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

TOM TẮT LUẬN VĂN THẠC SỸ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC

PGS.TSKH. HOÀNG ĐĂNG HẢI

Hà Nội - 2021

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS.TSKH. HOÀNG ĐĂNG HẢI

Phản biện 1: PGS. TS. Đỗ Trung Tuấn

Phản biện 2: TS. Tạ Quang Hùng

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

Vào lúc: 09h00 ngày 28 tháng 08 năm 2021

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.

Theo thống kê từ Trung tâm ứng cứu khẩn cấp máy tính Việt Nam (VNCERT) 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 nhập 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.

Trên cơ sở 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 nhập 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 nhập 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 nhập 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 3 chương chính với những nội dung sau:

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

Chương 2: Phát hiện truy nhập bất thường vào máy chủ web

Chương 3: 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

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.

1.1.2. Các thành phần cơ bản của máy chủ web

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.

1.1.3. Nguyên tắc hoạt động

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”.

1.1.4. 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.

1.1.5. 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

1.1.6. 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

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.

- **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…

- **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.

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

1.2.2. Các loại lỗ hổng phổ biến của Web

- Các loại lỗ hổng Web phổ biến theo OWASP

Theo OWASP năm 2013, 10 lỗ hổng bảo mật Web nghiêm trọng nhất là:

+ Chèn mã (Injection):

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

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

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

+ Cấu hình bảo mật kém (Security Misconfigtiration):

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

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

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

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

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

1.2.3. Phương pháp kiểm thử lỗ hổng

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.

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.

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.

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

- Tấn công chuyển dịch thư mục (Directory traversal attacks).
- Tấn công từ chối dịch vụ (Denial of Service Attacks).
- Tấn công chiếm giữ hệ thống tên miền (Domain Name System Hijacking).
- Tấn công nghe lén (Sniffing).
- Tấn công giả mạo (Phishing).
- Tấn công đầu độc (Pharming).
- Tấn công thay giao diện (Defacement).

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 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.

1.4. Phát hiện truy nhập bất thường vào máy chủ Web

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

Theo nguyên tắc, có hai cách để phát hiện truy nhập bất thường vào máy chủ Web. Phương pháp truyền thống là sử dụng hệ thống phát hiện xâm nhập (IDS – Intrusion Detection Systems).

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 nhập do máy chủ Web ghi liên tục theo thời gian và thực hiện phân tích để phát hiện các truy nhập bất thường. Những hành vi bất thường là dấu hiệu tiềm ẩn của một tấn công máy chủ Web.

Theo nguyên tắc trên, việc phân tích Weblog giúp cho việc phân tích, phát hiện truy nhập bất thường vào máy chủ Web. Đây là chủ đề nghiên cứu của bài luận văn và cũng là nội dung sẽ được trình bày trong các chương tiếp theo của bài.

1.5. 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, nguyên lý hoạt động, việc ghi nhật ký Weblog, giao thức HTTP, các lỗ hổng bảo mật Web, các loại tấn công vào máy chủ Web phát hiện truy nhập bất thường vào máy chủ Web và khả năng phân tích, phát hiện tấn công thông qua phân tích tệp nhật ký hoạt động của máy chủ Web (Web Log).

CHƯƠNG 2
PHÁT HIỆN TRUY NHẬP BẤT THƯỜNG VÀO MÁY CHỦ WEB

Phạm vi phân tích, phát hiện truy nhập bất thường vào 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 nhập 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.

2.1. Kiến trúc hệ thống phát hiện truy nhập bất thường

2.1.1. 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...

- **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.

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.

- **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.

