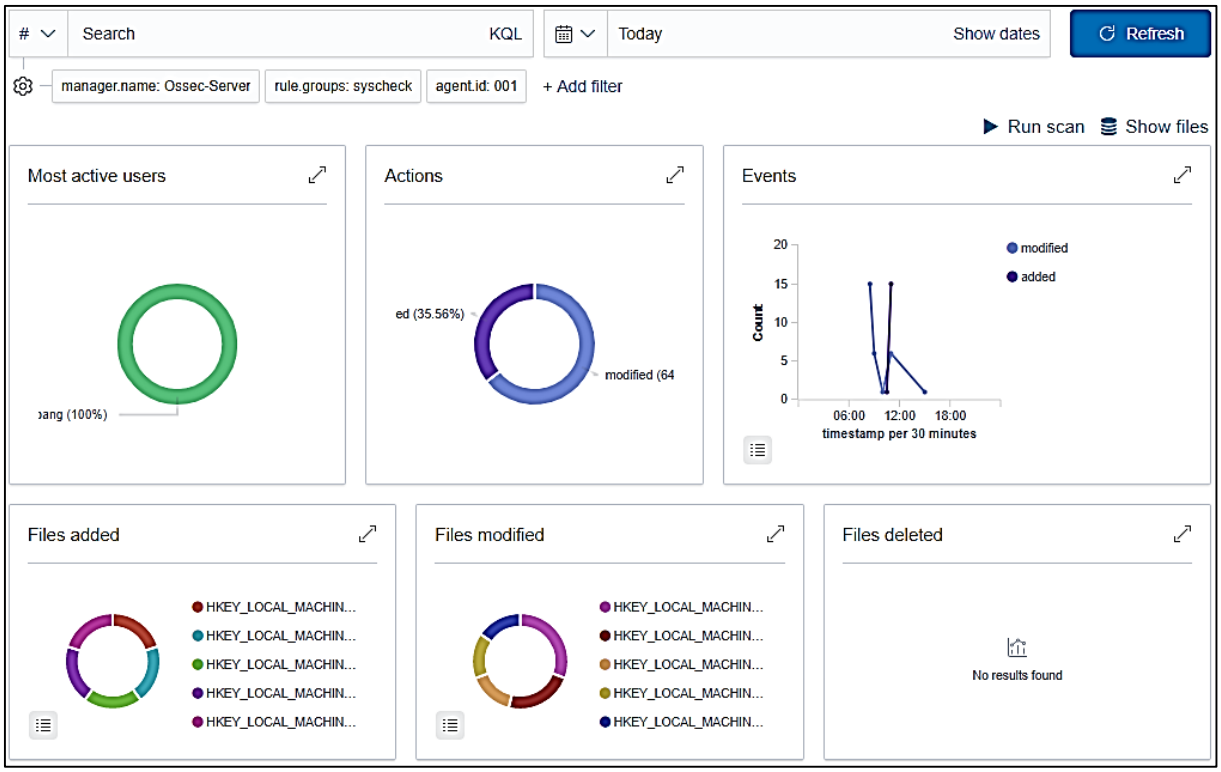
Hình 3.14. Giao diện hiển thị và quản lý các agent



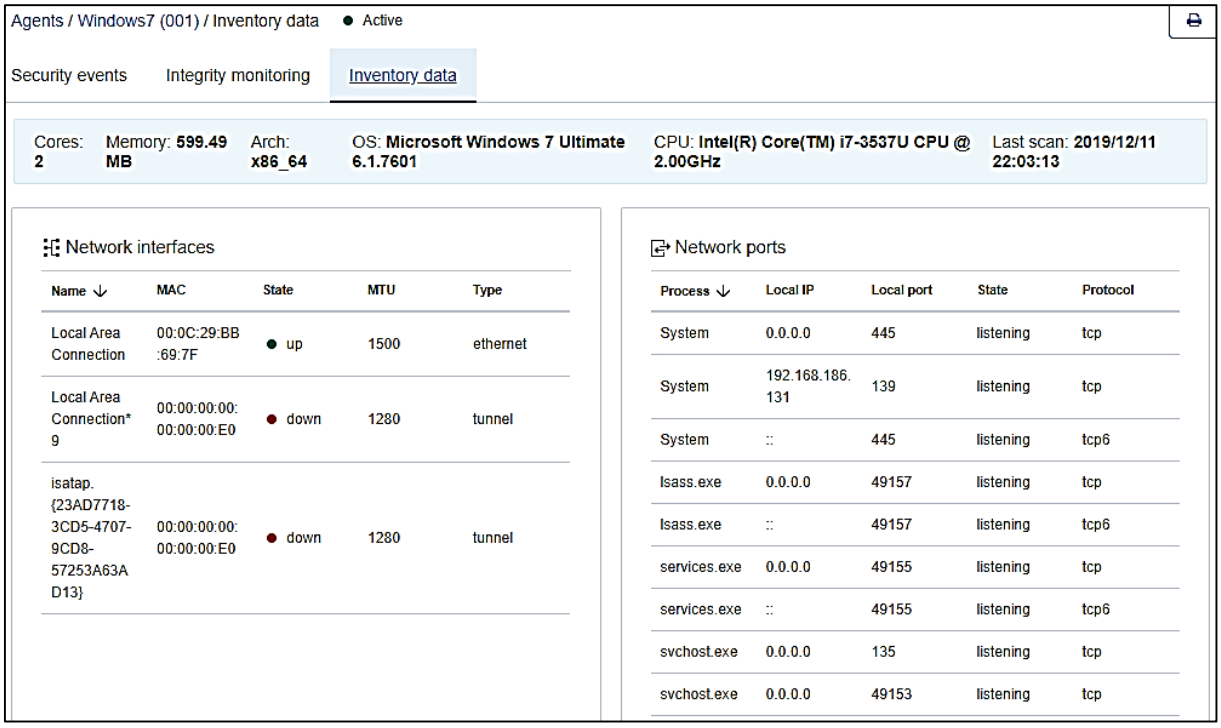
Hình 3.15. Hỗ trợ thêm agent



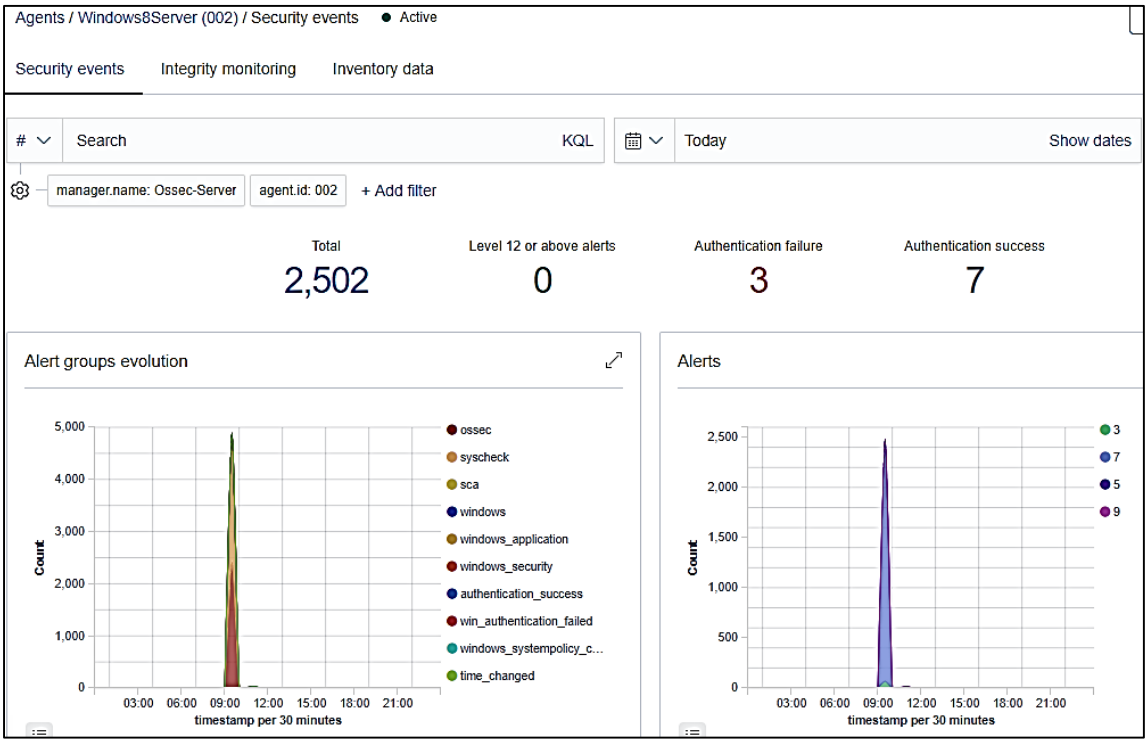
Hình 3.16. Các sự kiện an ninh từ agent số 001



