ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG

LÊ GIA HOÀNG THIỆN

22521387-MMTT2022.3

BÁO CÁO THỰC TẬP DOANH NGHIỆP

NGHIÊN CỨU VÀ PHÁT TRIỂN HỆ THỐNG OPENSOURCE SIEM

TP. HỒ CHÍ MINH, 01/2026

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG

BÁO CÁO THỰC TẬP DOANH NGHIỆP

NGHIÊN CỨU VÀ PHÁT TRIỂN HỆ THỐNG OPENSOURCE SIEM

Tên cơ quan thực tập:	**Công ty Cổ phần PopTech**
Người hướng dẫn tại công ty:	**Lê Phi**
Giảng viên hướng dẫn:	**ThS. Lê Anh Tuấn**
Tên sinh viên:	**Lê Gia Hoàng Thiện**
MSSV:	**22521387**
Lớp:	**MMTT2022.3**

TP. HỒ CHÍ MINH, 01/2026

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA MẠNG MÁY TÍNH
VÀ TRUYỀN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

Họ và tên sinh viên: .....

MSSV: .....

Công ty thực tập: .....

Thời gian thực tập:.....

Vị trí thực tập:.....

Nhiệm vụ được giao: .....

.....

.....

Đánh giá quá trình thực tập của sinh viên:

Các kết quả sinh viên đã thực hiện được:.....

.....

.....

.....

.....

.....

.....

Điểm: .....Bằng chữ:.....

……., ngày…..tháng …..năm ……

Giáo viên hướng dẫn

MỤC LỤC

Chương 1. GIỚI THIỆU VỀ NƠI THỰC TẬP 5
1.1. Giới thiệu Công Ty Cổ Phần Power Of Passion Technology Solutions (PopTech) 5
1.1.1. Thông tin chung về doanh nghiệp 5
1.1.2. Tầm nhìn, sứ mệnh và giá trị cốt lõi 6
1.1.3. Lĩnh vực hoạt động và dịch vụ chuyên môn 7
1.1.4. Môi trường làm việc 8
Chương 2. VỊ TRÍ THỰC TẬP 10
2.1. Điều kiện ứng tuyển 10
2.2. Chế độ làm việc và quyền lợi 10
2.3. Tổ chức công việc 10
2.4. Mô tả công việc 10
2.5. Nhật ký thực tập 11
Chương 3. KIẾN THỨC NỀN TẢNG 17
3.1. Tổng quan về SIEM 17
3.1.1. SIEM là gì? 17
3.1.2. Tại sao cần SIEM? 17
3.1.3. Cách SIEM hoạt động 18
3.1.4. Kiến trúc cơ bản của SIEM 18
3.2. OpenSearch - Công cụ lưu trữ và tìm kiếm 18
3.2.1. OpenSearch là gì? 19
3.2.2. Các khái niệm cơ bản 19
3.2.3. OpenSearch Dashboards 19

3.3. Công cụ thu thập Log -Vector 20
3.4. Wazuh - Agent thu thập Log 21
3.4.1. Wazuh là gì? 21
3.4.2. Thành phần sử dụng 21
3.4.3. Luồng hoạt động trong dự án 22
3.4.4. Lợi ích của việc sử dụng Wazuh Agent 22
3.5. ECS - Chuẩn hóa Log 22
3.5.1. ECS là gì? 22
3.5.2. Lợi ích của ECS 23
3.6. Docker 23
3.6.1. Docker là gì? 23
3.6.2. Tại sao dùng Docker cho SIEM? 23
3.6.3. Docker trong kiến trúc SIEM 24
3.7. High Availability và Load Balancing 24
3.7.1. High Availability là gì? 24
3.7.2. Load Balancing 24
3.7.3. Nginx Load Balancer 25
3.7.4. Failover 25
3.8. Các nguồn Log trong SIEM 25
Chương 4. NỘI DUNG THỰC TẬP 27
4.1. Giai đoạn 1: Nghiên cứu giải pháp và thiết kế kiến trúc (11/08 – 15/08/2025) 27
4.2. Xây dựng nền tảng Log và tích hợp Wazuh (15/08/2025 – 03/09/2025) 29

4.3. Đa dạng hóa nguồn logs và chuẩn hóa dữ liệu (04/09/2025 – 13/10/2025) 31

4.4. Vận hành production và tối ưu hóa (13/10/2025 – 31/10/2025) 33

4.5. Xử lý dữ liệu ESXi, tự động hóa và cân bằng tải (01/11/2025 – 04/12/2025) 34

4.6. Xử lý mạng phức tạp và hoàn thiện hệ thống (08/12/2025 – 18/12/2025) 36

Chương 5. KẾT LUẬN.....38

5.1. Kết quả đạt được và Bài học kinh nghiệm 38

5.1.1. Về mặt kiến thức và kỹ thuật: 38

5.1.2. Về kỹ năng mềm: 38

5.2. Thuận lợi và Khó khăn 39

5.2.1. Thuận lợi: 39

5.2.2. Khó khăn và Thách thức: 39

5.3. Định hướng phát triển 39

5.4. Kết luận chung.....40

DANH MỤC HÌNH VẼ

Hình 1.1 Logo PopTech 5
Hình 3.1 Hình ảnh về SIEM.....17
Hình 3.2 Logo OpenSearch.....18
Hình 3.3 Logo Vector 20
Hình 3.4 Logo Wazuh 21
Hình 3.5 Logo Docker.....23
Hình 3.6 Logo Nginx 25
Hình 4.1 Mô hình triển khai thu thập logs bằng fluentbit và fluentd.....28
Hình 4.2 Mô hình hệ thống thu thập logs bằng fluentd 29
Hình 4.3 Mô hình hệ thống thu thập logs bằng Vector.....29
Hình 4.4 Hình ảnh triển khai docker-compose.yml 30
Hình 4.5 Hình ảnh triển khai thành công theo sở đồ và hiển thị dashboard của opensearch.....31
Hình 4.6 Hình ảnh thu thập được nhiều nguồn logs khác nhau (hiển thị index) 32
Hình 4.7 Hình ảnh 1 logs ESXI được chuẩn hóa theo ECS.....35
Hình 4.8 Sơ đồ tổng thể triển khai cho khách hàng 36

DANH MỤC BẢNG

Bảng 2.1 Nhật ký thực tập.....11

DANH MỤC TỪ VIẾT TẮT

STT	Thuật ngữ	Giải nghĩa
1	SIEM	*Security Information and Event Management* – Hệ thống quản lý thông tin và sự kiện bảo mật.
2	ECS	*Elastic Common Schema* – Chuẩn hóa cấu trúc dữ liệu log dùng trong Elastic/OpenSearch.
3	SOC	*Security Operation Center* – Trung tâm điều hành an ninh mạng.
4	VMware vCenter	Nền tảng quản lý tập trung cho môi trường ảo hóa VMware.
5	Wazuh	Giải pháp mã nguồn mở dùng cho giám sát, phát hiện xâm nhập (IDS/IPS).
6	OpenSearch	Hệ thống tìm kiếm, phân tích log và hiển thị dữ liệu mã nguồn mở (fork từ Elasticsearch).
7	OpenDashboard	Giao diện dashboard của OpenSearch dùng để trực quan hóa dữ liệu log.
8	Vector	Công cụ thu thập và xử lý log hiệu suất cao (thay thế Fluentd trong hệ thống).
9	Fluentd	Công cụ thu thập, xử lý và chuyển tiếp log (log collector/forwarder).
10	Syslog	Giao thức chuẩn truyền và lưu trữ log giữa các thiết bị mạng.
11	Firewall	Tường lửa – thiết bị hoặc phần mềm bảo mật kiểm soát lưu lượng mạng.
12	Fortinet / Sophos	Các hãng cung cấp thiết bị bảo mật tường lửa được triển khai trong dự án.
13	PopTech	*Power of Passion Technology Solutions* – Công ty nơi thực tập.

14	Index	Chỉ mục dữ liệu trong OpenSearch dùng để lưu và truy vấn log.
15	Merge Index Policy	Chính sách hợp nhất chỉ mục cũ để tối ưu lưu trữ trong OpenSearch.
16	Pipeline	Chuỗi xử lý dữ liệu log từ thu thập → parsing → lưu trữ → hiển thị.
17	Parsing	Quá trình chuẩn hóa, trích xuất và ánh xạ dữ liệu log theo cấu trúc chuẩn.
18	Production environment	Môi trường sản xuất thật, khác với môi trường thử nghiệm (lab).
19	Lab environment	Môi trường thử nghiệm dùng để kiểm thử và mô phỏng hệ thống.
20	AIO	All-in-One – Mô hình triển khai tất cả các thành phần hệ thống trên cùng một máy chủ (thường dùng trong giai đoạn thử nghiệm OpenSearch).
21	HA	High Availability – Tính sẵn sàng cao, giải pháp thiết kế hệ thống nhằm giảm thiểu thời gian gián đoạn hoạt động (downtime).
22	LB	Load Balancing – Cân bằng tải, kỹ thuật phân phối lưu lượng mạng hoặc tải công việc đến nhiều máy chủ để tối ưu hiệu suất.
23	VRL	Vector Remap Language – Ngôn ngữ kịch bản chuyên dụng của Vector dùng để phân tích cú pháp (parsing) và chuyển đổi dữ liệu log.
24	VM	Virtual Machine – Máy ảo, hệ thống máy tính được mô phỏng bằng phần mềm (ví dụ các máy chạy trên ESXi).
25	VLAN	Virtual Local Area Network – Mạng cục bộ ảo, kỹ thuật chia một mạng vật lý thành nhiều mạng logic khác nhau.