- **Sumo Logic**

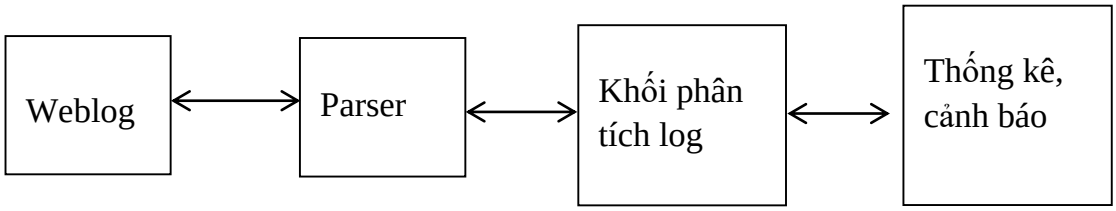
Sumo Logic là một dịch vụ phân tích, xử lý và quản lý log trên nền tảng điện toán đám mây. Sumo Logic có ưu điểm là cung cấp nhiều tính năng và có khả năng xử lý nhiều loại log, đồng thời việc cài đặt cũng tương đối dễ dàng do Sumo Logic dựa trên nền tảng điện toán đám mây, không đòi hỏi thiết bị chuyên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

- **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.

2.1.2. Kiến trúc hệ thống phát hiện truy nhập bất thường

Sơ đồ thiết kế hệ thống phân tích, phát hiện truy nhập 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 nhập máy chủ, bao gồm các sự kiện truy nhập 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 nhập bất thường bằng địa chỉ IP…từ đó cảnh báo tấn công máy chủ web.

2.2. 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 (-)

2.3. Nguyên tắc hoạt động của khối xử lý Parse

Parse service là dịch vụ đám mây đa nền tảng, được bắt đầu sử dụng từ năm 2012. Parse cung cấp các dịch vụ backend cho hầu hết các nền tảng phổ biến hiện nay như: iOS, android, window phone, Mac OS,...

Log parser là một công cụ mạnh mẽ, linh hoạt cung cấp khả năng truy cập, truy vấn tổng quát vào dữ liệu dựa trên văn bản, chẳng hạn như Log File, XML files và CSV files. Ngoài ra sử dụng các nguồn dữ liệu quan trọng trên hệ điều hành Windows chẳng hạn như Event Log, Registry,file system, Active Directory.

Có thể sử dụng Log Parser để cài đặt cho khối Parser trong hệ thống phát hiện truy nhập bất thường vào máy chủ Web. Sau khi đặt xong, có thể xem các tài liệu mô tả đầy đủ về các tính năng và các sử dụng Log Parser trên trang của Microsoft.

2.4. Thu thập thông tin Weblog cho phát hiện bất thường

2.4.1. 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.

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.

2.4.2. 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.

2.5. Phương pháp trích chọn đặc trưng dữ liệu

Về cơ bản, việc trích chọn các thuộc tính đặc trưng gồm hai phần. Thứ nhất là xây dựng các thuộc tính, thứ hai là lựa chọn các thuộc tính đặc trưng. Trong việc xử lý số liệu, xây dựng bộ các thuộc tính là một việc rất quan trọng. Khi xây dựng dữ liệu chúng ta cần phải đảm bảo thông tin không để bị mất quá nhiều và tiết kiệm chi phí. Mục đích của phần thứ hai là tìm ra những thuộc tính đại diện cho đối tượng và loại bỏ những thuộc tính không cần thiết, gây nhiễu để tăng hiệu suất của các thuật toán khai phá dữ liệu.

Giảm bớt số chiều của mẫu còn gọi là nén tập dữ liệu, thông qua trích chọn thuộc tính và lựa chọn thuộc tính. Trong việc tiền xử lý dữ liệu, giảm bớt số chiều của mẫu là bước cơ bản nhất. Lựa chọn thuộc tính có thể là một phần của trích chọn thuộc tính ví dụ như phương pháp phân tích thành phần chính hoặc là một thiết kế xử lý thuật toán chẳng hạn như trong thiết kế cây quyết định.

Lựa chọn thuộc tính là một quá trình tìm từ tập dữ liệu N ban đầu ra một tập con các thuộc tính từ M tập thuộc tính, vì vậy phải xác định tiêu chuẩn lựa chọn thuộc tính. Theo cách này, ta

có thể rút ngắn tối đa kích cỡ của không gian đặc trưng theo một tiêu chuẩn định lượng nhất định. Khi kích cỡ, số phần tử của tập N sẽ tăng lên, nên để tìm ra một tập đại diện tốt nhất là không đơn giản và tập được chọn sẽ có nhiều vấn đề liên quan đến. Một thuật toán trích chọn gồm 4 bước: Sinh tập con, lượng giá tập con, điều kiện dừng và xác nhận kết quả.

Quá trình sinh tập con là một thủ tục tìm kiếm, những tập con này dùng cho việc lượng giá. Gọi số các đại diện đặc trưng của tập dữ liệu gốc ban đầu là N, thì tổng các tập con có thể sinh ra là 2n . 2n tập này sẽ liệt kê toàn bộ các tập con của không gian, mỗi tập con được sinh ra bằng thuật toán cần được lượng giá trị bằng một tiêu chuẩn lượng giá trị nhất định và được so sánh với tập con tốt nhất đã tìm được trước nó. Thuật toán này có thể sẽ chạy mãi không dừng nếu không có điều kiện dừng phù hợp. Điều kiện dừng của một quá trình sinh phải thỏa mãn một trong số các yếu tố sau:

- Toàn bộ các phần tử của tập hợp đều phải được chọn.
- Một tập con nữa được sinh ra cũng không cho kết quả tốt hơn.
- Các phần tử chưa được chọn bị lặp lại.
- Số tập con thoả mãn điều kiện tiêu chuẩn đã được chọn đủ.

Tập con tốt nhất được chọn ra phải được lượng giá trong những trường hợp khác nhau và nó cùng với tập gốc phải biểu diễn được với dữ liệu thực tế.

Lựa chọn các thuộc tính có thể tiến hành theo hai cách. Cách thứ nhất là chọn ra tập con nhỏ nhất mà không làm giảm đi quá trình học(tự động xác định số lượng thuộc tính). Cách thứ hai là xếp loại các thuộc tính theo một tiêu chuẩn nào đó và lấy ra k thuộc tính đầu tiên (dựa vào ngưỡng để chọn thuộc tính).

Lựa chọn thuộc tính có thể dựa trên các chiến lược tìm kiếm, thước đo chất lượng thuộc tính và ước lượng, các mô hình. Có ba loại mô hình như Wrapper, Filter, 3 Embedded. Mô hình Embedded là mô hình tích hợp thuộc tính được lựa chọn khi xây dựng mô hình.

2.6. Cách thức phân tích Weblog phát hiện bất thường

Kỹ thuật phát hiện tấn công dựa vào dấu hiệu đã biết trước

Tấn công web có thể được xác định dựa trên các tập luật đã được xây dựng từ trước. Các tập luật này có thể phát hiện các tấn công vào website từ một số dấu hiệu nhất định trong log của web. Có thể là quy tắc đơn giản như phát hiện một số ký tự nhất định (tấn công SQL Injection chẳng hạn) hoặc các quy tắc phức tạp (tấn công chiếm phiên của người dùng). Kỹ thuật phân tích

tấn công này có thể được xây dựng từ hai cơ chế cơ bản: chủ động (Positive) và bị động (Negative).

Kỹ thuật phân tích bị động

Kỹ thuật phân tích tấn công bị động dựa trên các dấu hiệu từ một danh sách đen (black list hoặc rulebase) và một chính sách mặc định cho phép mọi thứ. Điều này có nghĩa là mọi request đều được chấp nhận. Danh sách đen này sẽ định nghĩa những dấu hiệu bị coi là không hợp lệ và được gắn cờ (dấu hiệu) là một cuộc tấn công web.

Kỹ thuật tấn công chủ động

Kỹ thuật phân tích tấn công chủ động ngược lại hoàn toàn với kỹ thuật phân tích bị động. Chính sách mặc định ở đây là từ chối tất cả và có một danh sách trắng (white list) các ký tự hợp lệ cho phép. Hầu hết các các tường lửa đều được cấu hình theo cách này.

Kỹ thuật phát hiện tấn công dựa vào dấu hiệu bất thường

Kỹ thuật phát hiện tấn công dựa trên sự bất thường của các request tới website thông qua các bộ luật (rules) được tự động xây dựng thông qua quá trình tự học (learnmod). Trong quá trình này, các request tới website sẽ được coi là lưu lượng sạch, nó được sử dụng làm cơ sở các mẫu để kiểm tra tính bất thường của các request. Các request khi đạt tới một ngưỡng sai lệch trong bộ luật sẽ bị đánh dấu là các request bất thường.

2.7. Kết luận chương

Trong chương 2, luận văn đã trình bày về phạm vi hệ thống, kiến trúc một số hệ thống điển hình cho thu thập, phân tích và phát hiện truy nhập bất thường vào máy chủ Web. Tiếp đó, bài đã trình bày về kiến trúc một hệ thống phát hiện truy nhập bất thường vào máy chủ Web với các thành phần Weblog, khối Parser thu thập và xử lý, khối phân tích Log, khối thống kê, cảnh báo. Luận văn đã trình bày chi tiết các nội dung cấu trúc Weblog, nguyên lý hoạt động của Parser, cách thức thu thập thông tin Weblog, cách thức trích chọn đặc trưng dữ liệu, phân tích và phát hiện truy nhập bất thường.

CHƯƠNG 3
THỬ NGHIỆM

3.1. Mô hình hệ thống máy chủ Web thử nghiệm

Hệ thống máy chủ Web thử nghiệm sử dụng cấu hình Master – Slave Database Replication để cải thiện hiệu suất do có số yêu cầu đọc thực hiện nhiều hơn yêu cầu ghi. Mô hình này đòi hỏi một master database và một hay nhiều slave database. Yêu cầu cập nhật dữ liệu sẽ được gửi cho master và các yêu cầu đọc dữ liệu sẽ được phân phối trên các slave.

3.2. Thử nghiệm phân tích, phát hiện bất thường với công cụ 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, tệp được truy cập, đường dẫn thông qua trang web, thông tin về trang giới thiệu, công cụ tìm kiếm, trình duyệt, hệ điều hành và hơn thế nữa. Chương trình tạo ra các báo cáo dễ đọc bao gồm cả thông tin văn bản (bảng) và biểu đồ.

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

3.3. Một số kết quả thử nghiệm với Weblog Expert

3.4. Kết luận chương

Trong chương 3, luận văn đã trình bày về một số kết quả thử nghiệm phân tích Weblog phát hiện truy nhập bất thường vào máy chủ We. 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 VÀ HƯỚNG PHÁT TRIỂN

Phân tích logfile, phát hiện các truy nhập bất thường vào máy chủ Web là một nhu cầu thực tế đặt ra giúp phán đoán nguy cơ xảy ra các cuộc tấn công vào máy chủ Web. Phát hiện truy cập bất thường là bước quan trọng để phát hiện ra tấn công vào máy chủ Web. Trên cơ sở đó 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.

Ngày nay có nhiều biện pháp phát hiện truy cập bất thường dựa trên nguyên lý chung là xây dựng một tập dấu hiệu bình thường của hệ thống, tiếp đó thu thập các hành vi truy nhập vào máy chủ, so sánh với tập dấu hiệu bình thường đã lưu sẵn. Nếu có sự khác biệt có nghĩa là có hành vi truy nhập bất thường. Một khả năng khác để thu thập thông tin về các hành vi là sử dụng công cụ thu thập thông tin về máy chủ Web và thực hiện phân tích, phát hiện dấu hiệu truy nhập bất thường vào máy chủ Web. Điển hình là công cụ WLELite.

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

Các kết quả nghiên cứu đạt được trong bài gồm:

- Nghiên cứu tổng quan về máy chủ web, logfile, truy cập bất thường.

- Nghiên cứu phương pháp phát hiện truy cập bất thường vào máy chủ web thông qua thu thập và phân tích Weblog.

- Ứng dụng thử nghiệm phần mềm WLELite trong việc thu thập, phân tích thông tin từ máy chủ web.

Hướng phát triển tiếp:

Luận văn là bài toán cơ sở để phát hiện dấu hiệu bất thường truy cập máy chủ web bất kì, các trung tâm dữ liệu. Kết quả thực hiện đề tài có ý nghĩa thiết thực khi được triển khai ứng dụng tại các nhà cung cấp dịch vụ