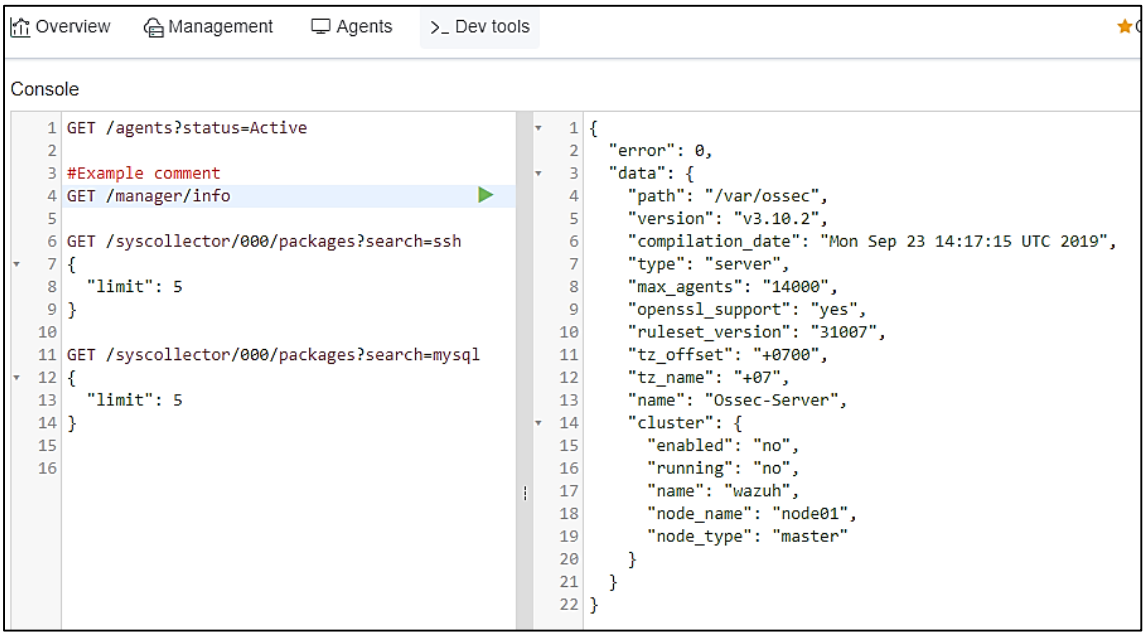
Hình 3.17. Giám sát tính toàn vẹn file từ agent 001



Hình 3.18. Giám sát sử dụng tài nguyên trên máy chạy agent 001



Hình 3.19. Giám sát tổng hợp từ máy chạy agent 002



Hình 3.20. Giao diện hỗ trợ phát triển – trực tiếp chạy các lệnh giám sát

3.3.3. Nhận xét

Từ các thử nghiệm tính năng và các kết quả mô tả trong các mục 3.3.1 và 3.3.2 rút ra một số nhật xét sau:

- Hệ thống thu thập, xử lý và phân tích log cho phép phát hiện bất thường và nguy cơ an toàn thông tin dựa trên sự tích hợp của OSSEC và ELK Stack có khả năng vận hành tốt, cung cấp các tính năng hữu ích cho người quản trị và người dùng thông thường. Cụ thể:

+ Giao diện quản trị và hiển thị đẹp, dễ sử dụng và vận hành;
+ Các thông tin về log và kết quả phân tích, xử lý được biểu diễn theo nhiều dạng trực quan và phong phú;
+ Hệ thống hỗ trợ các tính năng quản trị đơn giản và hiệu quả thông qua giao diện web thân thiện của Kibana, như quản trị tập luật, nhóm, chuỗi và từng agent.

- Hệ thống có khả năng mở rộng tốt:

+ Dễ dàng thêm mới hoặc loại bỏ các agent trên các máy cần giám sát;
+ Có khả năng mở rộng thành chuỗi giám sát thông qua nhiều máy chủ OSSEC kết nối với ELK Stack.

- Hệ thống hoàn toàn mã mở và miễn phí nên rất thích hợp cho triển khai tại các cơ quan, đơn vị có nguồn kinh phí hạn chế.

3.4. Kết luận chương

Chương 3 đã trình bày quá trình thử nghiệm triển khai cài đặt thử nghiệm hệ thống xử lý và phân tích log truy nhập cho phát hiện các bất thường và nguy cơ an toàn thông tin dựa trên việc tích hợp hệ thống phát hiện xâm nhập OSSEC và bộ công cụ xử lý phân tích log ELK. Các kết quả ban đầu cho thấy hệ thống tích hợp vận hành ổn định, có khả năng giám sát và phát hiện các bất thường và nguy cơ ATTT, cũng như hỗ trợ các tính năng quản trị đơn giản và hiển thị dữ liệu, kết quả đa dạng, có tính biểu diễn cao.

KẾT LUẬN

Các kết quả đạt được:

Luận văn này tập trung nghiên cứu về thu thập, xử lý, phân tích log truy cập, phục vụ phát hiện các hành vi bất thường và nguy cơ mất an toàn thông tin trong các hệ thống mạng. Các nội dung đã thực hiện trong luận văn bao gồm:

- Trình bày khái quát về log truy nhập, các dạng log truy nhập, vấn đề thu thập, xử lý và phân tích log truy nhập, cũng như ứng dụng của nó.
- Mô tả một số nền tảng và công cụ xử lý và phân tích log truy nhập, từ đó rút ra so sánh, đánh giá để tìm ra mô hình triển khai phù hợp.
- Trình bày mô hình xử lý và phân tích log khái quát, vấn đề tiền xử lý và chuẩn hóa log và các kỹ thuật phân tích log.
- Xây dựng, cài đặt và thử nghiệm thành công mô hình phân tích log dựa trên OSSEC kết hợp ELK Stack cho phát hiện bất thường và các nguy cơ ATTT trên hệ thống mạng mô phỏng.

Hướng phát triển:

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

- Triển khai thử nghiệm mô hình phân tích log dựa trên OSSEC kết hợp ELK Stack cho phát hiện bất thường và các nguy cơ ATTT trên hệ thống mạng thực.
- Xây dựng và bổ sung thêm các luật giám sát, phát hiện bất thường và các nguy cơ ATTT, đảm bảo khả năng phát hiện kịp thời các bất thường và nguy cơ ATTT.

TÀI LIỆU THAM KHẢO

[1] Access Log, https://xpolog.com/what-is-access-log-101.html, truy nhập tháng 11/2019.

[2] Windows Logging Basics, https://www.loggly.com/ultimate-guide/windows-logging-basics/, truy nhập tháng 11/2019.

[3] Linux/Unix Log, https://www.tutorialspoint.com/unix/unix-system-logging.htm, truy nhập tháng 11/2019.

[4] Linux Logs, https://stackify.com/linux-logs/, truy nhập tháng 11/2019.

[5] Phạm Duy Lộc, Hoàng Xuân Dậu (2017). 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.

[6] IBM QRadar, https://www.ibm.com/ms-en/marketplace/ibm-qradar-siem, truy nhập tháng 11/2019.

[7] VNCS – Giải pháp giám sát website tập trung, http://vncs.vn/portfolio/giai-phap-giam-sat-websites-tap-trung. Truy nhập tháng 11/2019.

[8] OSSEC, https://www.ossec.net, truy nhập tháng 11/2019.

[9] Sumo Logic, http://www.sumologic.com, truy nhập tháng 11/2019.

[10] Graylog, https://www.graylog.org, truy nhập tháng 11/2019.

[11] Webalizer, http://www.webalizer.org, truy nhập tháng 11/2019.

[12] Splunk, http://www.splunk.com, truy nhập tháng 11/2019.

[13] ELK Stack, https://www.elastic.co/what-is/elk-stack, truy nhập tháng 11/2019.

[14] Correlation Analysis, https://www.sciencedirect.com/topics/nursing-and-health-professions/correlation-analysis, truy nhập tháng 11/2019.

[15] Talentica, HIDS Implementation using Ossec, https://www.talentica.com/blogs/hids-implementation-using-ossec/, truy nhập tháng 11/2019.

[16] Wazuh, https://documentation.wazuh.com/3.10/getting-started/architecture.html, truy nhập tháng 11/2019.

[17] ELK Stack, https://www.guru99.com/elk-stack-tutorial.html, truy nhập 11/2019.

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

[19] 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.