26	IDS / IPS	Intrusion Detection System / Intrusion Prevention System – Hệ thống phát hiện và ngăn chặn xâm nhập (chức năng cốt lõi của Wazuh).
27	Agent	Phần mềm nhỏ gọn được cài đặt trên thiết bị đầu cuối (Endpoint) để thu thập dữ liệu và gửi về máy chủ quản lý (ví dụ: Wazuh Agent).

MỞ ĐẦU/LỜI CÁM ƠN

Trước hết, **em xin bày tỏ lòng biết ơn sâu sắc** đến các Thầy/Cô Khoa Mạng Máy Tính và Truyền Thông Dữ Liệu – Trường Đại học Công nghệ Thông tin, Đại học Quốc gia TP. Hồ Chí Minh đã tận tình giảng dạy và truyền đạt cho em những kiến thức nền tảng trong suốt quá trình học tập tại trường.

Đặc biệt, em xin gửi lời cảm ơn chân thành đến **ThS. Lê Anh Tuấn**, giảng viên hướng dẫn, người đã tận tình theo dõi, góp ý và định hướng học thuật cho em trong suốt quá trình thực hiện đề tài và hoàn thiện báo cáo thực tập.

Bên cạnh đó, em xin trân trọng cảm ơn **Ban lãnh đạo cùng toàn thể anh/chị tại Công ty Cổ phần Power of Passion Technology Solutions (PopTech)** đã tạo điều kiện thuận lợi để em được tham gia thực tập trong môi trường làm việc chuyên nghiệp. Em xin gửi lời cảm ơn đặc biệt đến **anh Lê Phi**, người hướng dẫn trực tiếp tại công ty, đã tận tình hỗ trợ, chia sẻ kinh nghiệm thực tế và định hướng kỹ thuật giúp em hoàn thành tốt các nhiệm vụ được giao.

Do thời gian và kiến thức còn hạn chế, báo cáo không tránh khỏi những thiếu sót. Em rất mong nhận được sự góp ý từ Quý Thầy/Cô và các anh/chị tại công ty để báo cáo được hoàn thiện hơn.

Em xin chân thành cảm ơn!

Chương 1. GIỚI THIỆU VỀ NƠI THỰC TẬP

1.1. Giới thiệu Công Ty Cổ Phần Power Of Passion Technology Solutions (PopTech)

Hình 1.1 Logo PopTech

1.1.1. Thông tin chung về doanh nghiệp

Công ty Cổ phần Power Of Passion Technology Solutions (gọi tắt là PopTech) là một trong những đơn vị tiên phong tại Việt Nam hoạt động theo mô hình "One-stop shop" trong lĩnh vực Công nghệ thông tin và Viễn thông. Được thành lập dựa trên triết lý cốt lõi là sự đam mê công nghệ và khát vọng mang lại giá trị thực tiễn cho doanh nghiệp, PopTech đã và đang khẳng định vị thế là nhà tích hợp hệ thống và phát triển phần mềm uy tín trên thị trường.

Khác với các công ty chỉ tập trung vào một mảng duy nhất, PopTech sở hữu năng lực toàn diện từ việc xây dựng hạ tầng phần cứng, đảm bảo an ninh mạng cho đến phát triển các giải pháp phần mềm nghiệp vụ và các công cụ tương tác khách hàng thông minh.

- **Tên tiếng Việt:** CÔNG TY CỔ PHẦN POWER OF PASSION TECHNOLOGY SOLUTIONS
- **Tên giao dịch quốc tế:** POWER OF PASSION TECHNOLOGY SOLUTIONS JOINT STOCK COMPANY
- **Tên viết tắt:** POPTECH
- **Mã số doanh nghiệp:** 0312104572
- **Ngày hoạt động:** 02/01/2013
- **Đại diện pháp luật:** Ông Trần Minh Hiển

- **Địa chỉ đăng ký kinh doanh:** 201 Nguyễn Thái Bình, Phường 4, Quận Tân Bình, Thành phố Hồ Chí Minh, Việt Nam.
- **Văn phòng làm việc :** Số 10, Đường T, Khu đô thị Lakeview City, Phường An Phú, Thành phố Thủ Đức, Thành phố Hồ Chí Minh.
- **Website:** https://poptech.vn/
- **Email liên hệ:** admin@poptech.vn
- **Slogan:** *Passion leads to Success*

1.1.2. Tầm nhìn, sứ mệnh và giá trị cốt lõi

Để hiểu rõ về định hướng phát triển của PopTech, cần nhìn vào kim chỉ nam hoạt động của công ty:

- **Tầm nhìn:** PopTech định hướng trở thành nhà cung cấp giải pháp công nghệ và chuyển đổi số hàng đầu khu vực, là đối tác chiến lược tin cậy giúp các doanh nghiệp, tổ chức tài chính và cơ quan chính phủ hiện đại hóa quy trình vận hành.
- **Sứ mệnh:**

  + **Đối với khách hàng:** Cung cấp các giải pháp công nghệ tối ưu nhất về chi phí và hiệu năng, giải quyết các "bài toán khó" trong quản trị và kinh doanh.

  + **Đối với nhân viên:** Xây dựng môi trường làm việc sáng tạo, nơi "Đam mê" được nuôi dưỡng và chuyển hóa thành năng lực thực tế.

  + **Đối với xã hội:** Đóng góp vào công cuộc chuyển đổi số quốc gia, nâng cao năng lực cạnh tranh của doanh nghiệp Việt Nam.

- **Giá trị cốt lõi :**

  + **Đổi mới:** Không ngừng cập nhật các xu hướng công nghệ mới như AI, Cloud Computing, Big Data.

  + **Chính trực:** Minh bạch trong mọi giao dịch và cam kết chất lượng với khách hàng.

+ Đồng đội): Tinh thần "Power of We" – Sức mạnh tập thể là nền tảng của mọi thành công.

1.1.3. Lĩnh vực hoạt động và dịch vụ chuyên môn

PopTech cung cấp một hệ sinh thái dịch vụ đa dạng, được chia thành các nhóm giải pháp chiến lược sau:

1. Giải pháp hạ tầng và tích hợp hệ thống : Đây là thế mạnh truyền thống của PopTech, bao gồm việc tư vấn, thiết kế và triển khai các hệ thống cốt lõi cho doanh nghiệp:

- **Hạ tầng mạng :** Triển khai hệ thống mạng LAN/WAN/SD-WAN hiệu suất cao sử dụng thiết bị từ Cisco, Juniper, Aruba. Tối ưu hóa luồng dữ liệu và đảm bảo kết nối thông suốt.
- **Bảo mật thông tin:** Cung cấp giải pháp bảo mật đa lớp bao gồm Tường lửa thế hệ mới (NGFW - Palo Alto, Fortinet), Bảo mật điểm cuối (Endpoint Security), và các giải pháp chống thất thoát dữ liệu (DLP).
- **Trung tâm dữ liệu và ảo hóa :**

  + Triển khai hạ tầng siêu hội tụ (HCI - Hyper-Converged Infrastructure) với công nghệ của Nutanix và VMware.

  + Giải pháp lưu trữ và sao lưu dự phòng đảm bảo an toàn dữ liệu tuyệt đối (Veeam, Commvault).

2. Dịch vụ phát triển Phần mềm: PopTech sở hữu đội ngũ kỹ sư phần mềm chất lượng cao, cung cấp các dịch vụ:

- **Gia công phần mềm (Outsourcing):** Cung cấp nhân sự và phát triển dự án theo mô hình ODC (Offshore Development Center) cho các thị trường quốc tế và nội địa.
- **Phát triển ứng dụng Mobile và Web:** Xây dựng các Super App, ứng dụng thương mại điện tử, và các cổng thông tin doanh nghiệp (Enterprise Portal) trên nền tảng iOS, Android, React Native, .NET, Java.

- **Chuyển đổi số quy trình (Business Process Digitalization):** Số hóa quy trình giấy tờ, quản lý quy trình nghiệp vụ (BPM), và tích hợp hệ thống ERP/CRM.

3. Giải pháp yương tác khách hàng: Đây là mảng công nghệ mũi nhọn giúp PopTech tạo nên sự khác biệt:

- Contact Center và Call Center: Triển khai tổng đài thông minh dựa trên nền tảng Avaya và Cisco.
- Omni-channel Platform: Hợp nhất các kênh giao tiếp (Thoại, Email, Chat, Social Media) vào một giao diện duy nhất giúp doanh nghiệp chăm sóc khách hàng liền mạch.
- AI và Automation (Sản phẩm Zóng): PopTech phát triển và triển khai các giải pháp như Zóng Bot (Chatbot/Voicebot AI), công nghệ chuyển đổi giọng nói thành văn bản (Speech-to-Text) và xác thực sinh trắc học giọng nói (Voice Biometrics), giúp tự động hóa quy trình CSKH.

1.1.4. Môi trường làm việc

Văn hóa doanh nghiệp và Môi trường làm việc:

- **Không gian mở:** Văn phòng tại Lakeview City được thiết kế như một không gian "Co-working space" hiện đại, gần gũi thiên nhiên, phá bỏ sự gò bó của các văn phòng truyền thống trong tòa nhà kính.
- **Triết lý "Work Smart, Play Hard":** PopTech không quản lý nhân viên bằng chấm công thẻ từ cứng nhắc mà quản lý bằng hiệu quả công việc (KPI/OKR). Công ty khuyến khích nhân viên chủ động thời gian, hỗ trợ làm việc từ xa (Remote/Hybrid working).
- **Đào tạo liên tục:** Văn hóa "Learning Organization" (Tổ chức học tập) được chú trọng. Nhân viên thường xuyên được tài trợ tham gia các khóa học và thi chứng chỉ quốc tế của Cisco, Microsoft, AWS.

- Hoạt động gắn kết: Các chương trình Teambuilding, Company Trip, Happy Hour diễn ra thường xuyên để gắn kết các thành viên (Popers), xây dựng một tập thể đoàn kết như gia đình.

Chương 2. VỊ TRÍ THỰC TẬP

2.1. Điều kiện ứng tuyển

- Hình thức: Full-time; có xem xét Intern/Part-time cho ứng viên phù hợp.
- Mức lương: Thỏa thuận theo năng lực và dự án.
- Kinh nghiệm: Không bắt buộc (Team sẽ đào tạo qua các dự án thực tế).
- Độ tuổi: < 30 tuổi

2.2. Chế độ làm việc và quyền lợi

- Thời gian thực tập: 28/08/2025 – 30/10/2025
- Hình thức: Trực tiếp tại văn phòng hoặc làm việc từ xa tại nha
- Lịch làm việc: 8h30 – 17h30 (gồm 1h30p nghỉ trưa), từ thứ 2 đến thứ 6.
- Quyền lợi:
  + Lương thực tập
  + Trợ cấp ăn trưa

2.3. Tổ chức công việc

- Bộ phận: Network và System
- Nhóm: Monitor và Log
- Chức vụ: SOC architecture
- Cấp bậc: Intern
- Quản lý trực tiếp: Lê Phi – Technical Sale

2.4. Mô tả công việc

- Nghiên cứu các giải pháp giám sát và phân tích log tập trung: OpenSearch (Mô hình All-in-One), OpenDashboard, Wazuh, Vector, Fluentd, OpenTelemetry.
- Tìm hiểu cấu trúc dữ liệu ECS Schema.
- Thiết kế mô hình kiến trúc OpenSearch và Wazuh triển khai trên nền tảng Docker.
- Triển khai thực tế kiến trúc All-in-One OpenSearch, cụm Cluster Wazuh và giải pháp cân bằng tải .

- Cấu hình thu thập log đa nguồn: Firewall, Docker, Web Server, SQL, Windows Server, Linux, ESXi.
- Chuyển đổi và tối ưu hóa hệ thống thu thập log từ Fluentd sang Vector.
- Thực hiện Parsing, Filtering, Tagging và Transforming log trên môi trường Linux.
- Gắn thẻ dữ liệu log để phục vụ việc giám sát và tra cứu.
- Giám sát, phát hiện và khắc phục các lỗi trên môi trường Production .
- Hỗ trợ team Parsing kiểm tra log, fix lỗi hệ thống và lỗi gắn tag.
- Viết tài liệu hướng dẫn triển khai (Wazuh, Vector, OpenSearch) và chia sẻ kiến thức trên diễn đàn công ty.

2.5. Nhật ký thực tập

Bảng 2.1 Nhật ký thực tập

Thời gian	Nội dung công việc	Kết quả đạt được
11/08/2025	Nghiên cứu mô hình OpenSearch All-in-One (AIO) và các công cụ liên quan (Dashboards, Fluentd...).	Hiểu rõ kiến trúc lưu trữ và đánh chỉ mục của OpenSearch.
14/08/2025 - 15/08/2025	Viết tài liệu thiết kế và hướng dẫn triển khai hệ thống lên diễn đàn nội bộ.	Hoàn thiện tài liệu chuẩn hóa quy trình triển khai cho team.
15/08/2025	Thiết kế và vẽ mô hình kiến trúc AIO OpenSearch triển khai trên nền tảng Docker. Triển khai giải pháp Fluentd và khắc phục lỗi thu thập log trên môi trường Windows.	Có sơ đồ kiến trúc tổng thể làm cơ sở triển khai. Fluentd hoạt động, log Windows được thu thập đúng định dạng.

15/08/2025	Triển khai giải pháp Fluentd và khắc phục lỗi thu thập log trên môi trường Windows.	Fluentd hoạt động, log Windows được thu thập đúng định dạng.
18/08/2025	Viết tài liệu kỹ thuật về giải pháp Vector trên diễn đàn (thay thế Fluentd).	Phân tích được ưu điểm hiệu năng của Vector (Rust) so với Fluentd.
21/08/2025	Tìm hiểu về Wazuh (Kiến trúc Manager, Agent, Indexer).	Nắm vững cơ chế hoạt động của giải pháp SIEM Wazuh.
22/08/2025	Triển khai kiến trúc AIO OpenSearch hoàn chỉnh trên Docker.	Hệ thống Core OpenSearch vận hành ổn định.
23/08/2025 - 26/08/2025	Triển khai Wazuh Manager và tích hợp với OpenSearch.	Hệ thống Wazuh được cài đặt và kết nối thành công với Fluentd
28/08/2025	Viết tài liệu "Tìm hiểu Wazuh" chia sẻ trên diễn đàn.	Tài liệu hóa kiến thức về Wazuh cho nội bộ.
03/09/2025	Viết tài liệu hướng dẫn triển khai chi tiết giải pháp Vector.	Có quy trình chuẩn để cài đặt Vector làm Aggregator.
03/09/2025 - 04/09/2025	Cấu hình thu thập log từ thiết bị Firewall.	Log mạng từ Firewall bắt đầu đổ về hệ thống.
09/09/2025	Triển khai chính thức giải pháp Vector vào hệ thống.	Vector thay thế vai trò thu thập log chính thức.
11/09/2025-12/09/2025	Cấu hình thu thập log từ ứng dụng Docker Container.	Giám sát được log output (stdout/stderr) của các container.

15/09/2025 - 16/09/2025	Cấu hình thu thập log ứng dụng Web (Nginx/Apache).	Thu thập được Access log và Error log của Web Server.
21/09 - 22/09(2 ngày)	Cấu hình thu thập log từ cơ sở dữ liệu SQL.	Ghi nhận hoạt động truy vấn của Database SQL.
25/09/2025 - 26/09/2025	Mở rộng thu thập log từ các thiết bị Windows (Server/PC).	Log Event Viewer (System, Security) được thu thập đầy đủ.
29/09/2025 - 30/09/2025	Mở rộng thu thập log từ các thiết bị Linux.	Log Syslog và Auth log của Linux Server được thu thập.
30/09/2025- 01/10/2025	Nghiên cứu về chuẩn hóa dữ liệu Elastic Common Schema (ECS).	Hiểu cấu trúc chuẩn để đồng nhất định dạng log.
01/10/2025- 03/10/2025	Thực hiện Parsing logs Linux (tách trường dữ liệu).	Log Linux được phân tích thành các trường JSON cụ thể.
04/10/2025 - 06/10/2025	Thực hiện Filtering logs Linux (lọc log nhiễu).	Loại bỏ log rác, tối ưu dung lượng lưu trữ.
07/10/2025 - 08/10/2025	Thực hiện Tagging logs Linux để chia index phục vụ SIEM.	Dữ liệu được phân loại, dễ dàng truy vấn theo tag.
08/10/2025 - 10/10/2025	Thực hiện Transform logs Linux (xóa trường dư thừa).	Bản tin log gọn nhẹ, chỉ giữ lại thông tin giá trị.
13/10/2025	Triển khai lại kiến trúc hệ thống theo file cấu hình Vector chuẩn.	Đồng bộ hóa cấu hình toàn hệ thống theo Vector.
12/10/2025 - 14/10/2025	Triển khai thu thập log Windows Server trên môi trường Production.	Hệ thống bắt đầu giám sát hạ tầng thực tế quan trọng.

15/10 /2025-16/10/2025	Kiểm tra kết nối Agents và phối hợp các bộ phận để thu thập log.	Đảm bảo luồng dữ liệu thông suốt giữa các phòng ban.
17/10/2025	Chuyển đổi (Convert) toàn bộ file config từ Fluentd sang Vector.	Hoàn tất migration sang Vector hiệu năng cao.
18/10/2025 - 20/10/2025	Triển khai phân nhóm (Groups) trên Wazuh Manager.	Quản lý Agent khoa học theo từng nhóm chức năng.
22/10/2025	Fix lỗi sập (Crash) Wazuh Manager do thiếu Rules/Decoders.	Hệ thống Wazuh ổn định, xử lý được lưu lượng log lớn.
22/10/2025 - 23/10/2025	Lấy log và thực hiện gắn tag bổ sung cho dữ liệu.	Dữ liệu log phong phú hơn, hỗ trợ tìm kiếm chi tiết.
26/10/2025 - 27/10/2025	Kiểm tra lỗi gắn tag và nghiên cứu khắc phục.	Sửa lỗi logic trong script gắn thẻ dữ liệu.
27/10/2025 - 29/10/2025	Hỗ trợ team parsing check log và sửa lỗi hệ thống tạm dừng.	Khôi phục hoạt động hệ thống nhanh chóng khi gặp sự cố.
02/11/2025 - 03/11/2025	Hỗ trợ gắn tag cho team parsing và bắt đầu thu log ESXi.	Dữ liệu ảo hóa ESXi bắt đầu được đưa vào quy trình.
04/11/2025 - 05/11/2025	Fix lỗi tràn bộ nhớ (Memory Leak) trên Production.	Vector hoạt động ổn định, RAM được kiểm soát tốt.
05/11/2025 - 06/11/2025	Tìm kiếm tài liệu và nghiên cứu cách Parsing log ESXi.	Xác định cấu trúc log đặc thù của VMware ESXi.
07/11/2025 - 10/11/2025	Thực hiện Parsing log ESXi (Regex phức tạp).	Log ESXi được bóc tách thành công các trường thông tin.

13/11/2025	Hỗ trợ team parsing tạo Group Parsing chuyên biệt.	Quy trình xử lý log được chuẩn hóa cho team vận hành.
12/11/2025 - 13/11/2025	Nghiên cứu triển khai log cho khách hàng (test trên máy ảo).	Có phương án kỹ thuật khả thi cho khách hàng.
13/11/2025 - 17/11/2025	Xây dựng phương án triển khai dựa trên sơ đồ của khách hàng.	Hoàn thiện kế hoạch triển khai chi tiết.
17/11/2025	Họp team chuẩn bị triển khai thực tế cho khách hàng.	Thống nhất phương án và phân công nhiệm vụ.
25/11/2025 - 26/11/2025	Nghiên cứu và triển khai cụm Cluster OpenSearch.	Hệ thống chuyển sang mô hình Cluster chịu lỗi tốt.
24/11/2025 - 26/11/2025	Nghiên cứu và triển khai Cân bằng tải (Load Balancing).	Hệ thống có khả năng phân phối tải, đảm bảo HA.
29/11/2025 - 01/12/2025	Tạo script chạy tự động Agent Linux (Failover).	Tự động hóa việc cài đặt Agent Linux, giảm thao tác tay.
27/11/2025 - 01/12/2025	Tiến hành thu thập log diện rộng.	Dữ liệu từ toàn bộ hệ thống được thu thập về trung tâm.
02/12/2025 - 03/12/2025	Họp triển khai thu log và xem xét đánh giá hệ thống.	Rà soát hiệu năng và chất lượng dữ liệu thu thập.
03/12/2025	Viết bài hướng dẫn triển khai Agent Linux tự động.	Tài liệu hóa quy trình automation cho Linux.
03/12/2025 - 04/12/2025	Tạo script chạy tự động Agent Windows (Failover).	Tự động hóa việc cài đặt Agent Windows.

08/12/2025	Điều chỉnh Load Balancing để lấy log từ VLAN khác.	Log được thu thập xuyên suốt qua các phân vùng mạng.
10/12/2025	Cấu hình Nginx dạng Stream Proxy cho Vector.	Giải quyết bài toán thu log qua vùng mạng bị hạn chế.
11/12/2025	Cấu hình Vector đẩy log hoàn chỉnh lên Dashboard.	Dữ liệu hiển thị trực quan trên OpenSearch Dashboards.
12/12/2025 - 15/12/2025	Thu thập và kiểm tra toàn bộ log từ các Agent.	Đảm bảo tính toàn vẹn dữ liệu từ tất cả các nguồn.
15/12/2025	Tìm hiểu kỹ thuật lấy Agent từ VLAN khác (bổ sung).	Tối ưu hóa giải pháp kết nối mạng đa vùng.
18/12/2025	Giải thích quy trình log Network/Firewall và bàn giao.	Hoàn tất dự án, chuyển giao hệ thống và tài liệu.

Chương 3. KIẾN THỨC NỀN TẢNG

3.1. Tổng quan về SIEM

Hình 3.1 Hình ảnh về SIEM

3.1.1. SIEM là gì?

SIEM (Security Information and Event Management) là hệ thống giám sát bảo mật tập trung, có nhiệm vụ thu thập log từ tất cả các thiết bị và ứng dụng trong hệ thống IT, sau đó phân tích để phát hiện các mối đe dọa và sự cố bảo mật.

3.1.2. Tại sao cần SIEM?

Trong một hệ thống IT lớn, có hàng trăm servers, thiết bị mạng, ứng dụng đều tạo ra log riêng. Nếu log nằm rải rác khắp nơi thì:

- Khó phát hiện tấn công khi hacker di chuyển qua nhiều hệ thống
- Tốn thời gian tìm kiếm log khi có sự cố
- Không thể tổng hợp được bức tranh toàn cảnh về bảo mật

SIEM giải quyết vấn đề này bằng cách tập trung tất cả log về một nơi, chuẩn hóa và phân tích chúng.

3.1.3. Cách SIEM hoạt động

Bước 1 - Thu thập log: Agents được cài trên các máy chủ, thiết bị để gửi log về SIEM.

Bước 2 - Chuẩn hóa log: Mỗi hệ thống có format log khác nhau. SIEM phải chuyển đổi tất cả về một định dạng chung để dễ tìm kiếm và so sánh.

Bước 3 - Lưu trữ và index: Log được lưu vào database đặc biệt (như OpenSearch) để có thể tìm kiếm cực nhanh trong hàng triệu log entries.

Bước 4 - Phân tích và cảnh báo: SIEM sử dụng rules để phát hiện các pattern nguy hiểm.

Bước 5 - Hiển thị và báo cáo: Dashboard trực quan cho phép theo dõi tình hình bảo mật real-time và tạo báo cáo định kỳ.

3.1.4. Kiến trúc cơ bản của SIEM

[Nguồn log] → [Agent thu thập] → [Xử lý log] → [Lưu trữ] → [Dashboard]

- **Nguồn log**: Servers, network devices, applications
- **Agent thu thập**: Fluentd, Vector, Filebeat
- **Xử lý log**: Parse, filter, enrich, normalize
- **Lưu trữ**: OpenSearch, Elasticsearch

3.2. OpenSearch - Công cụ lưu trữ và tìm kiếm

Hình 3.2 Logo OpenSearch

3.2.1. OpenSearch là gì?

OpenSearch là công cụ tìm kiếm mã nguồn mở, chuyên lưu trữ và tìm kiếm dữ liệu lớn. OpenSearch được fork từ Elasticsearch, rất phù hợp cho SIEM vì có thể:

- Lưu trữ hàng tỷ log entries
- Tìm kiếm trong vài giây thay vì vài phút
- Mở rộng dễ dàng bằng cách thêm server

3.2.2. Các khái niệm cơ bản

Cluster: Nhóm nhiều servers (nodes) làm việc cùng nhau như một hệ thống thống nhất.

Node: Một server trong cluster. Có các loại:

- Data node: Lưu trữ dữ liệu
- Master node: Điều phối cluster
- Ingest node: Xử lý dữ liệu trước khi lưu

Index: Giống như một "database" chứa log cùng loại. Ví dụ:

- firewall-logs: Chứa log từ firewall
- windows-logs: Chứa log từ Windows servers
- nginx-logs: Chứa log từ web server

Document: Một log entry cụ thể, được lưu dưới dạng JSON.

Shard: Index được chia thành nhiều phần nhỏ (shards) để phân tán trên nhiều servers, giúp tăng tốc độ và khả năng chịu lỗi.

3.2.3. OpenSearch Dashboards

Là giao diện web để làm việc với OpenSearch, gồm:

- **Discover**: Tìm kiếm và xem log
- **Visualize**: Tạo biểu đồ (pie chart, line chart, bar chart)
- **Dashboard**: Tổng hợp nhiều biểu đồ thành trang tổng quan
- **Alerting**: Tạo cảnh báo tự động

3.3. Công cụ thu thập Log -Vector

Hình 3.3 Logo Vector

Vector là công cụ thu thập log thế hệ mới, được viết bằng ngôn ngữ Rust, hiệu suất cao hơn Fluentd rất nhiều. **Trong dự án, Vector đóng vai trò chính trong việc parsing, filtering, transform và routing logs.**

Ưu điểm:

- Tiêu tốn ít RAM hơn
- Đảm bảo không mất log (exactly-once delivery)
- Có metrics để giám sát chính nó
- Xử lý và transform data mạnh mẽ với VRL (Vector Remap Language)

Vai trò trong kiến trúc:

[Wazuh Manager Log Files] → [Vector] → [OpenSearch]

Vector thực hiện:

1. **Đọc log files**: Từ Wazuh Manager hoặc trực tiếp từ sources khác
2. **Parsing**: Phân tích cú pháp log thành structured data
3. **Filtering**: Lọc bỏ logs không cần thiết
4. **Transform**: Chuẩn hóa theo ECS, thêm fields, xóa fields dư thừa
5. **Routing**: Gửi logs đến đúng index trong OpenSearch

6. **Output**: Đẩy logs đã xử lý lên OpenSearch

Vector Remap Language (VRL): Ngôn ngữ riêng của Vector để transform data một cách mạnh mẽ và an toàn.

3.4. Wazuh - Agent thu thập Log

Hình 3.4 Logo Wazuh

3.4.1. Wazuh là gì?

Wazuh là nền tảng bảo mật mã nguồn mở. Trong dự án này, Wazuh được sử dụng đơn giản như một agent để thu thập log từ các endpoints (Windows, Linux) và gửi về hệ thống xử lý log tập trung.

3.4.2. Thành phần sử dụng

Wazuh Agent: Phần mềm được cài trên các máy chủ cần thu thập log. Agent có khả năng:

- Đọc log từ các file theo đường dẫn được cấu hình
- Đọc Windows Event Logs
- Đọc log từ các ứng dụng
- Gửi log về Wazuh Manager

Wazuh Manager: Server trung tâm nhận log từ các agents. Manager không thực hiện phân tích phức tạp mà chỉ:

- Nhận log từ agents
- Lưu log vào file theo format JSON
- Cung cấp điểm thu thập tập trung cho Vector

3.4.3. Luồng hoạt động trong dự án

[Endpoint] → [Wazuh Agent] → [Wazuh Manager] → [Log Files] → [Vector] → [OpenSearch]

1. Wazuh Agent đọc log từ đường dẫn cấu hình (ví dụ: /var/log/app.log, Windows Event)
2. Agent gửi log đến Wazuh Manager qua port 1514
3. Manager lưu log vào file (thường tại /var/ossec/logs/archives/)
4. Vector đọc log files từ Manager
5. Vector thực hiện parsing, filtering, transform
6. Vector đẩy log đã xử lý lên OpenSearch

3.4.4. Lợi ích của việc sử dụng Wazuh Agent

- **Thu thập tập trung**: Một agent có thể đọc nhiều log files khác nhau
- **Hỗ trợ Windows**: Dễ dàng thu thập Windows Event Log
- **Reliable**: Agent có cơ chế buffer, retry khi mất kết nối
- **Lightweight**: Agent nhẹ, không tốn nhiều tài nguyên
- **Quản lý tập trung**: Quản lý tất cả agents từ Manager

3.5. ECS - Chuẩn hóa Log

3.5.1. ECS là gì?

ECS (Elastic Common Schema) là bộ quy tắc đặt tên và cấu trúc log chung. Thay vì mỗi hệ thống có cách đặt tên khác nhau, ECS thống nhất tất cả.

Ví dụ không dùng ECS:

- Firewall: src_ip, dst_ip

- Web server: client_ip, server_ip
- Windows: source_address, destination_address

Khi dùng ECS:

- Tất cả đều dùng: source.ip, destination.ip

3.5.2. Lợi ích của ECS

- **Tìm kiếm dễ dàng**: Một query hoạt động trên tất cả log sources
- **Dashboard tái sử dụng**: Tạo một dashboard có thể xem log từ nhiều nguồn
- **Correlation hiệu quả**: Dễ dàng liên kết events từ các hệ thống khác nhau
- **Security rules chung**: Một rule có thể áp dụng cho nhiều loại log

3.6. Docker

Hình 3.5 Logo Docker

3.6.1. Docker là gì?

Docker là công nghệ container hóa, cho phép đóng gói ứng dụng và tất cả dependencies vào một "container" độc lập, chạy được mọi nơi.

3.6.2. Tại sao dùng Docker cho SIEM?

- **Dễ triển khai**: Chỉ cần docker-compose up là cả stack SIEM chạy
- **Nhất quán**: Môi trường dev, test, production giống hệt nhau
- **Dễ mở rộng**: Muốn thêm node OpenSearch? Chỉ cần thêm container

- **Tiết kiệm tài nguyên**: Container nhẹ hơn máy ảo rất nhiều
- **Version control**: Dễ rollback về version cũ nếu có vấn đề

3.6.3. Docker trong kiến trúc SIEM

Các thành phần SIEM thường được containerize:

- OpenSearch nodes
- OpenSearch Dashboards
- Wazuh Manager
- Vector collectors

3.7. High Availability và Load Balancing

3.7.1. High Availability là gì?

High Availability (HA) nghĩa là hệ thống luôn sẵn sàng hoạt động, không bị gián đoạn ngay cả khi có server bị lỗi.

Trong SIEM, HA đạt được bằng cách:

- Có nhiều collectors: Nếu một collector chết, còn collectors khác
- Có nhiều OpenSearch nodes: Dữ liệu được nhân bản, một node chết không mất data
- Có nhiều Wazuh managers: Agents có thể kết nối đến manager dự phòng

3.7.2. Load Balancing

Load Balancing là kỹ thuật phân tán tải đều cho nhiều servers để:

- Không có server nào bị quá tải
- Tăng tốc độ xử lý
- Đảm bảo hệ thống không chết nếu một server gặp sự cố

3.7.3. Nginx Load Balancer

Hình 3.6 Logo Nginx

Nginx là web server và load balancer mạnh mẽ. Trong SIEM, Nginx được dùng để:

- Phân tán log từ agents đến nhiều collectors
- Health check: Tự động loại bỏ collector bị lỗi
- SSL termination: Xử lý mã hóa cho agents

3.7.4. Failover

Failover là cơ chế tự động chuyển sang hệ thống dự phòng khi hệ thống chính lỗi.

Ví dụ: Agent được cấu hình với 3 collector endpoints:

1. Agent gửi log đến collector1
2. Nếu collector1 không phản hồi, tự động chuyển sang collector2
3. Nếu collector2 cũng lỗi, chuyển sang collector3
4. Khi collector1 hoạt động trở lại, agent quay về dùng collector1

Nhờ failover, log không bao giờ bị mất ngay cả khi có sự cố.

3.8. Các nguồn Log trong SIEM

- Windows Logs
- Linux Logs
- Firewall Logs
- Web Server Logs
- Database Logs

- Docker Container Logs
- ESXi Logs
- Switchboard logs

Chương 4. NỘI DUNG THỰC TẬP

4.1. Giai đoạn 1: Nghiên cứu giải pháp và thiết kế kiến trúc (11/08 – 15/08/2025)

Mục tiêu:

Giai đoạn này nhằm xác định bài toán quản lý log tập trung trong môi trường doanh nghiệp, nghiên cứu và lựa chọn giải pháp SIEM mã nguồn mở phù hợp, đồng thời thiết kế kiến trúc hệ thống đảm bảo khả năng triển khai thực tế.

Nội dung thực hiện:

Hệ sinh thái OpenSearch được nghiên cứu toàn diện, bao gồm OpenSearch Engine và OpenSearch Dashboards, với trọng tâm là khả năng lưu trữ dữ liệu log phân tán, cơ chế đánh chỉ mục và truy vấn hiệu năng cao. Trên cơ sở phân tích đặc điểm hạ tầng và yêu cầu triển khai thực tế, mô hình OpenSearch All-in-One được lựa chọn cho giai đoạn khởi đầu nhằm tối ưu tài nguyên và đơn giản hóa công tác quản lý.

Song song đó, các công cụ thu thập log phổ biến như Fluentd và Vector được phân tích, so sánh dựa trên tiêu chí hiệu năng, mức tiêu thụ tài nguyên và khả năng mở rộng. Kết quả đánh giá cho thấy Vector có ưu thế rõ rệt trong xử lý log quy mô lớn, do đó được lựa chọn làm công cụ thu thập và xử lý log chính trong kiến trúc đề xuất.

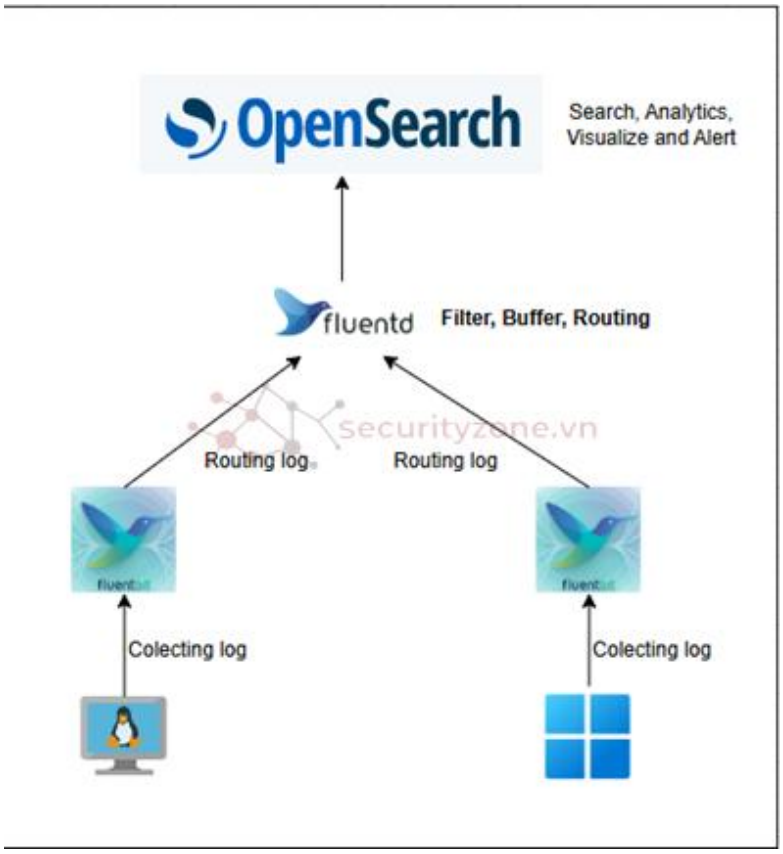
Kiến trúc tổng thể của hệ thống SIEM được thiết kế trên nền tảng Docker, xác định rõ luồng dữ liệu từ nguồn log (hệ điều hành, ứng dụng, thiết bị mạng) đến tầng xử lý, lưu trữ và trực quan hóa. Các thành phần mạng nội bộ, cơ chế lưu trữ dữ liệu bền vững và cấu hình triển khai được xây dựng làm nền tảng cho các giai đoạn tiếp theo.

Vai trò hướng dẫn:

- Anh Lê Phi (PopTech) đóng vai trò định hướng kỹ thuật, góp ý lựa chọn kiến trúc phù hợp với điều kiện hạ tầng doanh nghiệp, đảm bảo khả năng mở rộng và vận hành trên môi trường Production.
- ThS. Lê Anh Tuấn (GVHD) định hướng học thuật, góp ý về cấu trúc nội dung, tính logic và sự liên kết giữa cơ sở lý thuyết và kiến trúc triển khai.

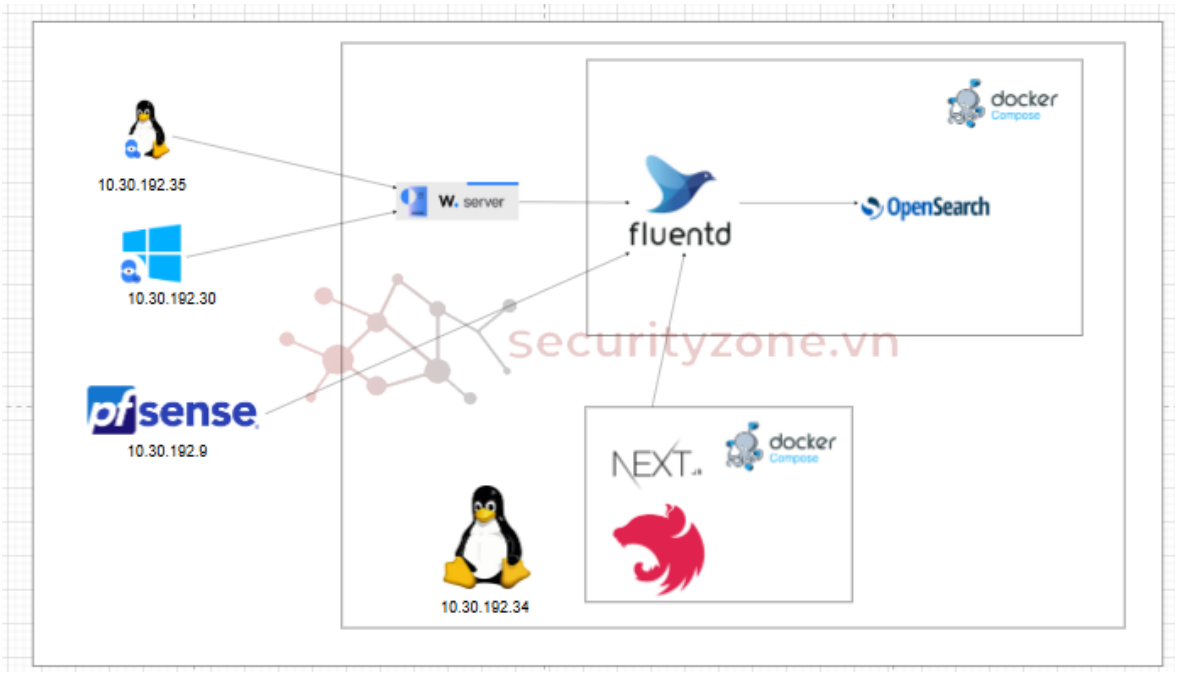
Kết quả đạt được:

- Hoàn thiện kiến trúc tổng thể của hệ thống SIEM mã nguồn mở.
- Xác định rõ vai trò của từng thành phần trong pipeline xử lý log, làm cơ sở cho triển khai thực tế.



Hình 4.1 Mô hình triển khai thu thập logs bằng fluentbit và fluentd

4.2. Xây dựng nền tảng Log và tích hợp Wazuh (15/08/2025 – 03/09/2025)



Hình 4.2 Mô hình hệ thống thu thập logs bằng fluentd

Hình 4.3 Mô hình hệ thống thu thập logs bằng Vector

Mục tiêu:

Hiện thực hóa kiến trúc đã thiết kế trên môi trường Lab, xây dựng nền tảng thu thập và lưu trữ log tập trung, đồng thời tích hợp Wazuh theo hướng tối ưu tài nguyên.

Nội dung thực hiện:

OpenSearch All-in-One được triển khai trên Docker và cấu hình tham số bộ nhớ JVM phù hợp nhằm đảm bảo hiệu năng đánh chỉ mục. OpenSearch Dashboards được cấu hình xác thực người dùng nội bộ để bảo mật truy cập hệ thống.

Wazuh Manager và Wazuh Agent được triển khai trên các máy thử nghiệm. Chế độ ghi log đầy đủ dưới dạng JSON được kích hoạt để phục vụ pipeline xử lý log, trong

khi các module không cần thiết được vô hiệu hóa nhằm giảm tải tài nguyên và đưa Wazuh về vai trò thu thập log.

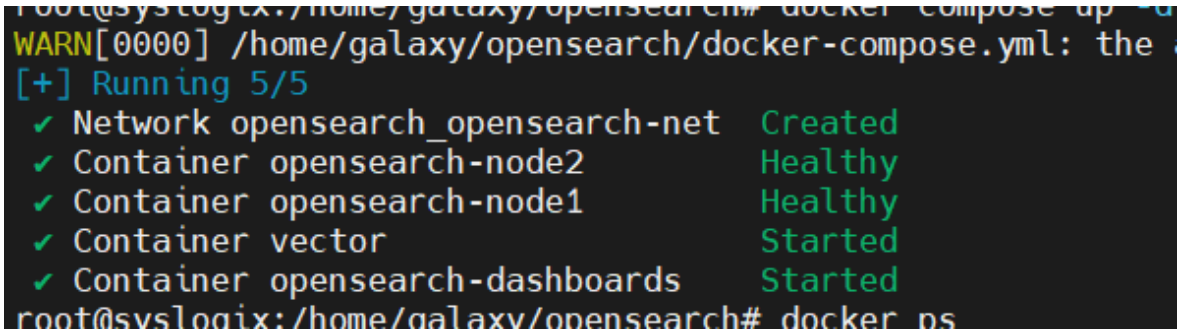
Vector được triển khai làm bộ chuyển tiếp trung tâm, thực hiện đọc log JSON từ Wazuh, parsing dữ liệu ban đầu và đẩy log vào OpenSearch thông qua Bulk API nhằm tối ưu hiệu năng ghi.

Vai trò hướng dẫn:

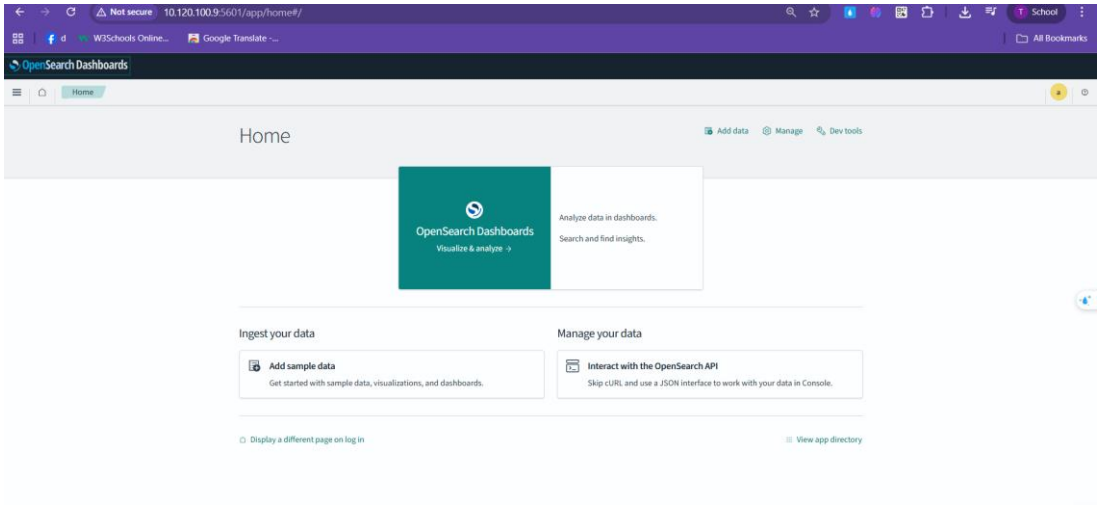
- Anh Lê Phi hỗ trợ định hướng cách tối ưu Wazuh trong vai trò thu thập log, tránh chồng chéo xử lý.
- ThS. Lê Anh Tuấn góp ý cách trình bày pipeline xử lý log và mối liên hệ giữa các thành phần trong hệ thống.

Kết quả đạt được:

- Hệ thống OpenSearch – Wazuh – Vector hoạt động ổn định trên môi trường Lab.
- Log được thu thập đầy đủ, độ trễ thấp, sẵn sàng mở rộng sang các nguồn khác.



Hình 4.4 Hình ảnh triển khai docker-compose.yml



Hình 4.5 Hình ảnh triển khai thành công theo sở đồ và hiển thị dashboard của opensearch

4.3. Đa dạng hóa nguồn logs và chuẩn hóa dữ liệu (04/09/2025 – 13/10/2025)

Mục tiêu:

Mở rộng phạm vi thu thập log từ nhiều lớp hạ tầng khác nhau, bao gồm thiết bị mạng, hệ điều hành và ứng dụng; đồng thời chuẩn hóa dữ liệu theo Elastic Common Schema (ECS) nhằm đảm bảo tính thống nhất và khả năng phân tích chéo trong hệ thống SIEM.

Nội dung thực hiện:

Luồng log từ các thiết bị mạng như Firewall Fortigate và Switch Cisco được cấu hình xuất log theo giao thức Syslog (UDP port 514). Vector được cấu hình mở cổng lắng nghe trực tiếp luồng Syslog này, cho phép tiếp nhận log mạng với lưu lượng lớn mà không cần thông qua Wazuh Manager, từ đó giảm thiểu nguy cơ quá tải tại tầng trung gian.

Đối với log hệ điều hành và ứng dụng, cơ chế thu thập thông qua Wazuh Agent tiếp tục được sử dụng. Các nguồn log bao gồm log hệ điều hành, log container Docker (json-file driver), log web server (Nginx access log) và log cơ sở dữ liệu (MySQL

error log). Việc thu thập đa nguồn giúp hệ thống có cái nhìn toàn diện hơn về trạng thái vận hành và an toàn của hạ tầng.

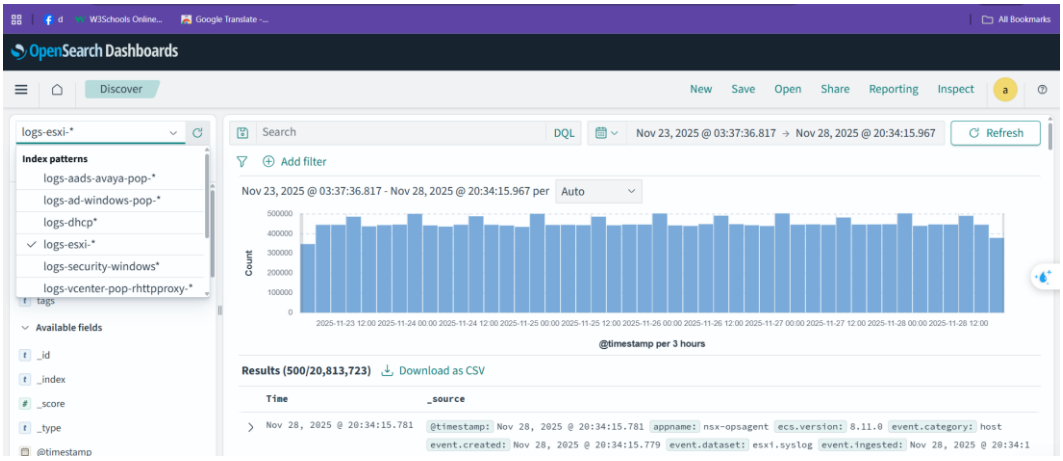
Chuẩn Elastic Common Schema được nghiên cứu và áp dụng để thống nhất cấu trúc dữ liệu log. Các script Vector Remap Language (VRL) được xây dựng nhằm phân loại nguồn log, parsing log thô thành các trường dữ liệu có cấu trúc, bổ sung thông tin cần thiết (enrichment) và loại bỏ các trường dư thừa. Quá trình tối ưu dữ liệu giúp giảm đáng kể dung lượng lưu trữ và nâng cao hiệu quả truy vấn trên OpenSearch.

Vai trò hướng dẫn:

- Anh Lê Phi (PopTech) góp ý về cách tổ chức luồng thu thập log mạng nhằm đảm bảo hiệu năng và tính ổn định khi xử lý lưu lượng lớn.
- ThS. Lê Anh Tuấn (GVHD) định hướng việc áp dụng chuẩn ECS, đảm bảo nội dung triển khai phù hợp với chuẩn học thuật và có khả năng mở rộng trong nghiên cứu.

Kết quả đạt được:

- Hệ thống tiếp nhận ổn định log từ nhiều nguồn khác nhau ở các lớp hạ tầng.
- Dữ liệu log được chuẩn hóa thống nhất theo ECS, hỗ trợ tìm kiếm và phân tích chéo hiệu quả.
- Dung lượng lưu trữ được tối ưu, nâng cao khả năng mở rộng của hệ thống trong môi trường thực tế.



Hình 4.6 Hình ảnh thu thập được nhiều nguồn logs khác nhau (hiển thị index)

4.4. Vận hành production và tối ưu hóa (13/10/2025 – 31/10/2025)

Mục tiêu:

Mở rộng phạm vi thu thập log từ nhiều lớp hạ tầng khác nhau, bao gồm thiết bị mạng, hệ điều hành và ứng dụng; đồng thời chuẩn hóa dữ liệu theo Elastic Common Schema (ECS) nhằm đảm bảo tính thống nhất và khả năng phân tích chéo trong hệ thống SIEM.

Nội dung thực hiện:

Luồng log từ các thiết bị mạng như Firewall Fortigate và Switch Cisco được cấu hình xuất log theo giao thức Syslog (UDP port 514). Vector được cấu hình mở cổng lắng nghe trực tiếp luồng Syslog này, cho phép tiếp nhận log mạng với lưu lượng lớn mà không cần thông qua Wazuh Manager, từ đó giảm thiểu nguy cơ quá tải tại tầng trung gian.

Đối với log hệ điều hành và ứng dụng, cơ chế thu thập thông qua Wazuh Agent tiếp tục được sử dụng. Các nguồn log bao gồm log hệ điều hành, log container Docker (json-file driver), log web server (Nginx access log) và log cơ sở dữ liệu (MySQL error log). Việc thu thập đa nguồn giúp hệ thống có cái nhìn toàn diện hơn về trạng thái vận hành và an toàn của hạ tầng.

Chuẩn Elastic Common Schema được nghiên cứu và áp dụng để thống nhất cấu trúc dữ liệu log. Các script Vector Remap Language (VRL) được xây dựng nhằm phân loại nguồn log, parsing log thô thành các trường dữ liệu có cấu trúc, bổ sung thông tin cần thiết (enrichment) và loại bỏ các trường dư thừa. Quá trình tối ưu dữ liệu giúp giảm đáng kể dung lượng lưu trữ và nâng cao hiệu quả truy vấn trên OpenSearch.

Vai trò hướng dẫn:

- Anh Lê Phi (PopTech) góp ý về cách tổ chức luồng thu thập log mạng nhằm đảm bảo hiệu năng và tính ổn định khi xử lý lưu lượng lớn.

- ThS. Lê Anh Tuấn (GVHD) định hướng việc áp dụng chuẩn ECS, đảm bảo nội dung triển khai phù hợp với chuẩn học thuật và có khả năng mở rộng trong nghiên cứu.

Kết quả đạt được:

- Hệ thống tiếp nhận ổn định log từ nhiều nguồn khác nhau ở các lớp hạ tầng.
- Dữ liệu log được chuẩn hóa thống nhất theo ECS, hỗ trợ tìm kiếm và phân tích chéo hiệu quả.
- Dung lượng lưu trữ được tối ưu, nâng cao khả năng mở rộng của hệ thống trong môi trường thực tế.

4.5. Xử lý dữ liệu ESXi, tự động hóa và cân bằng tải (01/11/2025 – 04/12/2025)

Mục tiêu:

Giải quyết bài toán xử lý log phức tạp từ hệ thống ảo hóa ESXi, đồng thời tự động hóa quy trình triển khai agent và nâng cao tính sẵn sàng của hệ thống SIEM.

Nội dung thực hiện:

Log từ hệ thống ảo hóa VMware ESXi được phân tích chi tiết, bao gồm nhiều định dạng khác nhau như Hostd, Vpxa và Rhttpproxy. Các quy tắc parsing chuyên dụng được xây dựng bằng Vector Remap Language với nhiều tầng Regex nhằm phân loại và trích xuất các thông tin quan trọng như tên máy ảo, người thao tác, trạng thái và cảnh báo phần cứng.

Trong quá trình triển khai, hệ thống phát sinh lỗi Memory Leak do một số biểu thức Regex chưa được tối ưu, gây hiện tượng tăng dần mức sử dụng bộ nhớ. Các biểu thức này được rà soát và tối ưu lại để loại bỏ hiện tượng backtracking phức tạp, qua đó khắc phục hoàn toàn lỗi trên.

Song song đó, các script tự động cài đặt agent cho Linux và Windows được xây dựng, thực hiện toàn bộ quy trình cài đặt, cấu hình và khởi động dịch vụ. Cơ chế failover

được tích hợp nhằm đảm bảo agent tự động chuyển sang máy chủ dự phòng khi xảy ra sự cố kết nối.

Hệ thống OpenSearch được mở rộng từ một node đơn lẻ sang cluster ba node, kết hợp với Nginx Load Balancer nhằm phân phối tải và đảm bảo tính sẵn sàng cao.

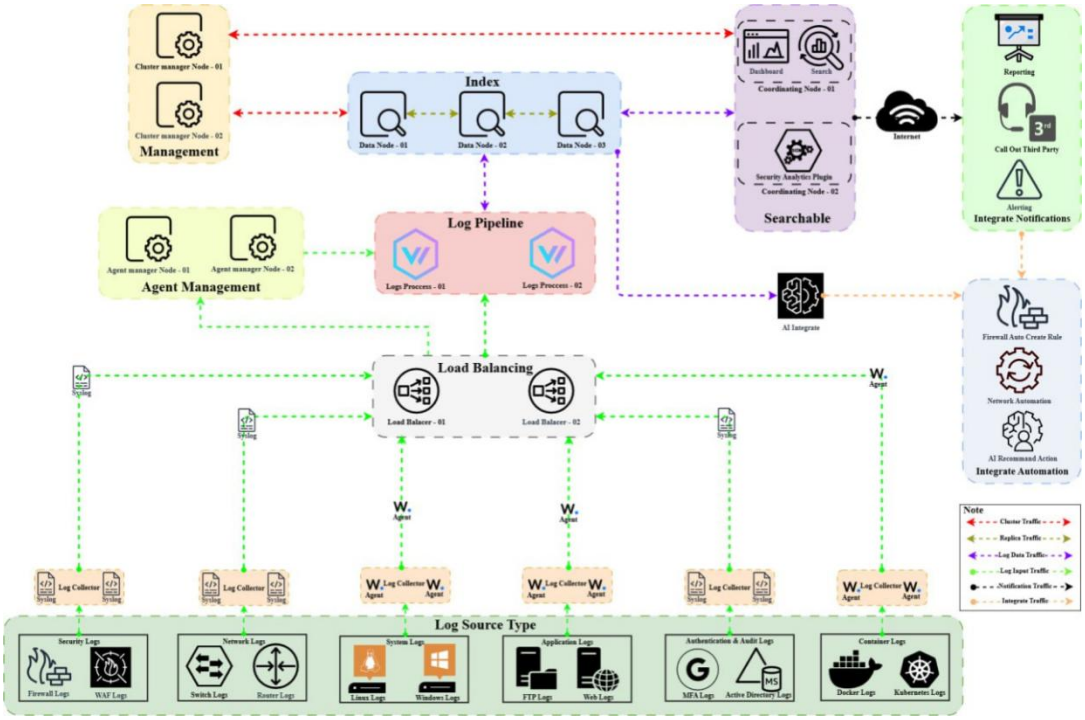
Vai trò hướng dẫn:

- Anh Lê Phi hỗ trợ phân tích log ESXi và đánh giá giải pháp parsing trong bối cảnh triển khai thực tế.
- ThS. Lê Anh Tuấn góp ý cách trình bày bài toán phức tạp và giải pháp kỹ thuật trong báo cáo.

Kết quả đạt được:

- Log ESXi được xử lý và chuẩn hóa thành công theo cấu trúc dữ liệu thống nhất.
- Thời gian triển khai agent được giảm đáng kể nhờ tự động hóa.
- Hệ thống SIEM đạt yêu cầu High Availability và khả năng chịu lỗi.

Hình 4.7 Hình ảnh 1 logs ESXI được chuẩn hóa theo ECS



Hình 4.8 Sơ đồ tổng thể triển khai cho khách hàng

4.6. Xử lý mạng phức tạp và hoàn thiện hệ thống (08/12/2025 – 18/12/2025)

Mục tiêu:

Hoàn thiện hệ thống SIEM, giải quyết bài toán thu thập log xuyên vùng mạng và chuẩn bị bàn giao hệ thống.

Nội dung thực hiện:

Đối với các máy chủ nằm trong vùng mạng DMZ hoặc VLAN bảo mật cao, giải pháp thu thập log thông qua Nginx Stream Proxy được triển khai. Proxy đóng vai trò trung gian chuyển tiếp toàn bộ lưu lượng log từ các vùng mạng bị giới hạn kết nối về hệ thống xử lý trung tâm thông qua một cổng duy nhất được firewall cho phép.

Các dashboard giám sát bảo mật và vận hành được thiết kế trên OpenSearch Dashboards, cung cấp cái nhìn trực quan về trạng thái hệ thống, các sự kiện an ninh và tình trạng tài nguyên. Quá trình đối soát dữ liệu giữa log gốc và log hiển thị được thực hiện nhằm đảm bảo độ chính xác của hệ thống.

Cuối cùng, hệ thống được bàn giao kèm tài liệu hướng dẫn vận hành, bao gồm các quy trình thêm mới agent, kiểm tra sự cố mất log và khôi phục hệ thống.

Vai trò hướng dẫn

- Anh Lê Phi hỗ trợ đánh giá khả năng vận hành thực tế và quy trình bàn giao hệ thống.
- ThS. Lê Anh Tuấn góp ý hoàn thiện nội dung tổng kết và đánh giá kết quả triển khai.

Kết quả đạt được

- Hệ thống SIEM giám sát xuyên suốt nhiều vùng mạng và lớp hạ tầng.
- Dashboard giám sát đáp ứng nhu cầu vận hành và an toàn thông tin.
- Hoàn thành kỳ thực tập với sản phẩm triển khai thực tế và sẵn sàng đưa vào sử dụng lâu dài.

Chương 5. KẾT LUẬN

5.1. Kết quả đạt được và Bài học kinh nghiệm

Sau quá trình thực tập tại Power of Passion Technology Solutions, em đã hoàn thành mục tiêu xây dựng một hệ thống giám sát an toàn thông tin (SIEM) hoàn chỉnh, từ khâu thiết kế kiến trúc đến vận hành thực tế. Các kết quả đạt được cụ thể như sau:

5.1.1. Về mặt kiến thức và kỹ thuật:

- **Làm chủ kiến trúc SIEM hiện đại:** Không chