HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

-------*-------**

NGUYỄN ANH MINH

NGHIÊN CỨU PHƯƠNG PHÁP PHÂN TÍCH, PHÁT HIỆN TRUY CẬP BẤT THƯỜNG DỰA TRÊN TẬP NHẬT KÝ WEB

LUẬN VĂN THẠC SỸ KỸ THUẬT

(Theo định hướng ứng dụng)

Hà Nội - 2021

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

-------*-------**

NGUYỄN ANH MINH

NGHIÊN CỨU PHƯƠNG PHÁP PHÂN TÍCH, PHÁT HIỆN TRUY CẬP BẤT THƯỜNG DỰA TRÊN TẬP NHẬT KÝ WEB

Chuyên ngành: Hệ thống thông tin
Mã số: 8480104

LUẬN VĂN THẠC SỸ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC

PGS.TSKH. HOÀNG ĐĂNG HẢI

Hà Nội - 2021

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi.

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.

Người viết luận văn

Nguyễn Anh Minh

LỜI CẢM ƠN

Luận văn này đã khép lại quá trình học tập, nghiên cứu của học viên tại Học viện Công nghệ Bưu chính Viễn thông. Học viên xin bày tỏ sự biết ơn sâu sắc tới Thầy hướng dẫn, PGS.TSKH.Hoàng Đăng Hải đã định hướng nghiên cứu và tận tình giúp đỡ, trực tiếp chỉ bảo trong suốt quá trình thực hiện luận văn. Đồng thời học viên cũng xin bày tỏ lòng biết ơn Lãnh đạo Học viện, các thầy cô của Khoa Đào tạo sau đại học, Khoa Công nghệ thông tin 1 tại Học viện Công nghệ Bưu chính Viễn thông.

Trân trọng!

Hà Nội, tháng 5 năm 2021

Học viên

Nguyễn Anh Minh

MỤC LỤC

THUẬT NGỮ VIẾT TẮT

TT	Từ viết tắt	Nghĩa tiếng anh	Nghĩa tiếng việt
1	CLF	Common Log File	Tệp nhật ký chung
2	CSRF	Cross-Site Request Forgery	Giả mạo yêu cầu liên kết trang
3	DoS	Denial of Services	Tấn công từ chối dịch vụ
4	HTTP	HyperText Transfer Protocol	giao thức truyền tải siêu văn bản
5	IDS	Intrusion Detection Systems	Hệ thống phát hiện xâm nhập
6	IIS	Internet Information Services	Dịch vụ thông tin Internet
7	OWASP	Open Web Application Security Project	Dự án mở về bảo mật ứng dụng web
8	SSL	Secure Sockets Layer	Lớp socket bảo mật
9	VNCERT	Vietnam Computer Emergency Response Team	Trung tâm ứng cứu khẩn cấp máy tính Việt Nam

DANH MỤC BẢNG

DANH MỤC HÌNH

MỞ ĐẦU

Ngày nay, khoa học công nghệ ngày càng phát triển, việc phòng, chống tội phạm sử dụng công nghệ cao, chiến tranh trên không gian mạng là vấn đề toàn cầu được nhiều quốc gia trong đó có Việt Nam xác định là một trong những nhiệm vụ trọng tâm trong việc phát triển và bảo vệ đất nước.

Trung tâm ứng cứu khẩn cấp máy tính Việt Nam (VNCERT) chỉ ra rằng chỉ trong tháng 11 của năm 2017, đã có tới gần 600 vụ tấn công, trong đó 248 sự cố Phishing (tấn công lừa đảo), 232 sự cố Deface (tấn công thay đổi giao diện) và 117 sự cố Malware (cài mã độc).. Máy chủ Web là một thành phần rất quan trọng, là mục tiêu của rất nhiều các cuộc tấn công. Vì vậy, việc phân tích các file log, từ đó phát hiện các truy cập bất thường vào máy chủ Web là một nhu cầu thực tế được đặt ra, giúp phán đoán nguy cơ xảy ra các cuộc tấn công vào máy chủ Web.

Dựa vào yêu cầu thực tiễn đặt ra, tôi đã chọn đề tài "**nghiên cứu phương pháp phân tích, phát hiện truy cập bất thường dựa trên tập nhật ký web".** Đây là đề tài có ý nghĩa thực tiễn đối với lĩnh vực an toàn thông tin nói chung và bảo đảm an toàn cho máy chủ Web nói riêng. Hiện tại, những cuộc tấn công vào các hệ thống mạng và hệ thống máy chủ Web đang diễn ra hàng ngày trên toàn thế giới. Vì vậy, đây là một vấn đề có tính cấp thiết, cần phải được nghiên cứu.

Phát hiện truy cập bất thường là bước quan trọng để phát hiện ra tấn công vào máy chủ Web. Đây là bước cơ sở để thực hiện các bước tiếp theo trong việc đảm bảo an toàn dịch vụ Web, phát hiện các hành động xâm nhập trái phép, các tấn công vào máy chủ Web.

Nguyên lý chung để phát hiện bất thường là xây dựng một tập dấu hiệu bình thường của hệ thống (trong điều kiện hoạt động bình thường, không có tấn công), tiếp đó thu thập các hành vi truy cập vào máy chủ, so sánh với tập dấu hiệu bình thường đã lưu sẵn. Nếu có sự khác biệt nghĩa là có hành vi truy cập bất thường.

Đối với máy chủ Web, khi thiết lập hệ thống có thể tạo tập dấu hiệu bình thường và lưu trữ trong máy (có thể trên một máy tính ở ngoài máy chủ). Mọi hành vi truy cập vào máy chủ Web đều được ghi vào Logfile ví dụ như Weblog. Thực hiện thu dữ liệu logfile và phân tích sẽ có thể thu được và tách ra những thông tin cần thiết để phát hiện truy cập bất thường.

Bài luận văn gồm 4 chương chính với những nội dung sau:

Chương 1: Máy chủ web và các vấn đề an toàn web

Chương 2: Phân tích bất thường dựa vào nhật ký máy chủ web

Chương 3: Phát hiện truy cập bất thường vào máy chủ web

Chương 4: Thử nghiệm

CHƯƠNG 1

MÁY CHỦ WEB VÀ CÁC VẤN ĐỀ VỀ AN TOÀN WEB

1.1. Tổng quan về lỗ hổng bảo mật Web

1.1.1. Giới thiệu về máy chủ web

Phần mềm máy chủ hoặc phần cứng dành riêng để chạy các phần mềm trên máy chủ có khả năng cung cấp các dịch vụ World Wide Web được gọi là máy chủ Web (Web server). Các yêu cầu (request) từ các client (mô hình server - client) được Web server xử lý thông qua giao thức HTTP và một số giao thức liên quan khác [1].

Máy chủ Web thường có dung lượng lớn, tốc độ cao, lưu trữ thông tin như một ngân hàng chứa dữ liệu, những website cùng với những thông tin liên quan khác, ví dụ như các chương trình dịch vụ và các file Multimedia, v.v.

Máy chủ Web có khả năng gửi đến máy khách những trang Web thông qua môi trường Internet (hoặc Intranet) qua giao thức HTTP (Hypertext Transfer Protocol) – giao thức được thiết kế để gửi các file đến trình duyệt Web (Web Browser), và các giao thức khác.

Các máy chủ Web đều có một tên miền (Domain Name) hoặc một địa chỉ IP (IP Address). Ví dụ khi đưa **https://qldt.ptit.edu.vn/** vào dòng địa chỉ trên trình duyệt nghĩa là gửi một yêu cầu đến một máy chủ Web có Domain Name là **qldt.ptit.edu.vn**.

Bất kỳ máy tính – máy chủ nào cũng có thể trở thành một máy chủ Web nếu cài đặt lên nó một phần mềm Web Server và có kết nối vào Internet. Khi máy tính của người dùng kết nối đến Web Server và gửi yêu cầu truy cập vào các thông tin trên một trang Web nào đó, Web Server sẽ nhận yêu cầu và gửi lại trình duyệt của người dùng những thông tin mà người dùng mong muốn.

Giống như những phần mềm khác, phần mềm máy chủ Web cũng chỉ là một ứng dụng phần mềm. Phầm mềm máy chủ Web được cài đặt và chạy trên máy tính – máy

chủ dùng để làm Web Server. Nhờ chương trình này, người dùng có thể truy cập vào các thông tin của trang Web từ một máy tính khác ở trên mạng (Internet, Intranet).

Phần mềm máy chủ Web có thể điều khiển việc kết nối vào cơ sở dữ liệu (CSDL) hay được tích hợp với CSDL để truy cập và tải thông tin từ CSDL lên các trang Web và truyền tải chúng đến người dùng.

Máy chủ Web thường sẽ hoạt động 24/24 giờ để phục vụ cho việc cung cấp thông tin liên tục. Vị trí đặt máy chủ Web đóng vai trò quan trọng trong chất lượng và tốc độ truyền thông tin từ máy chủ Web đến máy tính truy cập [2].

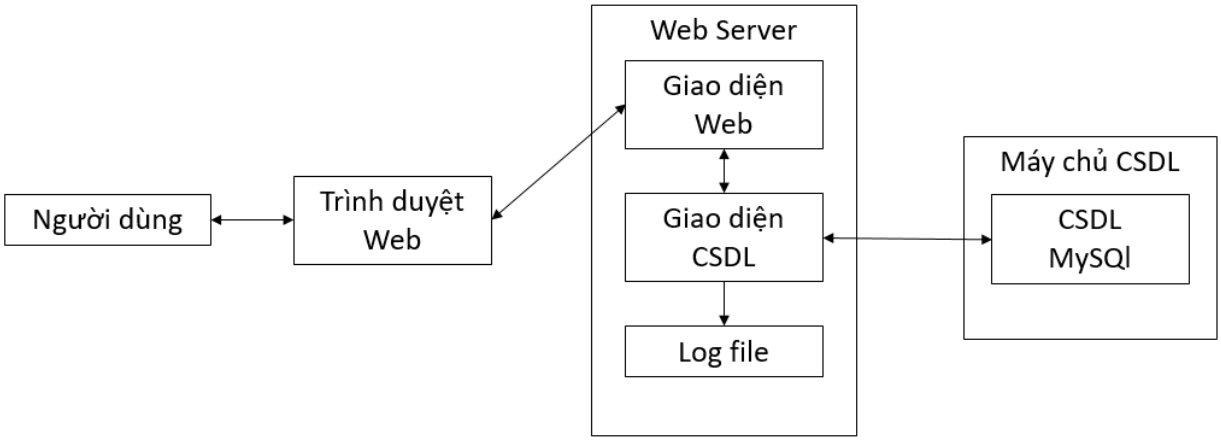
1.1.2. Các thành phần cơ bản của máy chủ web

Như đã nêu ở trên, một máy chủ Web bao gồm một phần mềm Web server cài đặt trên một máy tính hoặc máy chủ.

Máy tính hoặc máy chủ (server) là thiết bị phần cứng cần thiết để cài đặt phần mềm Web server. Tùy vào phạm vi ứng dụng, yêu cầu cung cấp dịch vụ, phần cứng máy chủ có thể có cấu hình từ đơn giản đến phức tạp. Quan trọng đối với hiệu năng của máy chủ Web là tốc độ CPU, dung lượng đĩa cứng lưu trữ và tốc độ kết nối mạng.

Ngoài ra để máy chủ Web hoạt động, còn cần thêm các thành phần quan trọng khác là máy chủ cơ sở dữ liệu (Database server) và máy chủ ứng dụng (Application server). Trong một hệ máy chủ Web đơn gian, hai thành phần nêu trên có thể cùng được cài đặt trên một phần cứng máy tính/máy chủ.

Hình 1.1 là kiến trúc đơn giản cho một hệ thống máy chủ Web. Web server bao gồm thành phần giao diện Web (Presentation layer hay Web interface), thành phần giao diện CSDL (Database interface). Máy chủ CSDL có thể đặt ngoài máy tính cài Web Server, được kết nối với nhau qua giao diện TCP/IP. Giao diện giữa máy chủ Web và trình duyệt của người dùng (Web Browser) được thực hiện qua kết nối HTTP hoặc HTTPS. Toàn bộ hoạt động của máy chủ Web được ghi vào nhật ký (Log file) phục vụ cho việc theo dõi, giám sát hoạt động và tìm lỗi).



Hình 1.1 Kiến trúc hệ thống của Web Server

1.2. Các lỗ hổng bảo mật Web

1.2.1. Khái niệm lỗ hổng bảo mật

Lỗ hổng bảo mật (Security Vulnerability): là một điểm yếu trong hệ thống cho phép kẻ tấn công khai tác gây tổn hại đến an ninh, an toàn hệ thống.

Lỗ hổng bảo mật Web có thể liên quan đến tính toàn vẹn, bí mật, sẵn sàng

Những nguy cơ tiềm ẩn, đối với an toàn ứng dụng web đến từ nhiều nguyên nhân khách quan và chủ quan sau đây:

- Tai họa bất ngờ: tác động đến từ bên ngoài, ảnh hưởng ở mức vật lý của trang web như hỏa hoạn, động đất, lũ lụt, tai nạn lao động...

- Sự cố máy tính: sự cố vật lý ảnh hưởng đến vận hành của trang web như mất điện, hỏng phần cứng, thiết bị nối mạng trục trặc, môi trường vận hành không đảm bảo,...

- Sự cố vô tình: sự cố ảnh hưởng đến hệ thống do yếu tố con người như thiếu kiến thức về bảo mật, chủ quan, cẩu thả trong khâu quản lý hệ thống....

- Sự cố có chủ ý: hoạt động phá hoại, khai thác tấn công làm ảnh hưởng đến an toàn của trang web, bao gồm tội phạm mạng sử dụng công nghệ cao, nhân viên mâu thuẫn với tổ chức, nội gián bán thông tin để nhận hối lộ, nhân viên bị đánh lừa tài khoản hệ thống...

Phải tốn nhiều thời gian và công sức để loại bỏ các mối đe dọa trên. Trước hết, cần nhận thức đầy đủ và rõ ràng về hiểm họa này, sau đó lên kế hoạch thực hiện và phòng tránh rủi ro thích hợp trước mỗi loại nguy cơ.

1.2.2. Các loại lỗ hổng phổ biến của Web

- Các loại lỗ hổng Web phổ biến theo OWASP

Hiểu rõ lỗ hổng bảo mật là một việc có ý nghĩa tối quan trọng. OWASP (Open Web Application Security Project) là một dự án mở về bảo mật ứng dụng web nhằm đảm bảo các ứng dụng web một cách an toàn. Hàng năm OWASP công bố 10 lỗ hổng phổ biến nhất của ứng dụng web trong năm đó. Danh sách này luôn được cập nhật thường xuyên do thay đổi về ảnh hưởng của các lỗ hổng. Theo OWASP năm 2013, 10 lỗ hổng bảo mật Web nghiêm trọng nhất là:

+ Chèn mã (Injection):

Sai sót trong nhập liệu, ví dụ SQL injection, OS injection, LDAP injection... Thông tin không chính xác được đưa vào cùng với các dữ liệu đầu vào như một phần của câu truy vấn. Tin tặc lợi dụng lỗ hổng này để thực hiện các lệnh không hợp pháp hay truy cập các dữ liệu trái phép.

+ Lỗi xác thực, quản lý phiên (Broken Authentication and Session Management):

Lỗ hổng này cho phép tin tặc lợi dụng để lấy password, khóa hay phiên làm việc, từ đó giả mạo phiên làm việc và danh tính của người dùng.

+ Lỗi chéo trang-XSS (Cross-Site Scripting):

Nguyên nhân tồn tại lỗ hổng này là do không kiểm soát dữ liệu nhập vào cẩn thận. Các dữ liệu của kẻ tấn công gửi đến trình duyệt web mà không cần xác nhận. Điều này cho phép hắn thực thi các kịch bản trên trình duyệt web của nạn nhân, thay đổi nội dung, chuyển hướng trang web hay lấy phiên làm việc được người dùng lưu trên trình duyệt...

+ Tham chiếu trực tiếp đối tượng không an toàn (Insecure Direct Object References):

Nhà phát triển ứng dụng web đưa ra tham chiếu đến một đối tượng bên trong ứng dụng chẳng hạn như là một tập tin, một thư mục hay một khóa cơ sở dữ liệu. Nếu quá trình tham chiếu này không được kiểm tra an toàn, hacker có thể dựa vào để tham chiếu đến các dữ liệu không được cấp quyền truy cập

+ Cấu hình bảo mật kém (Security Misconfigtiration):

Một hệ thống bảo mật tốt cần trang bị cho máy chủ ứng dụng, khung ứng dụng, máy chủ cơ sở dữ liệu, nền tảng... các biện pháp bảo mật an toàn, cần thiết và liên kết với nhau. Việc này nhằm giảm nguy cơ bị kẻ tấn công khai thác vào ứng dụng, có thể để lộ những thông tin bí mật khi trao đổi các gói tin.

+ Lộ dữ liệu nhạy cảm (Sensitive Data Exposure):

Các dữ liệu nhạy cảm như thẻ visa, tài khoản banking, mật khẩu tài khoản mạng xã hội... nếu không lưu trữ và bảo vệ an toàn, kẻ gian có thể đánh cắp hoặc chỉnh sửa những thông tin này. Dữ liệu nhạy cảm cần được lưu trữ và bảo vệ một cách cẩn thận, nên mã hoá và sao lưu định kỳ.

+ Thiếu kiểm soát truy cập mức chức năng (Missing Function Level Access Control):

Trong việc phân quyền quản trị các mức, nếu thiếu các điều khoản sẽ dẫn đến việc tin tặc có thể tìm ra các điểm yếu trên hệ thống hay lợi dụng để thay đổi phân quyền.

+ Giả mạo yêu cầu liên kết trang (Cross-Site Request Forgery - CSRF):

Lợi dụng sơ hở của nạn nhân, kẻ tấn công có thể giăng bẫy khiến nạn nhân truy cập vào trang Web giả mạo mà không hề hay biết. Hậu quả có thể là mất tiền, lộ thông tin cá nhân hay mất tài khoản ngân hàng, tài khoản mạng xã hội,…

+ Sử dụng lỗ hổng đã biết (Using Known Vulnerable Components):

Tin tặc sử dụng các thư viện, plugin, module... chứa các lỗ hổng đã được công bố, từ đó tấn công vào hệ thống một cách nhanh chóng.

+ Chuyển hướng không an toàn (Unvalidated Redirects and Forwards):

Chuyển hướng người dùng đến một đường dẫn bên ngoài có thể bị tin tặc lợi dụng để chuyển hướng người dùng đến đường dẫn được hắn chuẩn bị sẵn

OWASP Top 10 - 2013		OWASP Top 10 - 2017
A1 – Injection	→	A1:2017-Injection
A2 – Broken Authentication and Session Management	→	A2:2017-Broken Authentication
A3 – Cross-Site Scripting (XSS)	↘	A3:2017-Sensitive Data Exposure
A4 – Insecure Direct Object References [Merged+A7]	U	A4:2017-XML External Entities (XXE) [NEW]
A5 – Security Misconfiguration	↘	A5:2017-Broken Access Control [Merged]
A6 – Sensitive Data Exposure	↗	A6:2017-Security Misconfiguration
A7 – Missing Function Level Access Contr [Merged+A4]	U	A7:2017-Cross-Site Scripting (XSS)
A8 – Cross-Site Request Forgery (CSRF)	☒	A8:2017-Insecure Deserialization [NEW, Community]
A9 – Using Components with Known Vulnerabilities	→	A9:2017-Using Components with Known Vulnerabilities
A10 – Unvalidated Redirects and Forwards	☒	A10:2017-Insufficient Logging&Monitoring [NEW,Comm.]

Hình 1.2. Các lỗ hổng bảo mật phổ biến nhất

1.3. Tấn công vào máy chủ Web

1.3.1. Giới thiệu về tấn công vào máy chủ Web

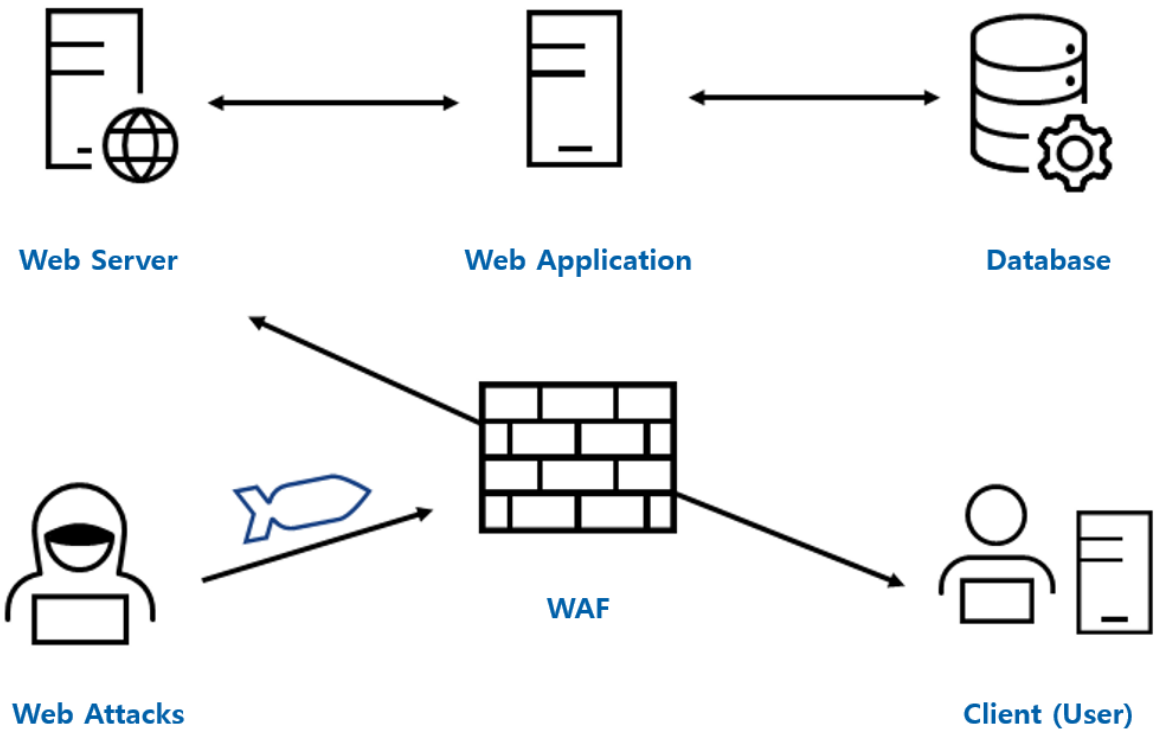
Tấn công vào máy chủ Web là hình thức kẻ tấn công tìm cách khai thác các lỗ hổng đã biết hoặc chưa biết trên máy chủ Web nhằm đánh cắp thông tin từ máy chủ, phá hoại hoạt động hoặc gây gián đoạn, ngưng trệ dịch vụ Web. Đối tượng bị tấn công có thể là cá nhân, doanh nghiệp, tổ chức hoặc cơ quan nhà nước.

Kẻ tấn công có thể dùng công cụ bắt gói tin tự động, rà quét các lỗ hổng trong hệ thống, quét cổng, và kiểm tra các dịch vụ đang chạy với mục đích là thăm dò, thu thập thông tin về hệ thống. Thông qua các lỗ hổng trong dịch vụ web, đường truyền, dịch vụ xác thực, kẻ tấn công có thể dễ dàng truy cập vào các tài khoản của quản trị viên như trong cơ sở dữ liệu, website, ứng dụng, phần mềm quản lý… để lấy đi những thông tin, dữ liệu quan trọng.

Tin tặc có thể quấy phá hoạt động bằng cách xóa dữ liệu, mã hóa dữ liệu, tống tiền chủ sở hữu máy chủ để lấy lại được dữ liệu cần thiết. Ngoài ra, phương pháp xâm nhập như email lừa đảo, đường link lạ, thông báo trúng thưởng giả mạo thông qua các đường link mã độc cũng là một kỹ thuật được tin tặc áp dụng thường xuyên.

Với kiểu tấn công DDoS, kẻ tấn công sẽ gửi hàng loạt những yêu cầu với số lượng cực lớn (vượt quá khả năng xử lý tới hệ thống nạn nhân), làm cho hoạt động của hệ thống máy chủ Web bị tạm dừng

Ví dụ về tấn công máy chủ Web: Trong cùng một hệ thống máy chủ Web nội bộ, kẻ tấn công có thể xâm nhập vào hệ thống Web của công ty bằng cách đóng vai như một người dùng thật trong hệ thống, sau đó tiến hành xâm nhập vào tệp chứa tài liệu bí mật của công ty. Tin tặc có thể rút sạch tiền, hoặc thay đổi các con số, ẩn file…



Hình 1.3. Mô hình tấn công mạng theo phương pháp truy cập trực tiếp

Các tấn công có thể gây ra những thiệt hại, tác động lớn đến hoạt động của một tổ chức.

- Danh tiếng có thể bị hủy hoại: tin tặc chỉnh sửa các nội dung trên trang web thành các thông tin vi phạm pháp luật hoặc liên kết đến một trang web khiêu dâm.
- Tin tặc có thể lợi dụng các máy chủ web để cài đặt phần mềm độc hại. Các phần mềm độc hại được tải về máy tính của người dùng có thể là virus, Trojan, worm hoặc phần mềm botnet, v..v…
- Dữ liệu người dùng bị lộ lọt: có thể dẫn đến kinh doanh suy giảm hoặc bị kiện bởi những người dùng cung cấp thông tin của họ cho tổ chức.

1.3.2. Một số loại tấn công điển hình vào máy chủ Web

- Tấn công chuyển dịch thư mục (Directory traversal attacks). Đây là loại tấn công khai thác lỗi trong máy chủ web để truy cập trái phép vào các tệp tin và thư mục. Khi tin tặc đã đạt được quyền truy cập, chúng có thể đăng tải những thông tin nhạy cảm, thực thi lệnh trên máy chủ hoặc cài đặt phần mềm độc hại.
- Tấn công từ chối dịch vụ (Denial of Service Attacks). Với kiểu tấn công này, máy chủ web có thể bị sập hoặc trở nên không có tính sẵn dùng cho người sử dụng hợp pháp.
- Tấn công chiếm giữ hệ thống tên miền (Domain Name System Hijacking). Các thiết lập DNS được thay đổi để trỏ đến trang web của kẻ tấn công. Tất cả lưu lượng lẽ ra phải được gửi đến máy chủ web bị chuyển sai hướng.
- Tấn công nghe lén (Sniffing). Dữ liệu không mã hóa gửi qua mạng có thể bị chặn và được sử dụng để truy cập trái phép vào máy chủ web để nghe lén thông tin.
- Tấn công giả mạo (Phishing). Kẻ tấn công giả các trang web và chuyển hướng đến một trang web giả mạo.
- Tấn công đầu độc (Pharming). là một loại tấn công mạng liên quan đến việc chuyển hướng lưu lượng truy cập web từ trang hợp pháp sang một trang giả mạo. Trang giả mạo này được thiết kế để trông giống như trang web hợp pháp, do đó người dùng sẽ bị lừa khi đăng nhập và nhập thông tin chi tiết của mình vào đó. Những chi tiết này sau đó được thu thập bởi các "pharmer" và sử dụng cho các hoạt động bất hợp pháp.

- Tấn công thay giao diện (Defacement). Kẻ tấn công thay thế trang web của tổ chức với một trang khác có chứa tên, hình ảnh và có thể bao gồm nhạc nền và tin nhắn của kẻ tấn công.

1.3.3. Một số biện pháp điển hình chống tấn công vào máy chủ Web

Các biện pháp chống tấn công vào máy chủ Web có thể được phân loại theo: các biện pháp quản lý và các biện pháp kỹ thuật.

Các biện pháp quản lý bao trùm phạm vi khá rộng từ: bảo vệ vật lý cho đến các chính sách bảo mật trong tổ chức, quản lý. Các biện pháp kỹ thuật cũng rất đa dạng, từ việc mã hóa dữ liệu, lưu trữ dự phòng và bảo vệ dữ liệu hệ thống, bảo vệ hạ tầng thiết bị và đường truyền, cập nhật phần mềm, sử dụng các thiết bị bảo vệ như tường lửa, hệ thống chống tấn công,…

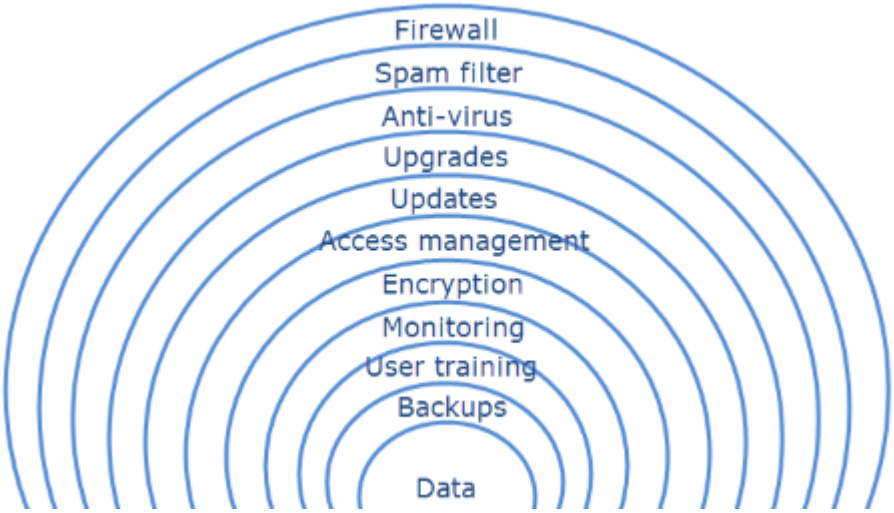
Nhìn chung, một số biện pháp cơ bản có thể liệt kê gồm:

- Cập nhật và cài đặt các bản vá lỗi thường xuyên để giúp đảm bảo các máy chủ.
- Bảo mật các cài đặt và cấu hình của hệ điều hành.
- Bảo mật các cài đặt và cấu hình của phần mềm máy chủ web.
- Sử dụng các công cụ như Snort, Nmap, Scanner Access Now Easy (SANE).
- Sử dụng firewall.
- Sử dụng phần mềm diệt virus, mã độc như BKAV, Kaspersky,...
- Vô hiệu hóa chức năng quản trị từ xa.
- Xóa các tài khoản mặc định và không sử dụng khỏi hệ thống.
- Cổng và cài đặt mặc định (như FTP ở cổng 21) nên được thay đổi (cổng FTP 5069).

Các biện pháp bảo vệ theo chiều sâu bao gồm: bảo vệ vòng ngoài với tường lửa, tiếp đến là vòng trong với hệ thống phát hiện và ngăn chặn xâm nhập trái phép (IDS/ISP), bảo vệ miền DMZ. Vành đai bảo vệ tiếp theo là hệ thống chống virus, mã độc. Tiếp đến

là bảo vệ cục bộ với việc kiểm soát, xác thực truy cập (login) vào hệ thống dữ liệu, CSDL. Cuối cùng là lớp bảo vệ các ứng dụng, dịch vụ Web.

Các biện pháp rà quét, kiểm tra thường xuyên các lỗ hổng bảo mật máy chủ Web cũng đóng vai trò quan trọng trong tăng cường bảo mật, chống tấn công. Ngoài ra, quản trị hệ thống có thể tắt các dịch vụ không cần thiết để hạn chế tối đa khả năng khai thác của tin tặc.



Hình 1.4 Các biện pháp bảo vệ theo chiều sâu

1.4. Kết luận chương

Trong chương này, luận văn đã trình bày về các nội dung: giới thiệu về máy chủ Web, các thành phần cơ bản của máy chủ Web. Ngoài ra luận văn cũng đã trình bày về các lỗ hổng bảo mật Web, các loại tấn công vào máy chủ Web.

CHƯƠNG 2
PHÂN TÍCH BẤT THƯỜNG DỰA VÀO NHẬT KÝ MÁY CHỦ WEB

2.1 Một số nền tảng Apache, IIS, Nginx

- **Nền tảng Apache**

Apache là một máy chủ Web phổ biến nhất trên thế giới cho phép thiết lập một website dễ dàng không tốn nhiều công sức. Các doanh nghiệp nhỏ, các ứng dụng quy mô nhỏ thường chọn máy chủ web này

Có thể cài đặt một trang tin WordPress trên máy chủ Apache Web Server mà không phải tùy chỉnh bất cứ gì. Hơn nữa, Apache server hoạt động tốt với các hệ thống quản trị nội dung lớn trên thế giới như Joomla, Drupal, …, web frameworks (Django, Laravel, etc.), và các ngôn ngữ lập trình khác. Điều này giúp Apache giữ vững vị trí số một trong số các nền tảng web hosting, đặc biệt là đối với VPS hoặc shared hosting.

Apache là phần mềm Web Server miễn phí mã nguồn mở, đang chiếm khoảng 46% thị phần trang Web trên toàn thế giới. Tên đầy đủ của Apache là Apache HTTP Server, được điều hành và phát triển bởi công ty Apache Software Foundation.

Apache Web Server là lựa chọn ưu việt để tạo ra một website ổn định và có thể tùy chỉnh linh hoạt

- **Nền tảng IIS**

IIS (viết tắt của Internet Information Services) được đính kèm cùng với các phiên bản của Windows. IIS gồm các dịch vụ máy chủ chạy trên nền hệ điều hành Window, cung cấp và phân phối các thông tin lên mạng. IIS gồm có nhiều dịch vụ khác nhau như Web Server, FTP Server…

Các thành phần chính của IIS như sau.

+ Khối quản trị IIS (IIS Manager):

IIS Manager dùng để quản trị IIS Server. Nó quản lý tài nguyên các file, thư mục và các thiết lập cho các ứng dụng như về security, performance và các tính năng khác.

+ Khối chính sách kiểm toán (Audit Policy):

Cần thiết lập Audit Policy trên IIS Server trong môi trường làm việc đảm bảo toàn bộ thông tin của người dùng khi log vào hệ thống sẽ đều được ghi lại. Tất cả những dữ liệu được truy cập đều được log lại.

+ Khối cấp quyền người dùng (*User Rights Assignments):*

Cần cài đặt "Deny access to this computer from the network". Cài đặt này quyết định những người dùng nào bị cấm truy cập tới IIS Server từ mạng.

Thiết lập khiến cho tài khoản người dùng sẽ bị hạn chế và đảm bảo tính bảo mật cao hơn. Điển hình là các cấu hình: ANONOYMOUS LOGON, Built-in Administrator, Suport_388945a0, ..

+ Khối nhật ký sự kiện (*Event Log):*

Trên IIS Servers, toàn bộ các sự kiện cần lưu lại theo một thể thống nhất với các tham số tuỳ vào yêu cầu. Quá trình đăng nhập vào hệ thống và ghi lại những đối tượng được truy cập là hai yêu cầu cần được ghi lại (kể cả lỗi hay không trong quá trình đăng nhập).

+ Khối dịch vụ hệ thống (*SYSTEM Services):*

Các dịch vụ trong Microsoft Windows Server cần được thiết lập ở chế độ tự động gồm:

- HTTP SSL
- IIS Admin Service
- World Wide Web Publishing Service

+ Thiết lập quyền trong IIS (*IIS Web Site Permissions):*

IIS có thể thiết lập Website permissions để quyết định cho phép hay không những hành động truy cập vào website. Một số quyền có thể gán trong Web Site Permissions

- Read: Chỉ được xem các nội dung và các thuộc tính của các thư mục hoặc tập tin.
- Write: Có khả năng thay đổi nội dung và thuộc tính của các thư mục hoặc tập tin.

- Log Visits: Ghi lại truy cập của người dùng vào Website
- Excute: Thực thi script.

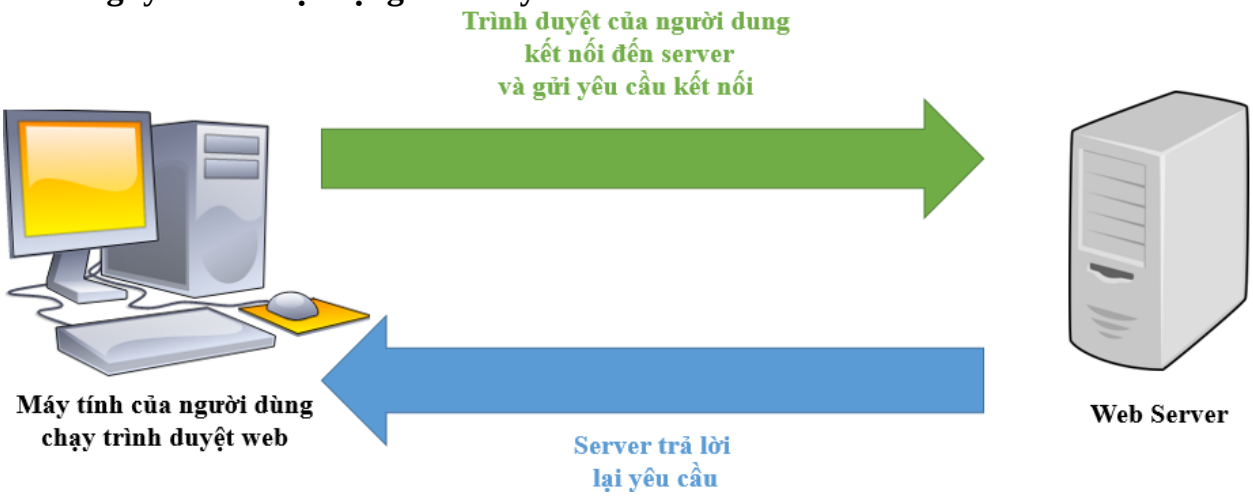
- **Nền tảng Nginx**

Nginx là một nền tảng máy chủ Web sử dụng phổ biến giao thức HTTP, HTTPS, SMTP, POP3, IMAP đồng thời tạo cân bằng tải. **Nginx** chú trọng vào việc phục vụ số lượng lớn kết nối đồng thời, sử dụng bộ nhớ thấp và đạt hiệu suất cao. Nginx có sự ổn định lớn, cấu hình đơn giản, nhiều tính năng và tiết kiệm tài nguyên.

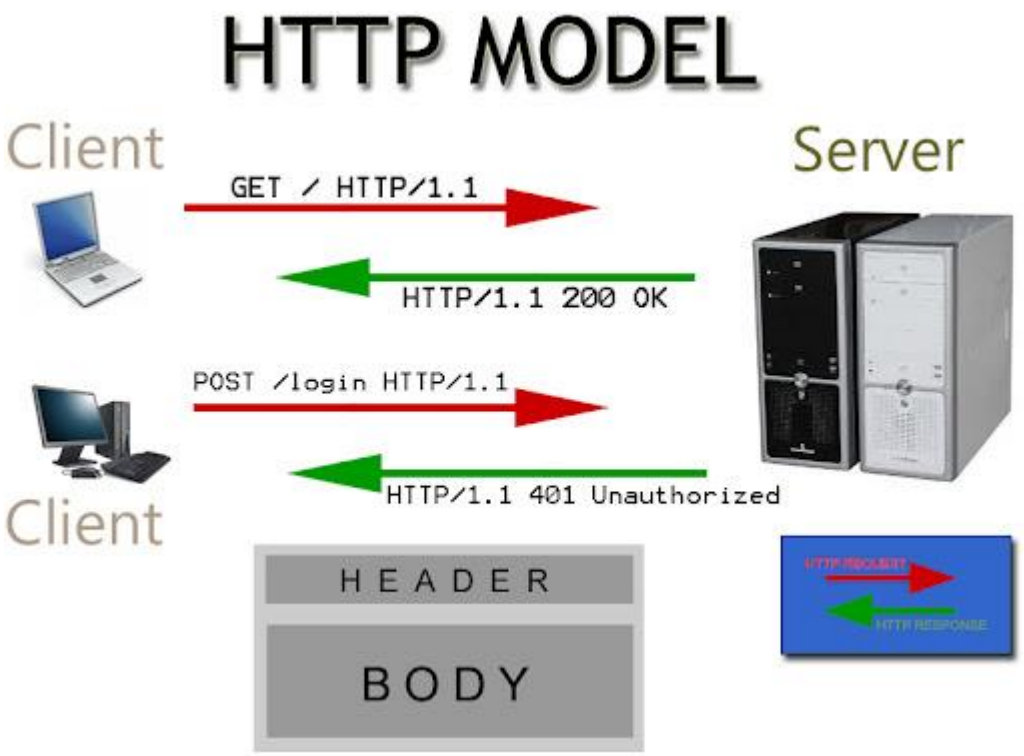
Không giống các nền tảng máy chủ khác, thay vì dựa vào luồng (threads) để xử lý các truy vấn (request), Nginx sử dụng kiến trúc hướng sự kiện (event-driven) không đồng bộ và có khả năng mở rộng. Do hiệu suất cao và yêu cầu bộ nhớ thấp, Nginx vẫn nên được sử dụng ngay cả khi không cần phải xử lý hàng ngàn truy vấn đồng thời. Nginx có thể được sử dụng trên máy chủ cấu hình thấp nhất cho đến một hệ thống lớn như trên đám mây.

2.2 Phương pháp ghi nhật ký máy chủ Web

2.2.1 Nguyên tắc hoạt động của máy chủ Web



Hình 2.1 Các bước trong tiến trình truyền tải web



Hình 2.2. Yêu cầu, phản hồi của HTTP

Nguyên tắc hoạt động cơ bản của máy chủ Web được thể hiện trên hình 2.1. Người dùng gõ dòng địa chỉ máy chủ Web vào trình duyệt web và ấn Enter, trang web sẽ hiển thị trên màn hình máy tính người dùng. Để trang web có thể hiển thị được thì cơ chế hoạt động của máy chủ web được thể hiện qua các bước cơ bản trong tiến trình truyền tải trang web đến màn hình người dùng như sau:

Tiến trình cơ bản:

Browser thực hiện kết nối tớiWeb server, yêu cầu một trang web và nhận lại nó. Trình tự từng bước xảy ra như sau:

Trình duyệt web tách địa chỉ của một website làm 3 phần như sau:

- Tên giao thức: “http”
- Tên miền của máy chủ web: “**qldt.ptit.edu.vn/**”
- Tên tệp HTML: “web-server.htm”

* Trình duyệt gửi yêu cầu kết nối tới máy chủ Web (bản tin HTTP request).

* Máy chủ Web trả lời bằng bản tin HTTP response.

* Căn cứ thông tin trong bản tin yêu cầu, máy chủ Web liên hệ với máy chủ tên miền (DNS Server) để chuyển đổi tên miền "**https://qldt.ptit.edu.vn/**" ra địa chỉ IP tương ứng. Tiếp theo, trình duyệt gửi tiếp một kết nối tới máy chủ của website có địa chỉ IP này qua cổng 80. Nhờ giao thức HTTP, browser gửi yêu cầu GET đến máy chủ, yêu cầu tệp HTML "web-server.htm".

Lưu ý: một cookies cũng được gửi kèm từ trình duyệt web đến máy chủ.

Sau đó, một file văn bản có các thẻ HTML sẽ được máy chủ gửi đến trình duyệt web của máy yêu cầu (máy chủ cũng gửi kèm theo một cookies khác tới trình duyệt web, cookies này được ghi trên đầu trang của mỗi trang web).

Các thẻ HTML sẽ được trình duyệt web đọc để xác định định dạng (cách thức trình bày) trang web và hiển thị nội dung ra trên màn hình yêu cầu.

Trong giao thức HTTP nguyên bản, cần cung cấp đầy đủ đường dẫn tên tệp. Ví dụ như "/" hoặc "/tên tệp.htm". Sau đó, giao thức sẽ tự điều chỉnh để cho ra một địa chỉ URL đầy đủ. Các công ty kinh doanh dịch vụ lưu trữ có thể lưu trữ nhiều tên miền ảo (virtual domains) nhờ việc này. Có nghĩa là có thể có nhiều tên miền cùng tồn tại trên một máy chủ và sử dụng cùng một địa chỉ IP duy nhất. Ví dụ, trên máy chủ của Học viện Công nghệ Bưu chính Viễn thông, địa chỉ IP là 123.30.171.25, nhưng nó có rất nhiều tên miền khác nhau cùng tồn tại.

Nhiều máy chủ web thêm vào một số chế độ bảo mật trong tiến trình xử lý. Ví dụ, khi ta truy cập vào một trang web và trình duyệt đưa ra một hộp hội thoại yêu cầu nhập tên truy cập và mật khẩu, lúc này trang web mà ta truy cập đã được bảo vệ bằng mật khẩu.

Máy chủ web hỗ trợ người quản lý trang web duy trì một danh sách bao gồm tên và password cho phép những người được phép truy cập vào trang web. Với máy chủ web chuyên nghiệp, yêu cầu mức độ bảo mật lớn ví dụ như: chỉ cho kết nối đã được mã hóa

giữa máy chủ và trình duyệt. Do đó những thông tin nhạy cảm như password… có thể được truyền tải lên Internet.

Bên trên là những vấn đề cơ bản về cách thức hoạt động của máy chủ Web để truyền tải các trang web chuẩn, hay còn gọi là trang web tĩnh. Các trang web tĩnh là những trang web không thay đổi (Chỉ thay đổi khi người tạo ra thay đổi lại).

2.2.2 Giao thức HTTP

HTTP (HyperText Transfer Protocol) có nghĩa là giao thức truyền tải siêu văn bản, là một trong năm giao thức chuẩn của mạng Internet, dùng để liên hệ thông tin giữa máy chủ Web và máy sử dụng dịch vụ (Web client hay Web browser) trong mô hình Client-Server dùng cho World Wide Web-WWW. HTTP là một giao thức ứng dụng của bộ giao thức TCP/IP. HTTP nằm trong tầng Application Layer, được sử dụng để truyền tải nội dung trang Web từ Web Server đến trình duyệt Web ở Client.

Request-Response là cơ chế hoạt động chính của HTTP: Web Client sẽ gửi Request đến máy chủ web, máy chủ web xử lý và trả về cho Web Client Response

Phiên bản hoàn chỉnh đầu tiên của HTTP là **HTTP 0.9** (Ra đời năm 1991), tiếp theo là **HTTP 1.0** (1996), **HTTP 1.1** (1997) và mới nhất là **HTTP 2.0**. Các phiên bản sau ra đời thay thế phiên bản trước, kế thừa những chức năng cốt lõi của phiên bản trước nhưng có nhiều cải tiến và bổ sung. Hiện nay, HTTP 2.0 chưa được dùng phổ biến do còn khá mới và các doanh nghiệp Web không tiện chuyển đổi. Vì vậy, giao thức HTTP phổ biến nhất vẫn đang là HTTP 1.1. HTTP 1.0 vẫn còn được ưa chuộng trong nhiều trong hệ thống Proxy và một số ứng dụng cũ (wget).

Đặc điểm nổi bật của HTTP là "phi trạng thái" (Stateless), có nghĩa là mỗi request được xem là một phiên giao dịch độc lập, không lưu giữ bất kỳ thông tin nào liên quan đến các request trước đó cũng như các phiên làm việc trước.

Phương thức truy vấn (Request Method) phổ dụng của HTTP gồm: GET, HEAD, POST, PUT, DELETE, TRACE, OPTIONS, CONNECT, PATCH. Ngoài

Method, URI (Địa chỉ định danh của tài nguyên), HTTP Version thì trong Request Header có một số trường thông dụng sau:

- **Accept**: loại nội dung có thể nhận được từ thông điệp response. Ví dụ: text/plain, text/html…
- **Accept-Encoding**: các kiểu nén được chấp nhận. Ví dụ: gzip, deflate, xz, exi…
- **Connection**: điều khiển kết nối. Ví dụ: keep-alive, Upgrade…
- **Cookie**: thông tin HTTP Cookie từ server.
- **User-Agent**: thông tin về user agent của người dùng.

Hai phương thức được sử dụng nhiều nhất trong HTTP request là GET và POST. Với GET, câu truy vấn sẽ được đính kèm vào đường dẫn của HTTP request. Ví dụ: /?username=”abc”&password=”def”. Một số đặc điểm của phương thức GET như sau.

- GET request có thể được cached, bookmark và lưu trong lịch sử của trình duyệt.
- GET request bị giới hạn về chiều dài vì chiều dài của URL là có hạn.
- GET request không nên dùng với dữ liệu quan trọng, chỉ dùng để nhận dữ liệu.

Ngược lại, với POST thì câu truy vấn sẽ được gửi trong phần message body của HTTP request, một số đặc điểm của POST như sau.

- POST không thể cached, bookmark hay lưu trong lịch sử trình duyệt.
- Độ dài của POST không bị giới hạn.

Hình 2.3. Request

Response Code

Response Code chỉ ra trạng thái của một Response khi phản hồi một request từ Web Client - Thành công hay thất bại? Có lỗi hay không? Lỗi ở đâu?

```
HTTP/1.1 200 OK
Connection: close
Date: Thu, 06 Aug 1998 12:00:15 GMT
Server: Apache/1.3.0 (Unix)
Last-Modified: Mon, 22 Jun 1998 …...
Content-Length: 6821
Content-Type: text/html

data data data data data ...
```

Hình 2.4. Response

Các thông báo lỗi cụ thể như sau.

- 1xx - Informational
- 2xx - Success (200 - OK, 202 - Accepted, 204 - No Content)
- 3xx - Redirection (301 Moved Permanently: tài nguyên đã được chuyển hoàn toàn tới địa chỉ Location trong HTTP response. 303 See other: tài nguyên đã được chuyển tạm thời tới địa chỉ Location trong HTTP response. 304 Not Modified: tài nguyên không thay đổi từ lần cuối client request, nên client có thể sử dụng đã lưu trong cache.)

- 4xx - Client Error (400 - Bad Request, 401 - Unauthorized, 403 - Forbidden, 404 - Not Found, 405 - Method Not Allowed, 408 - Request Timeout, 429 - Too many requests)

5xx - Server Error (500 - Internal Server Error, 503 - Service Unavailable, 504 - Gateway Timeout, 509 - Bandwidth Limit Exceed)

2.2.3. Ghi nhật ký (Web log)

Web log chính là các tệp nhật ký tự động được tạo và duy trì bởi một máy chủ Web. Mỗi lần truy cập vào trang Web, bao gồm mỗi lần xem một tài liệu HTML, hình ảnh hoặc các đối tượng của website đều được web server ghi lại. Các máy chủ web như IIS, Apache hay Nginx đều có các web log file để ghi lại các nhật ký hoạt động của website.

Định dạng tệp nhật ký web thô chủ yếu là một dòng văn bản cho mỗi lần truy cập vào trang web. Điều này chứa thông tin về ai đã ghé thăm trang web, nơi họ truy cập và chính xác những gì họ đang làm trên trang web.

Với webserver có 2 dạng log file quan trọng:

- Log truy cập (access log) ghi lại những thông tin người dùng truy cập vào website.

- Log lỗi (error log) ghi lại các cảnh báo các lỗi xảy ra với dịch vụ liên quan web server.

- **Access_log:**

Tất cả những yêu cầu được xử lý bởi server được ghi lại ở Access_log. CustomLog directive điều khiển vị trí và nội dung của access log. Có thể định dạng nội dung của tập tin access_log bằng cách dùng LogFormat directive. LogFormat chứa những thông tin server cần theo dõi để ghi lại trong access log.

- **Error_log:**

Error_log là nơi mà httpd sẽ gửi những thông tin nhận dạng và bất kỳ lỗi nào gặp phải trong quá trình xử lý những yêu cầu. Tập tin này là nơi cần xem xét đầu tiên khi

gặp những lỗi khởi động httpd hay những thao tác của server, vì nó lưu những thông tin chi tiết về lỗi và cách sửa lỗi. Định dạng của tập tin error_log không bị bó buộc.

- **Định dạng các tệp log:**

+ Đối với Access_log:

Để thay đổi định dạng access log, có thể dùng directive log_format. Chỉ thị này mặc định nằm trong block http {…}. Để xem được log format ví dụ như của Nginx, có thể vào chỉ dẫn thư mục /etc/nginx/nginx.conf (đối với hệ điều hành CentOS và Ubuntu).

Mẫu log_format mặc định trong Nginx là access combined như sau:

```
log_format combined '$remote_addr - $remote_user [$time_local] '
    '"$request" $status $body_bytes_sent '
    '"$http_referer" "$http_user_agent"';
```

Chỉ thị log_format để định dạng access log cho Nginx, ý nghĩa các biến như sau:

- ***$remote_addr*** địa chỉ IP truy cập website của người dùng.
- ***$remote_user*** ghi lại tài khoản truy cập web nếu trang web có xác thực người dùng.
- ***$time_local*** thời gian người dùng truy cập.
- ***$request*** đoạn đầu của request.
- ***$status*** trạng thái của response.
- ***$body_bytes_sent*** kích thước body mà server response.
- ***$http_referer*** URL được tham chiếu.
- ***$http_user_agent*** thông tin trình duyệt, hệ điều hành mà người dùng truy cập.

Dưới đây là ví dụ về một bản ghi log trong Access_log (sample log):

```
117.65.256.123 - - [/Dec/2020:06:02:11 +0700] "GET /wp-content/plugins/wp-tab-widget-pro/css/font-awesome.min.css HTTP/1.1" 200 7027 "https://portal.ptit.edu.vn/gioi-thieu/.html" "Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/47.0.2526.106 Safari/537.36"
```

+ Đối với Error_log:

Log này để ghi lại thông tin các lỗi cài đặt cấu hình hay đơn giản chỉ là những cảnh báo giữa Web Server và các dịch vụ của nó.

Ví dụ một bản ghi trong file error_log trên máy chủ Web Apache như sau:

[Fri Jan 08 17:12:44 2020] [error] [client 127.0.0.1] client denied by server configuration: /export/home/live/ap/htdocs/test

Cột đầu tiên chỉ ra ngày giờ truy cập này được tạo ra. Cột thứ 2 chỉ ra đây là truy cập lỗi. Cột thứ 3 hiển thị địa chỉ IP của client tạo ra lỗi. Tiếp theo là message có nội dung chỉ ra rằng việc truy cập của client bị từ chối vì server được cấu hình như vậy. Tiếp đến là đường dẫn của document mà client cần truy cập.

Trong máy chủ Web Nginx, thực hiện ghi error_log bằng cách thêm chỉ thị error_log vào block http {…} như sau.

```
server {
error_log /var/log/nginx/error.log error;
...
}
```

Cú pháp chung của error log là:

```
error_log log_file log_level
```

Trong đó:

- ***emerg*** log ở level này mang tính khẩn cấp, dạng như server gần như đã sập.
- ***alert*** cảnh báo các vấn cần cần được xử lý ngay.
- ***crit*** các vấn đề quan trong nhưng không nhất thiết phải xử lý ngay lập tức, để theo dõi thêm.
- ***error*** ghi lại thông tin lỗi như đăng nhập hoặc cấu hình sai, mức độ thấp hơn crit.
- ***warn*** ở mức độ cảnh báo không phải lỗi.
- ***notice*** để thông báo cái gì đó.
- ***info*** ghi thông tin hệ thống, không có gì cả.

- ***debug*** ghi lại tất cả mọi thứ, dùng để dò lỗi.

Ví dụ các trường trong một file log trên máy chủ IIS như sau:

Date , Time , ClientIP , UserName , ServerIP , Method , UriStem , UriQuery , TimeTaken , Htt pStatus , Win32Status , ServerPort , UserAgent , HttpSubStatus , and Referer

2.3. Phương pháp phân tích dựa trên kiểm thử

Kiểm thử lỗ hổng bảo mật đối với ứng dụng web chính là việc làm thế nào để có thể chỉ ra những lỗ hổng đang tồn tại trên hệ thống một cách đầy đủ và khoa học nhất. Công việc này là một công việc rất khó khăn. Chính vì vậy, người ta đã tìm cách đưa ra các phương pháp kỹ thuật kiểm thử để nhằm đơn giản hoá công việc này, đồng thời đảm bảo đủ tin cậy rằng hệ thống sau khi được kiểm thử sẽ có được một báo cáo đầy đủ và chính xác nhất có thể.

Các phương pháp kiểm thử lỗ hổng bảo mật phổ biến hiện nay là: kiểm thử hộp đen, hộp trắng và hộp xám. Các phương pháp kiểm thử trên đều có những ưu, nhược điểm và đều có thể áp dụng cho các lỗ hổng bảo mật máy chủ Web.

- **Phương pháp kiểm thử hộp đen**

Kiểm thử hộp đen *(Black Box Testing*) dựa trên đầu vào và đầu ra của hệ thống để kiểm thử. Phương pháp này không cần biết mã nguồn bên trong như nào.

Với phương pháp kiểm thử hộp đen, các lỗ hổng bảo mật trên ứng dụng web thực hiện kiểm thử các ứng dụng từ bên ngoài, phía giao diện người dùng. Tức là quan sát các dữ liệu được chuyển đến ứng dụng và các dữ liệu từ ứng dụng xuất ra mà không biết mã nguồn hay hệ thống bên trong. Quá trình chuyển dữ liệu từ bên ngoài đến ứng dụng có thể thực hiện bằng cách sử dụng công cụ tự động hoặc bằng phương pháp thủ công

- **Phương pháp kiểm thử hộp trắng**

Kiểm thử hộp trắng (White Box Testing) dựa vào thuật toán, cấu trúc mã nguồn của chương trình. Mục đích của phương pháp này là để chắc chắn rằng tất cả các câu lệnh và điều kiện sẽ được thực hiện ít nhất một lần.

Việc kiểm thử các lỗ hổng bảo mật trên ứng dụng web được thực hiện thông qua việc phân tích, tìm kiếm lỗi trực tiếp trên mã nguồn của ứng dụng. Phương pháp này thường được thực hiện bởi nhà phát triển. Quá trình xác định lỗ hổng dựa trên mã nguồn có thể được thực hiện thủ công hoặc bằng công cụ tự động. Việc thực hiện thủ công với số lượng lớn các dòng lệnh có cấu trúc phức tạp sẽ gặp rất nhiều khó khăn. Do đó, rất cần phải có công cụ hỗ trợ cho kiểm thử viên để phân loại, tiếp cận nhanh chóng những điểm mà ứng dụng có khả năng bị lỗi. Khi kiểm thử, các công cụ sẽ tiến hành quét toàn bộ mã nguồn của ứng dụng và dựa trên tập nhận biết các hàm, các chỉ dẫn có khả năng gây ra lỗi bởi ngôn ngữ lập trình phát triển ứng dụng web. Một công cụ miễn phí được sử dụng rất phổ biến để quét mã nguồn là AppCodeScan do Blueinjy Solutions Pvt. Ltd. phát triển.

- **Phương pháp kiểm thử hộp xám**

Kiểm thử hộp xám (Grey Box Testing) được sử dụng để kiểm thử khi những thông tin được biết bên trong hệ thống mang tính hạn chế. Thực chất đây là phương pháp kết hợp giữa kiểm thử Black Box Testing và White Box Testing. Trong phương pháp này, kiểm thử viên có thể được xem tài liệu thiết kế, truy cập CSDL. Với những thông tin có được, kiểm thử viên có thể có kịch bản kiểm thử tốt hơn khi lên kế hoạch kiểm thử. Việc kiểm thử có thể được tiến hành với vai trò người dùng cuối hoặc nhà phát triển phần mềm.

- **Phân tích, so sánh các phương pháp kiểm thử lỗ hổng**

Ba phương pháp kiểm thử lỗ hổng bảo mật như đã trình bày trong phần trên nhìn chung đều có ưu và nhược điểm riêng và đều có thể áp dụng trong việc kiểm thử các lỗ hổng bảo mật máy chủ Web.

Việc lựa chọn phương pháp kiểm thử phụ thuộc vào điều kiện thực tế. Chẳng hạn, một tổ chức tự phát triển một ứng dụng web từ khâu thiết kế đến việc lập trình, các chuyên gia có đủ năng lực (lập trình an toàn) thì kiểm thử bằng White Box Testing là

phù hợp. Thực tế, phần lớn các website chỉ được xây dựng để đảm bảo các chức năng cho mục đích sử dụng và chưa qua bước kiểm thử tính bảo mật trước khi đưa vào sử dụng. Hơn nữa các lập trình viên thường không có kỹ năng lập trình an toàn, do đó ngay cả những nhà phát triển cũng không có đủ năng lực để kiểm thử về các điểm yếu an toàn.

Phương pháp kiểm thử hộp xám cũng có những lợi thế nhất định. Tuy nhiên, chính những lợi thế đó đôi khi dẫn đến những sai sót trong khi kiểm thử. Vì người kiểm thử sẽ không đảm bảo tính khách quan, rất khó có thể nhìn nhận trên quan điểm của kẻ tấn công.

Đối với kiểm thử hộp đen, người kiểm thử hoàn toàn đứng trên quan điểm kẻ tấn công, đây là một yêu cầu rất quan trọng trong quá trình kiểm thử, vì mục tiêu của việc kiểm thử là tìm ra những điểm yếu mà từ đó kẻ tấn công có thể xâm nhập vào hệ thống. Mặt khác, việc chi phí về thời gian cũng như về tài chính sẽ nằm trong phạm vi cho phép đối với nhiều tổ chức. Đối với những lỗ hổng điển hình, các công cụ ngày nay có thể xác định chính xác đến 100%.

Như đã phân tích ở trên, kiểm thử theo phương pháp hộp đen là phù hợp nhất trong điều kiện thực tế hiện nay. Phương pháp này sẽ được chọn đi sâu trong luận văn và sẽ được trình bày chi tiết hơn ở phần dưới.

2.4. Phương pháp phân tích truy cập bất thường dựa vào nhật ký đã ghi

Các truy cập bất thường vào một máy chủ Web có nguy cơ là một cuộc tấn công, do vậy việc phát hiện truy cập bất thường vào một máy chủ Web có vai trò quan trọng trong việc phát hiện sớm tấn công vào máy chủ Web [1, 7].

Một truy cập bất thường được định nghĩa là một hành vi "khác biệt" so với các hành vi truy cập bình thường khác [1]. Theo định nghĩa này, tất cả truy cập bình thường vào một máy chủ Web được định nghĩa từ trước thông qua các chính sách truy cập được thiết lập trong cấu hình máy chủ. Những chính sách này quy định cấu hình máy chủ, cấu trúc thư mục và các tệp, các dịch vụ có thể cung cấp bởi hệ thống, các giao thức có thể sử dụng, các cổng kết nối được mở cho từng loại dịch vụ tương ứng, các quyền truy cập

được cung cấp cho từng phân lớp người dùng và người quản trị. Mọi truy cập vi phạm các chính sách trên được gọi là truy cập bất thường vào máy chủ Web. Định nghĩa này sẽ xuyên suốt toàn bài luận văn trong quá trình xây dựng và thử nghiệm phát hiện truy cập bất thường thông qua nhật ký hệ thống máy chủ. [6]

Ví dụ về truy cập bất thường vào máy chủ Web như sau.

Ví dụ 1: Máy chủ Web cung cấp quyền quản trị các thư mục và tệp. Người dùng bên ngoài chỉ được quyền truy cập đến thư mục cấp quyền Public (ví dụ tệp index.html) và được truy cập qua giao thức HTTP, cổng 80. Khi người dùng cố tình vi phạm chính sách, truy cập sâu hơn vào thư mục khác của hệ thống hoặc tìm cách truy cập qua một cổng khác cổng 80, máy chủ Web sẽ thông báo lỗi và ghi hành vi bất thường này vào nhật ký hệ thống (Weblog).

Ví dụ 2: Máy chủ Web ghi nhận một hành vi nhập các dữ liệu bất thường vào phần đăng nhập, ví dụ như quá số lượng ký tự cho phép. Hành vi bất thường này có thể là dấu hiệu tấn công chèn mã (SQL Injection).

Ví dụ 3: Máy chủ Web có thể ghi nhận hàng loạt truy cập vào các địa chỉ và các cổng khác nhau của hệ thống. Đây là dấu hiệu rà quét hệ thống để khai thác thông tin, là giai đoạn bắt đầu của một cuộc tấn công footprinting.

Ví dụ 4: Máy chủ Web ghi nhận hiện tượng một đoạn mã script được cài vào bản tin POST/ Request gửi đến từ trình duyệt người dùng. Đây là rất có thể là một cuộc tấn công mã độc.

Theo nguyên tắc, có hai cách để phát hiện truy cập bất thường vào máy chủ Web. Phương pháp truyền thống là sử dụng hệ thống phát hiện xâm nhập (IDS – Intrusion Detection Systems). Theo cách này, hệ thống IDS được cấu hình với một tập luật (tập dấu hiệu – còn gọi là Signature) hỗ trợ cho việc phát hiện xâm nhập trái phép [2, 4]. Tuy nhiên, hạn chế của phương pháp này là phải biết các dấu hiệu tấn công từ trước. Mặt khác, với sự gia tăng của các tấn công mới, tập dấu hiệu sẽ phải cập nhật liên tục theo thời gian.

Cách thứ hai là phát hiện các hành vi bất thường, nghĩa là các hành vi vi phạm chính sách an ninh của hệ thống. Hệ thống phát hiện có thể được cấu hình theo một số mẫu có sẵn và tự cập nhật theo quá trình phát hiện các hành vi vi phạm mới. Một cách khác, hệ thống có thể duy trì việc phát hiện trên cơ sở phát hiện các vi phạm đối với các chính sách an ninh đã đặt trước. Theo cách này, hệ thống phát hiện có thể sử dụng các tập nhật ký ghi lại lỗi truy cập do máy chủ Web ghi liên tục theo thời gian và thực hiện phân tích để phát hiện các truy cập bất thường. Những hành vi bất thường là dấu hiệu tiềm ẩn của một tấn công máy chủ Web.

2.5 Kết luận chương

Trong chương 2, luận văn đã trình bày về một số nền tảng Apache, IIS, Nginx. Tiếp đó, bài đã trình bày về phương pháp ghi nhật ký máy chủ Web, Phương pháp phân tích dựa trên kiểm thử, Phương pháp phân tích truy cập bất thường dựa vào nhật ký đã ghi.

CHƯƠNG 3:
PHÁT HIỆN TRUY CẬP BẤT THƯỜNG VÀO MÁY CHỦ WEB

3.1. Quy trình và nguyên tắc phát hiện bất thường truy cập web

3.1.1. Phạm vi phân tích, phát hiện truy cập bất thường vào máy chủ Web

Ta cần thu thập dữ liệu từ Weblog để thực hiện nhiệm vụ phát hiện truy cập bất thường vào máy chủ Web. Về mặt nguyên tắc, có thể thu thập được các nhật ký (access log, error log, v.v… gọi chung là Weblog) từ máy chủ Web, trang tin, mạng nói chung và các log truy cập các dịch vụ mạng. Tuy nhiên, trong khuôn khổ, luận văn chỉ tập trung vào việc thu thập và phân tích Weblog để phát hiện truy cập bất thường vào máy chủ Web.

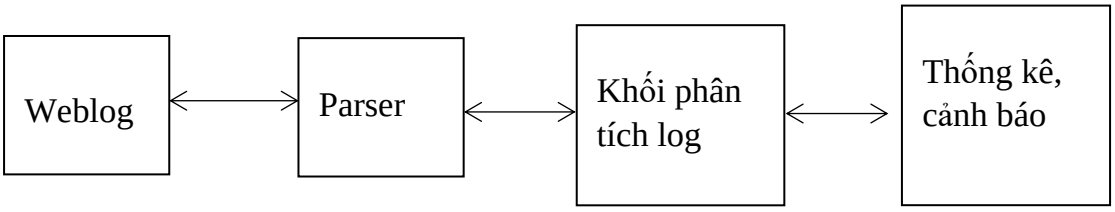
Qua quá trình xử lý, phân tích dữ liệu log thô thu thập được, ta có thể trích xuất được các thông tin quan trọng về dấu hiệu, khả năng xuất hiện của các hành vi truy cập bất thường. Từ kết quả phân tích, ta phát hiện dấu hiệu truy cập bất thường có thể giúp nhận biết các loại mã độc, tấn công, xâm nhập vào hệ thống máy chủ Web.

Hiện có rất nhiều công cụ, phần mềm và hệ thống phục vụ cho việc thu thập, xử lý, phân tích và phát hiện dấu hiệu truy cập bất thường vào máy chủ Web. Trong khuôn khổ của bài, luận văn không đi vào chi tiết trình bày các công cụ phần mềm và hệ thống đã có, mà chỉ trình bày khái quát một số phương pháp, công cụ điển hình. Trên cơ sở đó, phần tiếp theo của chương 2 sẽ trình bày theo các nội dung sau: kiến trúc hệ thống, cấu trúc và định dạng của Weblog, phương pháp thu thập dữ liệu Weblog, phương pháp trích chọn mẫu và đặc trưng, phương pháp phân tích phát hiện bất thường.

3.1.2. Quy trình và nguyên tắc phát hiện

Để quản trị một Website hiệu quả, tránh nguy cơ xảy ra các cuộc tấn công vào máy chủ Web. Máy chủ web ghi log, căn cứ log có thể phân tích, xử lý, thống kê, cảnh báo.

Sơ đồ thiết kế hệ thống phân tích, phát hiện truy cập bất thường vào máy chủ Web được mô tả như sau:



Weblog: Khối này đặc tả ghi weblog, ghi lại thông tin về các sự kiện xảy ra trong truy cập máy chủ, bao gồm các sự kiện truy cập bất thường.

Parse: Là khối xử lý sơ bộ, định dạng Weblog và truyền về trung tâm phân tích.

Khối phân tích Log: Phân tích dấu hiệu bất thường của Weblog

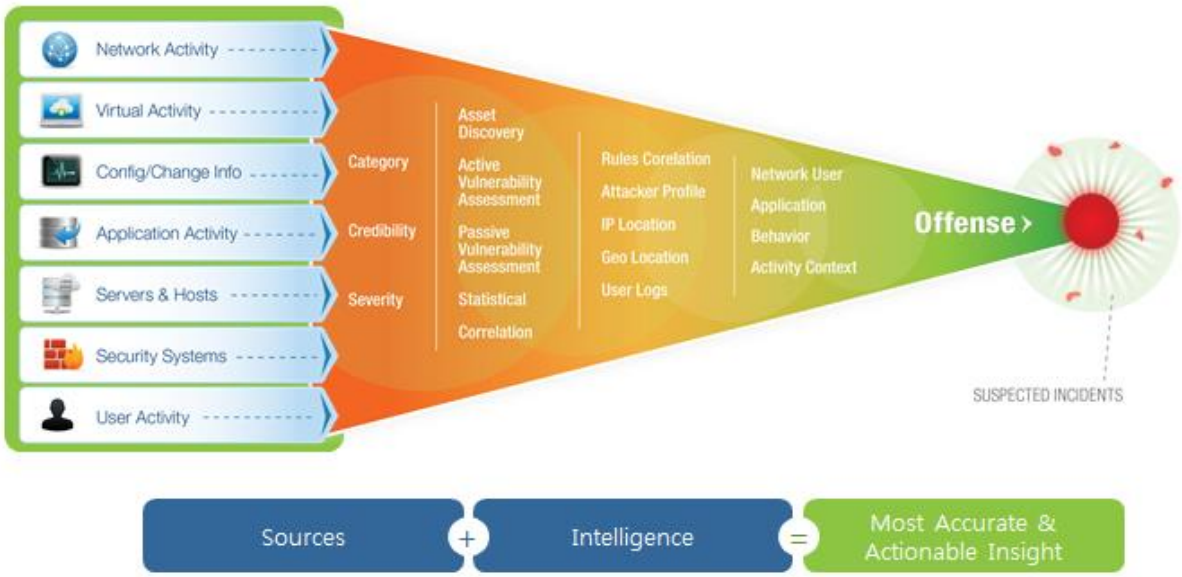
Khối thống kê, cảnh báo: Đưa ra thống kê, cảnh báo: Sau khi đã phân tích filelog đưa ra thống kê các truy cập bất thường bằng địa chỉ IP…từ đó cảnh báo tấn công máy chủ web.

3.1.3. Tham khảo một số mô hình kiến trúc hệ thống

Hiện nay, có nhiều nền tảng và công cụ xử lý, phân tích log truy cập thương mại cũng như mã mở được cung cấp như IBM QRadar SIEM, Splunk, Sumo Logic, VNCS Web Monitoring, Logstash, Graylog, LOGalyze, Webalizer... Để tham khảo các mô hình kiến trúc này, phần sau đây khảo sát chi tiết tính năng, các ưu nhược điểm của các nền tảng và công cụ kể trên. Các tiêu chí quan trọng được xem xét bao gồm: 1) Khả năng thu thập, xử lý các dạng log truy cập từ nhiều nguồn; 2) Khả năng phát hiện và cảnh báo các truy cập bất thường và các dạng tấn công, xâm nhập vào hệ thống Web; 3) Khả năng quản lý, lưu trữ, tìm kiếm log và tạo các dạng báo cáo, thống kê [7].

- **IBM QRadar SIEM**

QRadar SIEM (Security Information and Event Management) là hệ thống quản lý các thông tin và sự cố an ninh được phát triển và cung cấp bởi hãng IBM.



Hình 3.1. Nguyên lý hoạt động của IBM QRadar SIEM

QRadar SIEM có các tính năng tiêu biểu như sau:

+ Khả năng phát hiện giả mạo, các nguy cơ bên trong và bên ngoài;

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

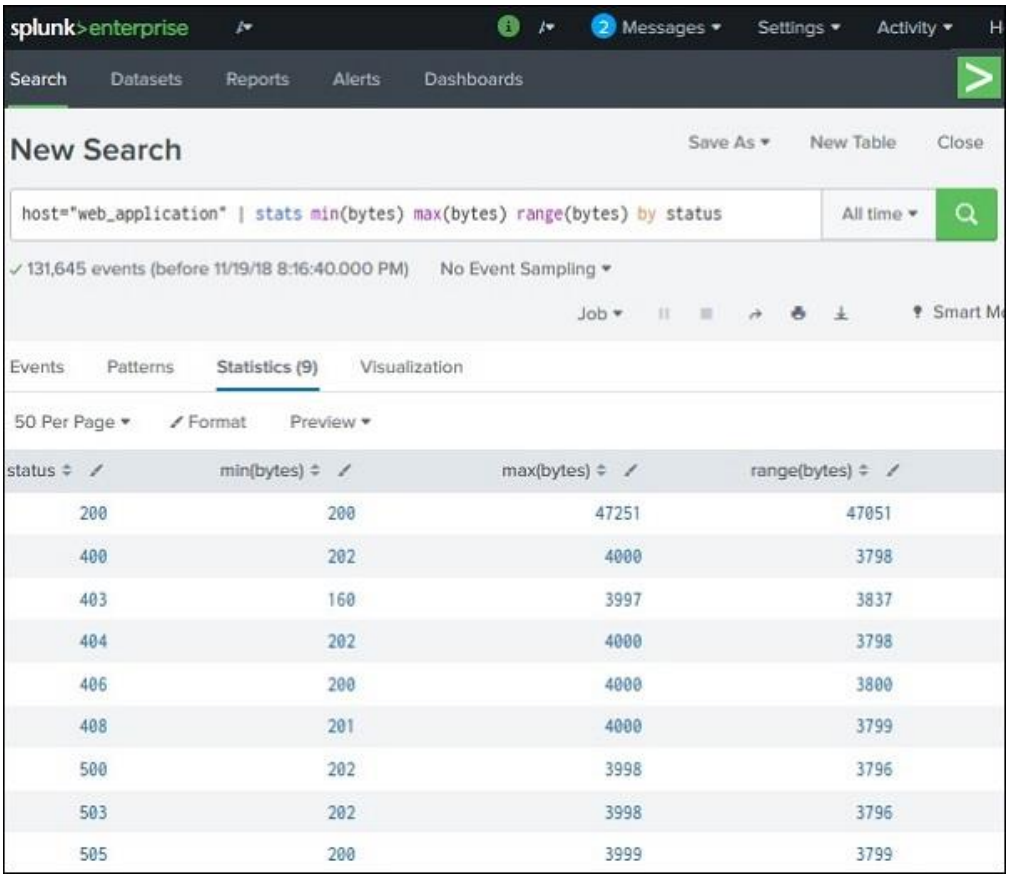
+ Khả năng theo dõi và liên kết các sự cố và nguy cơ;

+ Có thể dễ dàng mở rộng tính năng lưu trữ, xử lý.

Tuy nhiên, QRadar SIEM có hạn chế lớn 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 [9].

- **Splunk**

Splunk là một nền tảng xử lý và phân tích log rất mạnh, được cung cấp bởi hãng Splunk Inc., Hoa Kỳ. Splunk có hàng trăm công cụ tích hợp, cho phép xử lý nhiều loại log khác nhau với khối lượng lớn theo thời gian thực. Để phục vụ đảm bảo an toàn thông tin, Splunk có thể xử lý, phân tích log, cũng như trích rút thông tin hỗ trợ cho các hoạt động kinh doanh. Splunk cung cấp các công cụ tìm kiếm và biểu đồ cho phép biểu diễn kết qua đầu ra theo nhiều dạng. Hình 3.2 hiển thị màn hình thống kê của Splunk



Hình 3.2. Thống kê của Splunk

Chi phí cài đặt lớn là hạn chế lớn nhất của Splunk, do khoản đầu tư ban đầu cho hệ thống thiết bị chuyên dụng có độ phức tạp cao. Một vấn đề khác là phí bản quyền hàng năm của Splunk cũng rất đắt đỏ (ước tính có thể lên đến hàng chục ngàn đô-la Mỹ mỗi năm), nên Splunk cũng 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 [9].

- **Sumo Logic**

Sumo Logic là một dịch vụ phân tích, xử lý và quản lý log trên nền tảng điện toán đám mây. Sumo Logic có ưu điểm là cung cấp nhiều tính năng và có khả năng xử lý nhiều loại log, đồng thời việc cài đặt cũng tương đối dễ dàng do Sumo Logic dựa trên nền tảng điện toán đám mây, không đòi hỏi thiết bị chuyên dụng. Log được thu thập từ

hệ thống của khách hàng sử dụng các Agent/Collector và được tải lên hệ thống xử lý và phân tích của Sumo Logic. Nhược điểm lớn nhất của Sumo Logic là việc phải tải khối lượng lớn log (có thể lên đến hàng chục GB/ngày) từ hệ thống sinh log lên hệ thống dịch vụ Sumo Logic để xử lý. Việc này đòi hỏi chi phí lớn cho đường truyền, có thể gây ra chậm trễ trong quá trình xử lý và tiềm ẩn nguy cơ rò rỉ dữ liệu nhạy cảm chứa trong log [9].

- **Hệ thống giám sát Web của VNCS**

VNCS Web monitoring là giải pháp cho phép giám sát nhiều Website đồng thời dựa trên thu thập, xử lý và phân tích log truy cập sử dụng nền tảng Splunk do Công ty cổ phần Công nghệ An ninh không gian mạng Việt Nam phát triển.

Hình 2.3 là một màn hình thống kê của VNCS Web monitoring VNCS Web monitoring thu thập Web log từ các máy chủ cần giám sát, sau đó chuyển về hệ thống trung tâm để xử lý, phân tích. Hệ thống này cho phép quản lý log tập trung, hỗ trợ phân tích log thủ công để tìm sự cố, hỗ trợ giám sát và cảnh báo trạng thái hoạt động của Website, hỗ trợ phát hiện các dạng tấn công thay đổi nội dung, thay đổi giao diện, tấn công chèn mã SQL (SQL Injection - SQLi), tấn công chèn mã script liên miền (Cross Site Scripting - XSS) và phát hiện mã độc trên Website. Hạn chế của VNCS Web monitoring là chỉ có khả năng xử lý và phân tích Web log [9].

Hình 3.3. Thống kê của VNCS Web monitoring

3.2 Thu thập thông tin nhật ký web cho phát hiện bất thường

3.2.1. Cấu trúc Weblog

Tệp nhật ký Weblog có định dạng chuẩn CLF (Common Log File), chứa các dòng thông điệp cho mỗi một gói HTTP request, cấu tạo như sau:

Host Ident Authuser Date Request Status Bytes

Trong đó:

- Host: Tên miền đầy đủ của client hoặc IP
- Ident: Nếu chỉ thị IdentityCheck được kích hoạt và client chạy identd, thì đây là thông tin nhận dạng được client báo cáo
- Authuser: Nếu URL yêu cầu xác thực HTTP thì tên người dùng là giá trị của mã thông báo này
- Date: Ngày và giờ thực hiện yêu cầu
- Request: Dòng yêu cầu của client, được đặt trong dấu ngoặc kép (“”)
- Status: Mã trạng thái (gồm ba chữ số)
- Bytes: số bytes trong đối tượng trả về cho client, ngoại trừ các HTTP header
- Mỗi yêu cầu có thể chứa các các dữ liệu bổ sung như đường liên kết hoặc chuỗi

ký tự của người dùng.

Nếu mã thông báo không có giá trị, thì mã thông báo được biểu thị bằng một dấu gạch ngang (-). Ví dụ:

192.168.0.111 - uche [21/Dec/2020:12:30:45 +0700] "GET/index.html HTTP/1.1" 200

Giải thích chi tiết các trường:

Bảng 3.1. Giải thích chi tiết các trường trong Weblog

Tên trường	Giá trị mẫu	Mô tả
host	192.168.0.111	Địa chỉ IP hoặc tên máy chủ của ứng dụng khách HTTP đã đưa ra yêu cầu
identd	-	Mã định danh giao thức máy chủ xác thực (RFC 931) cho máy khách; trường này hiếm khi được sử dụng. Nếu không sử dụng nó được cho là "-".
username	uche	Tên người dùng được xác thực HTTP (qua bắt tay phản hồi 401); đây là hộp thoại đăng nhập và mật khẩu người dùng thấy trên một số trang web, trái ngược với biểu mẫu đăng nhập được nhúng trong trang Web, nơi thông tin ID của người dùng được lưu trữ trong phiên phía máy chủ. Nếu không

Tên trường	Giá trị mẫu	Mô tả
		được sử dụng (ví dụ, khi yêu cầu cho một nguồn tài nguyên không hạn chế) nó được cho là "-".
date/time	[21/Dec/2020:08:53:33 +0700]	Ngày thì giờ thì múi giờ, theo định dạng [dd / MMM / yyyy: hh: mm: ss + -hhmm]
request line	"GET /index.html HTTP/1.1"	Hàng đầu của yêu cầu HTTP, bao gồm phương thức ("GET"), tài nguyên được yêu cầu và phiên bản giao thức HTTP
status code	200	Mã số được sử dụng trong phản hồi để biểu thị yêu cầu của yêu cầu, ví dụ để chỉ ra thành công, thất bại, chuyển hướng hoặc yêu cầu xác thực
bytes		Số byte được chuyển trong phần thân của phản hồi

* Định dạng log kết hợp

1 192.168.0.111 - uche [21/Dec/2020:08:53:33 +0700] "GET /index.html 2HTTP/1.1" 200

3 2345

4 "http://www.google.com/" "Mozilla/5.0 (X11; U; Win64 x64; en-US; 5rv:1.9a8)

Gecko/2007100619

GranParadiso/3.0a8" "USERID=Anhminh;IMPID=12345"

Định dạng nhật ký kết hợp là định dạng chung cộng với ba trường bổ sung — liên kết giới thiệu, tác nhân người dùng và cookie.

Các trường bổ sung trong một dòng định dạng log được kết hợp

Bảng 3.2. Giải thích chi tiết các trường bổ sung

Tên trường	Giá trị mẫu	Mô tả
referrer	"http://www.google.com/"	Khi người dùng theo một liên kết từ trang này đến trang khác, trang này thường báo cáo đến trang web thứ hai mà URL đã giới thiệu.
user agent	"Mozilla / 5.0 (X11; U; Win64 x64; en-US; rv: 1.9a8) Gecko / 2007100619 GranParadiso / 3.0a8"	Chuỗi cung cấp thông tin về tác nhân người dùng đã thực hiện yêu cầu (ví dụ: phiên bản trình duyệt hoặc trình thu thập thông tin web)
cookie	"USERID = Anhminh; IMPID = 12345"	Cặp khóa / giá trị thực tế của bất kỳ cookie nào được gửi bởi máy chủ HTTP có thể gửi lại cho máy khách trong phản hồi.

3.2.2. Thu thập thông tin từ logfile hệ thống

Trong một hệ thống mạng lớn, người quản trị thường phải thu thập một lượng dữ liệu lớn như log thiết bị, hệ thống, các thông điệp cảnh báo, điều khiển được tạo ra trên

mạng lưới bởi các thiết bị hoặc ứng dụng. Những dữ liệu này thường phức tạp và đa dạng vì trong hệ thống có rất nhiều thiết bị tham gia vào. Các hệ điều hành khác nhau với các máy chủ cũng tạo ra một lượng lớn dữ liệu log. Ngoài ra còn có các log của những ứng dụng hoạt động trên hệ thống.

Một số loại logfile hệ thống sau:

- Firewall logs: Là nơi ghi lại trạng thái hoạt động của Firewall, các kết nối ra vào hệ thống, các hành động mà Firewall đã cho phép hoặc không đối với từng kết nối, thông tin về nguồn và đích kết nối,v.v…

- STB logs: Là các log mô tả lại những hoạt động của người dùng cuối như bật, tắt STB, yêu cầu nội dung, v.v…

- Router syslogs: Được tạo bởi các router, mô tả một loạt các sự kiện được router ghi lại.

Các log ứng dụng, đặc biệt là các log ứng dụng web, cho phép khai thác dữ liệu khi người dùng truy cập máy chủ web để thực hiện một số hành động như đăng ký gói dịch vụ truyền hình, đăng ký và sử dụng dịch vụ giá trị gia tăng, v.v…Các log ứng dụng khác có thể được sử dụng cho những yêu cầu cụ thể.

3.2.3. Thu thập thông tin từ công cụ

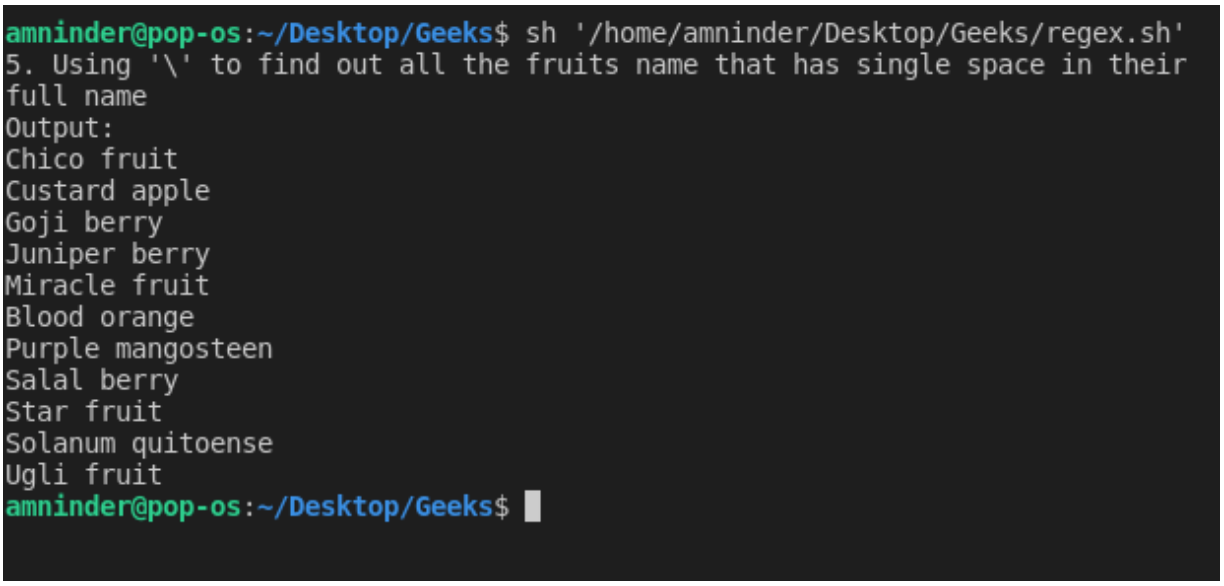
Các phương pháp phân tích tập tin nhật ký thủ công & phát hiện tấn công theo dấu hiệu luôn là các phương pháp hiệu quả về mặt kết quả, tuy nhiên sẽ mất rất nhiều thời gian và công sức để phân tích log file, vì log file thường chứa rất nhiều dòng nhật ký. Vì vậy Regular expression là lựa chọn phù hợp.

Regular (Regex) cho phép xử lý các chuỗi ký tự linh hoạt, hiệu quả và mạnh mẽ. Regex cho phép mô tả và phân tích chuỗi ký tự với các bản mẫu tương tự như một ngôn ngữ lập trình nho nhỏ. Regex có trong nhiều dạng công cụ, nhưng sức mạnh của nó chỉ được thể hiện tối đa khi là 1 phần của một ngôn ngữ lập trình.

Dưới đây là ví dụ một đoạn code viết bằng python, sử dụng Regular Expression trong việc phân tích tập tin nhật ký web phát hiện tấn công XSS.

```
import os, sys, re
from collections import
Counter from subprocess
import call

PATH =
sys.argv[1]
TYPE =
sys.argv[2]

if TYPE == 'access':
        log = 'access.log'

elif TYPE == 'error':
        log = 'error.log'

f =
open(PATH+log,
'r') ipList = []
```

```
xss_match='(.+(POST\s+|GET\s+|HEAD\s+|PUT\s+|OPTION\s+).+?=.+?((S|
s)(C|c)(R|r)(I|i)(P|p)(T|t)|(S|
s)(E|e)(L|l)(F|f)|(A|a)(L|l)(E|e)(R|r)(T|t)).+?HTTP/[0-9]\.[0-9].+)'
time_regex = re.compile("([0-9]{2}:[0-9]{2}:[0-9]{2}\s+)")
date_regex =
re.compile("((\d{2}|\d{4})/(\d{2}|\w{3})/(\d{2}|\d{4}))(?:
\:|\s+)") ip_regex = "(\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})"
ip_regsearch = re.compile(ip_regex)
xss_payload_regex=re.compile("((POST\s+|GET\s+|HEAD\s+|PUT\s+|OPTIO
N\s+).+?=.+?((S|s)(C|c)(
R|r)(I|i)(P|p)(T|t)|(S|s)(E|e)(L|l)(F|f)|(A|a)(L|l)(E|e)(R|r)(T|t)|(J|j)(A|a)(V|v)(A|a)(
S|s)(C|c)(R|r)(I|i)(P|p)(T|t
)\:|(X|x)(S|S)(S|s)).+?HTTP/[0-9]\.[0-9].+)")

for line in f.read().split('\n'):

if re.match(xss_match, line):

dateData = date_regex.search(line)

timeData = time_regex.search(line)

ipData = ip_regsearch.search(line) if

re.match(xss_match, line):

payloadType = "XSS"

payloadData = xss_payload_regex.search(line)

print "["+payloadType+"] "+dateData.group(0)+" | "+timeData.group(0)+"|"+
ipData.group(0)+" | "+payloadData.group(0)
```



Hình 3.4. Kết quả sau khi ứng dụng Regex

3.3. Kết luận chương

Trong chương 3, luận văn đã trình bày về quy trình và nguyên tắc phát hiện bất thường truy cập Web và một số mô hình kiến trúc hệ thống. Tiếp đó, bài đã trình bày về cấu trúc Weblog, thu thập thông tin từ logfile và công cụ.

CHƯƠNG 4: THỬ NGHIỆM

4.1. Giới thiệu công cụ Weblog Expert

Có một số công cụ thu thập thông tin logflie hệ thống phổ biến mã nguồn mở như: FireStats, Open Web Analytics, Weblog Expert, Go Access, Web Forensik,…. Tuy nhiên trong bài luận văn này, tác giả lựa chọn phần mềm Weblog Expert.

WebLog Expert là một chương trình phân tích nhật ký truy cập nhanh và mạnh mẽ. Nó sẽ cung cấp thông tin về khách truy cập trang web: thống kê hoạt động, đường dẫn được truy cập, đường dẫn thông qua trang web, thông tin về trang, công cụ tìm kiếm, trình duyệt, hệ điều hành,.... Chương trình giúp tạo ra các báo cáo dễ đọc bao gồm thông tin dưới dạng văn bản (bảng) và biểu đồ.

4.2. Mô hình hệ thống máy chủ Web thử nghiệm

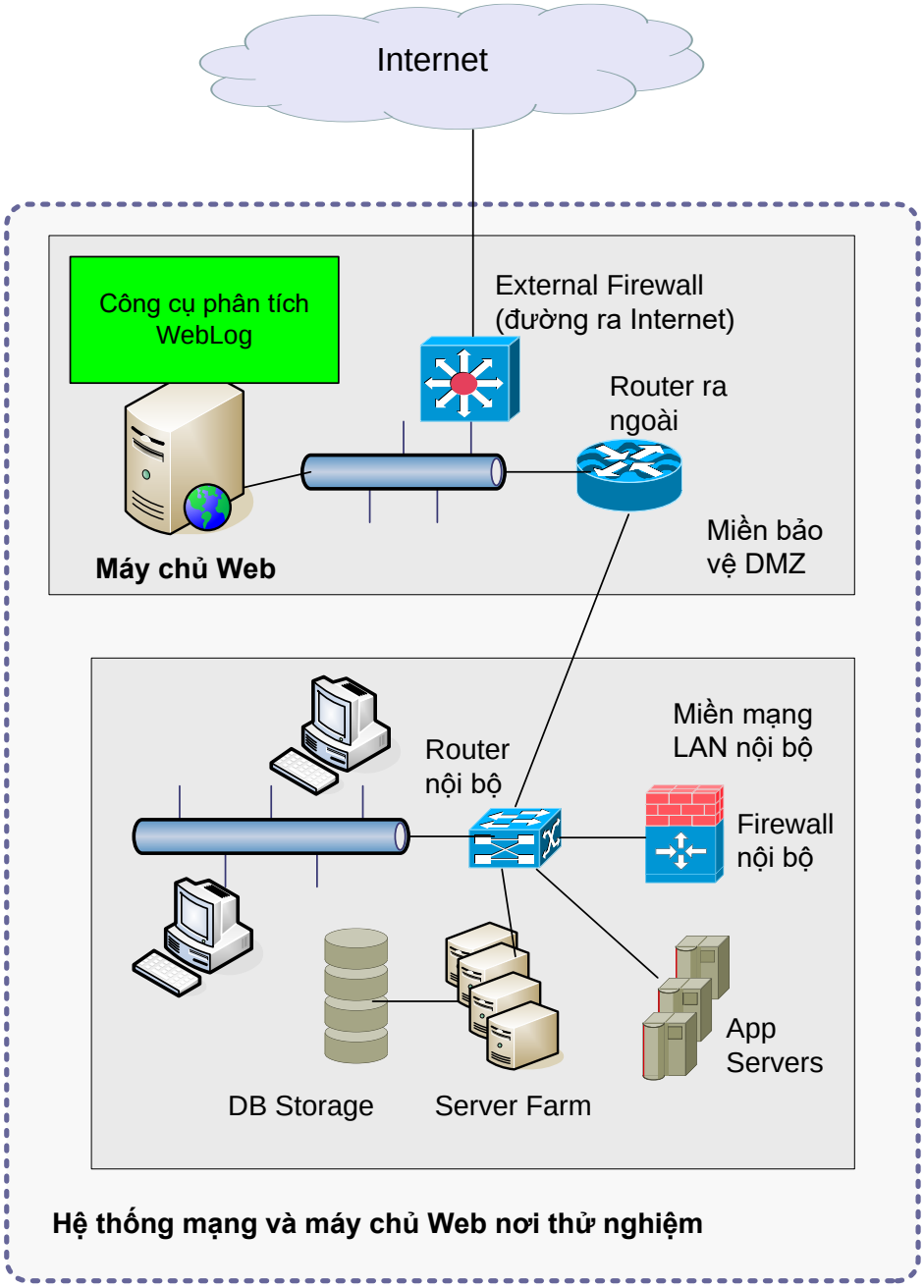
Qua khảo sát tại các hệ thống máy chủ Web cài đặt ở các cơ quan, có thể nhận thấy có một số đặc điểm chung như sau:

- Tuy mức độ đầu tư về hạ tầng CNTT khác nhau tại mỗi điểm phụ thuộc vào điều kiện, quy mô, phạm vi, song máy chủ Web thường được đặt trong phân vùng mạng có bảo vệ DMZ (Demilitarized Zone). Phân vùng mạng này cho phép mọi truy cập từ mạng Internet, nghĩa là từ bất kỳ máy tính nào. Đây là phân vùng mạng có nhiều nguy cơ tấn công.
- Các phân vùng mạng khác gồm phân vùng mạng trục tạo kết nối giữa các hệ thống lớn, mạng LAN của các phòng ban trực thuộc, phân vùng mạng quản trị hệ thống, phân vùng mạng riêng, phân vùng mạng lưu trữ nội bộ, v.v.
- Các máy chủ ứng dụng, máy chủ cơ sở dữ liệu, máy chủ Email, máy chủ dịch vụ công cũng thường được đặt trong miền mạng DMZ có bảo vệ. Một số máy chủ cơ sở dữ liệu và máy chủ DHCP được đặt trong phân vùng mạng nội bộ. Máy chủ quản trị hệ thống thường đặt trong phân vùng quản trị.

- Phân vùng mạng riêng thường dành cho những dịch vụ đặc biệt khác. Phân vùng này thường gồm các máy trạm nội bộ, phục vụ các hoạt động quản trị hành chính và thông tin nội bộ.

Luận văn tập trung vào máy chủ Web, là nơi sẽ tiến hành cài đặt công cụ thu thập và phân tích Weblog. Do vậy, luận văn chỉ tập trung vào phân vùng mạng DMZ, nơi đặt máy chủ Web và kết nối của phân vùng này với Internet. Phân vùng này có thể có một số thiết bị bảo vệ như tường lửa, IDS/IPS tùy vào mức độ đầu tư hạ tầng CNTT của các cơ quan tổ chức. Máy chủ Web có thể gồm cả tường lửa lớp ứng dụng (Web Application Firewal - WAF) và DNS server, nơi có thể thu thập dữ liệu giám sát hoạt động của máy chủ Web.

Từ những lý do trên, luận văn xây dựng mô hình hệ thống thử nghiệm với máy chủ Web như trên hình 4.1 như sau.



Hình 4.1. Mô hình thử nghiệm phân tích Weblog máy chủ Web

Máy chủ Web có cấu hình như sau:

- Phần cứng máy chủ: Dual Intel Xeon, 8 Core, 20 Threads, tốc độ 2 x 2.1 GHz, đĩa cứng 1.2 TB.

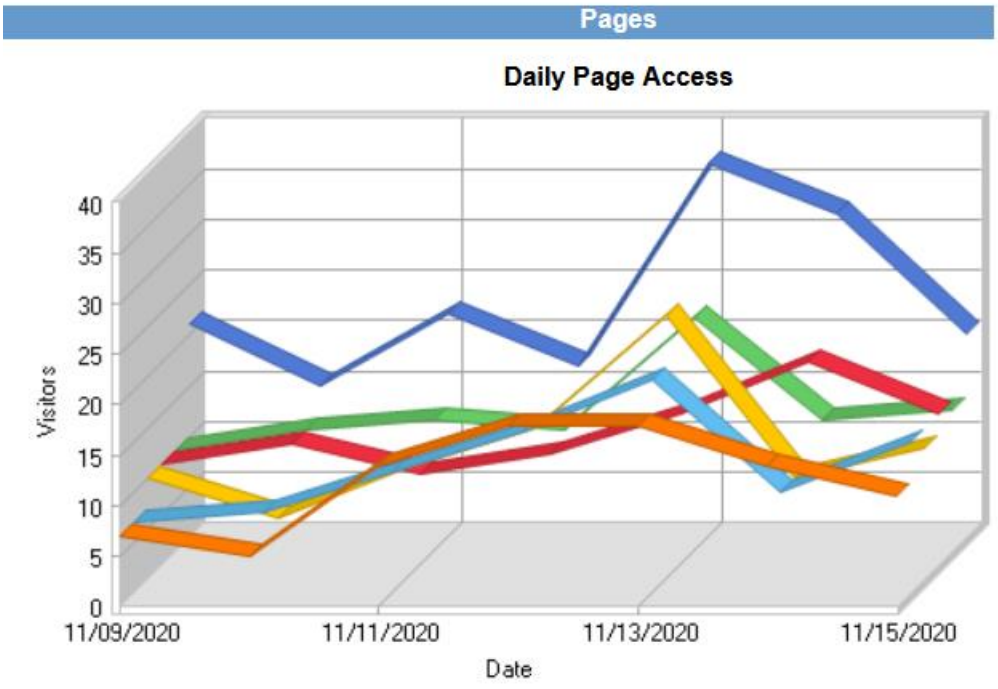
- Phần mềm: Hệ điều hành máy chủ là Microsoft Windows Server 2012 R2. Hệ thống WebServer được xây dựng trên nền tảng máy chủ Web Microsoft IIS, sử dụng MySQL.

Công cụ thu thập và phân tích Weblog được chọn thử nghiệm là:

- Thử nghiệm phân tích, phát hiện bất thường sử dụng công cụ Weblog Expert.

4.3. Thử nghiệm phân tích, phát hiện bất thường với công cụ Weblog Expert

Báo cáo mẫu Weblog Expert để nhận ý tưởng chung về nhiều thông tin khác nhau về việc sử dụng trang web mà nó có thể cung cấp:



Hình 4.2. Báo cáo các truy cập trang hàng ngày của Weblog Expert

Bảng 4.1. Thống kê báo cáo của Weblog Expert

Summary

Hits	
Total Hits	3250
Visitor Hits	3112
Spider Hits	88
Average Hits per Day	470
Average Hits per Visitor	6.54
Cached Requests	0
Failed Requests	15
Page Views	
Total Page Views	1022
Average Page Views per Day	145
Average Page Views per Visitor	1.98
Visitors	
Total Visitors	504
Average Visitors per Day	74
Total Unique Ips	426
Bandwidth	
Total Bandwidth	964.36 MB
Visitor Bandwidth	884.31 MB
Spider Bandwidth	18.51 MB
Average Bandwidth per Day	144.91 MB
Average Bandwidth per Hit	305.63 KB
Average Bandwidth per Visitor	2.08 MB

Trình phân tích nhật ký có thể tạo báo cáo ở định dạng HTML, PDF và CSV. Nó cũng bao gồm một máy chủ web hỗ trợ các báo cáo HTML động.

WebLog Expert có thể phân tích nhật ký của các máy chủ web Apache, IIS và Nginx. Nó có thể đọc các tệp nhật ký nén GZ và ZIP, do đó sẽ không cần giải nén chúng theo cách thủ công.

Trình thủ thuật tích hợp sẽ giúp ta nhanh chóng và dễ dàng tạo tiểu sử cho trang web và phân tích nó.

* Tính năng, đặc điểm của WebLog Expert

- Về báo cáo

+ Thống kê chung

+ Thống kê hoạt động: hàng ngày, theo giờ trong ngày, theo ngày trong tuần, theo tuần và theo tháng.

+ Truy cập thống kê: thống kê cho trang, tệp, hình ảnh, thư mục, truy vấn, thời gian xem, trang nhập, trang thoát, thư bị trả lại, đường dẫn qua trang web, loại tệp, miền ảo và máy chủ cân bằng tải.

+ Thông tin về khách truy cập: máy chủ, tên miền cấp cao nhất, quốc gia, tiểu bang, thành phố, người dùng được xác thực, độ phân giải màn hình, độ sâu màu và ngôn ngữ.

+ Liên kết giới thiệu: giới thiệu các trang web, URL, công cụ tìm kiếm (bao gồm thông tin về cụm từ tìm kiếm và từ khóa)

+ Trình duyệt, hệ điều hành, loại thiết bị và số liệu thống kê

+ Thông tin về lỗi: loại lỗi, thông tin lỗi chi tiết

+ Thống kê mục tiêu

+ Số liệu thống kê tệp đã theo dõi

+ Báo cáo lớp phủ nhấp

+ Hỗ trợ báo cáo tùy chỉnh

- Bộ lọc

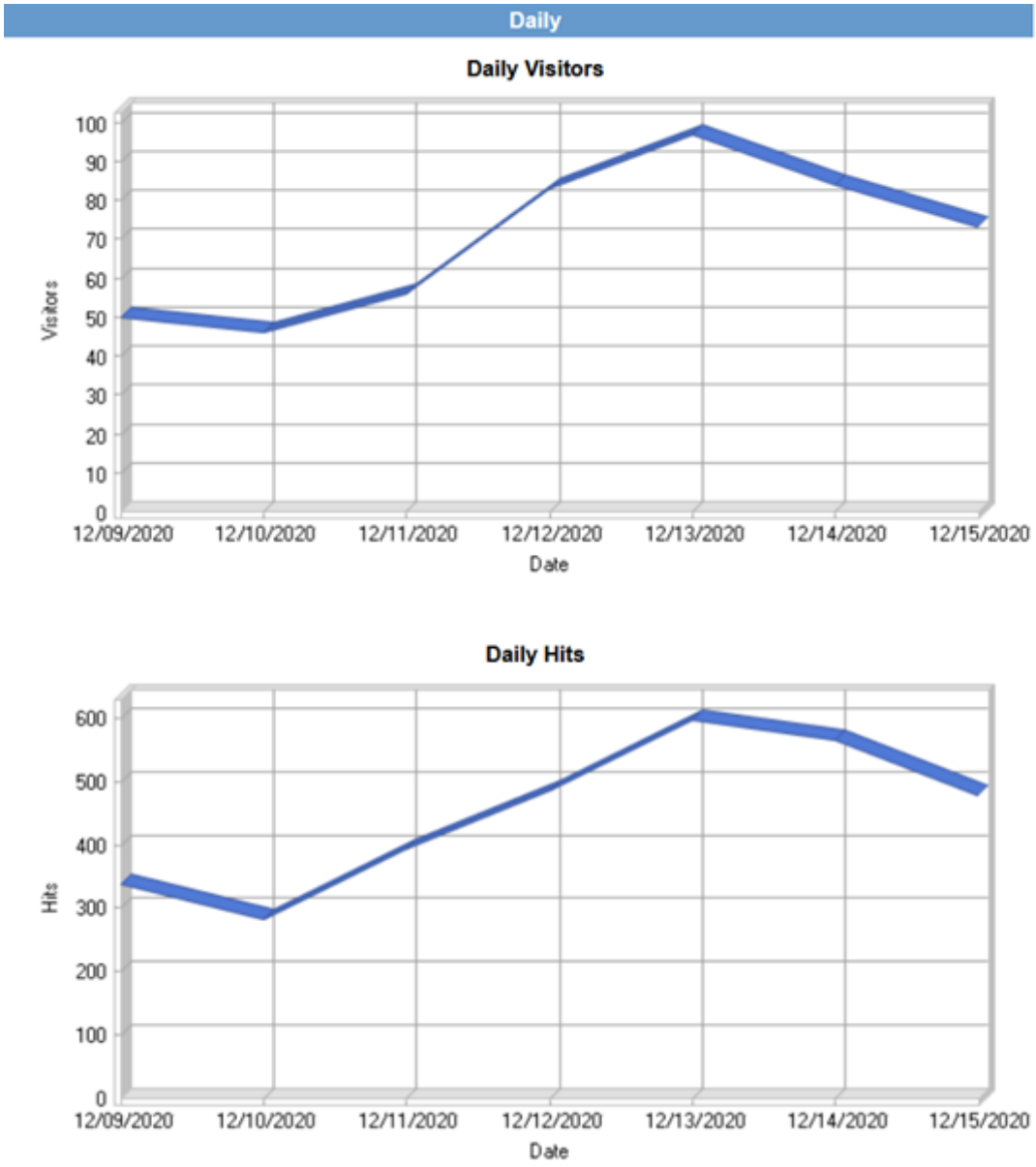
+ Bộ lọc (truy cập nhật ký): tệp, truy vấn, máy chủ, liên kết giới thiệu, mã trạng thái, phương thức, cổng, máy chủ, hệ điều hành, trình duyệt, loại thiết bị, spider, tác nhân người dùng, ngày trong tuần, giờ trong ngày, quốc gia, tiểu bang, thành phố, tổ chức, người dùng được xác thực, miền ảo, thời gian thực hiện

+ Bộ lọc khách truy cập: khách truy cập đã truy cập tệp cụ thể, khách truy cập có trang nhập cụ thể, khách truy cập có trang thoát cụ thể, khách truy cập đến từ URL giới thiệu cụ thể, khách truy cập đến từ một công cụ / cụm từ tìm kiếm cụ thể.

4.4. Một số kết quả thử nghiệm với Weblog Expert

Sau đây là một số kết quả thử nghiệm phát hiện bất thường với công cụ Weblog Expert.

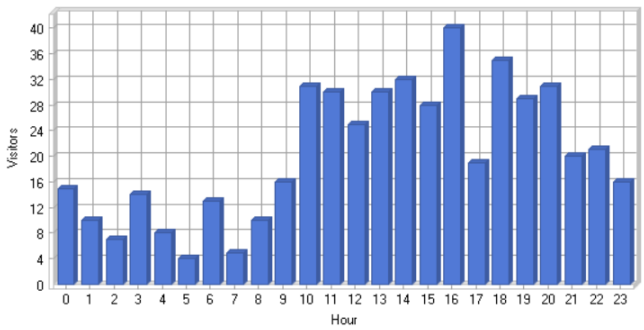
Hình 4.3 mô tả các truy cập vào Web theo từng ngày, trong đó lưu lượng truy cập PhP có độ tăng đột biến thể hiện một tấn công Web.



Hình 4.3. Mô tả truy cập Web theo từng ngày

By Hour of Day

Activity by Hour of Day



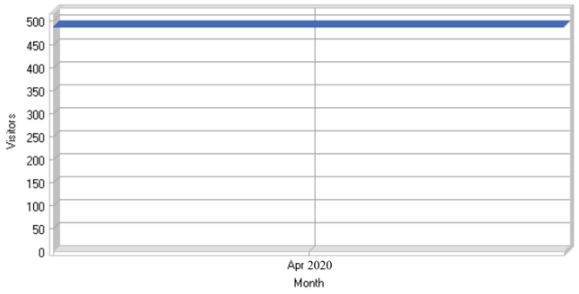
Activity by Hour of Day

Hour	Hits	Page Views	Visitors	Bandwidth (KB)
00:00 - 00:59	104	32	15	38,110
01:00 - 01:59	53	20	10	24,162
02:00 - 02:59	50	13	7	6,124
03:00 - 03:59	94	28	14	19,703
04:00 - 04:59	61	16	8	14,067
05:00 - 05:59	24	7	4	5,656
06:00 - 06:59	114	27	13	25,342
07:00 - 07:59	29	11	5	8,301
08:00 - 08:59	47	19	10	8,601
09:00 - 09:59	115	33	16	9,840
10:00 - 10:59	184	55	31	29,269
11:00 - 11:59	156	64	30	54,423

Hình 4.4. Các hành vi truy cập thống kê theo giờ trong ngày

By Month

Activity by Month



Activity by Month

Month	Hits	Page Views	Visitors	Bandwidth (KB)
Apr 2020	3,135	957	489	876,415
Total	**3,135**	**957**	**489**	**876,415**

Hình 4.5 Các hành vi truy cập theo tháng

4.5. Kết luận chương

Trong chương 4, luận văn đã trình bày về một số kết quả thử nghiệm phân tích Weblog phát hiện truy cập bất thường vào máy chủ Web. Luận văn đã trình bày cụ thể

một số đặc tả dữ liệu Weblog máy chủ ghi nhận được, trình bày tóm tắt về công cụ Weblog Expert dùng để thu thập, phân tích dấu hiệu Weblog. Tiếp đó, bài đã trình bày một số kết quả thử nghiệm.

KẾT LUẬN

Việc phân tích logfile để phát hiện các truy cập bất thường vào máy chủ Web là một bước quan trọng để dự đoán nguy cơ xảy ra các cuộc tấn công vào máy chủ Web.

Hiện nay có rất nhiều phương pháp phát hiện truy cập bất thường dựa trên nguyên tắc xây dựng một tập dấu hiệu bình thường của hệ thống, sau đó thu thập các hành vi truy cập vào máy chủ rồi so sánh với tập dấu hiệu bình thường đã có. Một hành vi truy cập được coi là khác thường khi hành vi đó khác với tập dấu hiệu bình thường.

Mục đích của luận văn là nghiên cứu về vấn đề truy cập bất thường vào máy chủ Web, các phương pháp thu thập dữ liệu Weblog và phát hiện dấu hiệu bất thường.

Các kết quả nghiên cứu đạt được trong bài gồm:

- Nghiên cứu tổng quan về máy chủ web, lỗ hổng bảo mật.
- Nghiên cứu phương pháp ghi nhật ký máy chủ web, phân tích truy cập bất thường
- Nghiên cứu nguyên tắc phát hiện truy cập bất thường Web
- Ứng dụng phần mềm Weblog Expert trong việc thu thập, phân tích logfile từ máy chủ web.

DANH MỤC TÀI LIỆU THAM KHẢO

1. Hongxin Hu, Gail-Joon Ahn and Ketan Kulkarni. Anomaly Discovery and Resolution in Web Access Control Policies. SACMAT'11. Proceedings of the 16th ACM symposium on Access control models and technologies. Pp. 165-174.
2. Sipola, Tuomo; Juvonen, Antti; Lehtonen, Joel. Anomaly detection from network logs using diffusion maps. *Engineering Applications of Neural Networks* (pp. 172-181). IFIP Advances in Information and Communication Technology (363).
3. Shilin He, Jieming Zhu, Pinjia He, and Michael R. Lyu. Experience Report: System Log Analysis for Anomaly Detection. IEEE 27th International Symposium on Software Reliability Engineering (ISSRE), 2016.
4. Yi Xie and Shun-Zheng Yu. Monitoring the Application-Layer DDoS Attacks for Popular Websites. IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 17, NO. 1, FEBRUARY 2009. Pp. 15-26
5. Juan M. Estévez-Tapiador. Pedro García-Teodoro. Jesús E. Díaz-Verdejo. Detection of Web-based Attacks through Markovian Protocol Parsing. ISCC 2005. Proceedings. 10th IEEE Symposium on Computers and Communications, 2005.
6. Christopher Kruegel, Giovanni Vigna. Anomaly Detection of Webbased Attacks. CCS '03 Proceedings of the 10th ACM conference on Computer and communications security. Pp. 251-261.
7. Shaimaa Ezzat Salama. Web Server Logs Preprocessing for Web Intrusion Detection. Computer and Information Science, Vol. 4, No. 4; July 2011. Pp. 123-134.

BẢN CAM ĐOAN

Tôi cam đoan đã thực hiện việc kiểm tra mức độ tương đồng nội dung luận văn qua phần mềm DoIT một cách trung thực và đạt kết quả mức độ tương đồng **19%** toàn bộ nội dung luận văn. Bản luận văn kiểm tra qua phần mềm là bản cứng luận văn đã nộp bảo vệ trước hội đồng. Nếu sai tôi xin chịu các hình thức kỷ luật theo quy định hiện hành của Học viện

Hà Nội, 18 tháng 5 năm 2021

HỌC VIÊN CAO HỌC

Hệ thống hỗ trợ nâng cao chất lượng tài liệu

KẾT QUẢ KIỂM TRA TRÙNG LẶP TÀI LIỆU

THÔNG TIN TÀI LIỆU

Tác giả	minh1
Tên tài liệu	Luận văn Minh
Thời gian kiểm tra	18-05-2021, 02:15:56
Thời gian tạo báo cáo	18-05-2021, 03:48:47

KẾT QUẢ KIỂM TRA TRÙNG LẶP

Điểm	19
Nguồn trùng lặp tiêu biểu	[text.123doc.org, vi.wikipedia.org, viblo.asia]

(*) Kết quả trùng lặp phụ thuộc vào dữ liệu hệ thống tại thời điểm kiểm tra

HỌC VIÊN — NGƯỜI HƯỚNG DẪN KHOA HỌC

Nguyễn Anh Minh — **PGS. TSKH. Hoàng Đăng Hải**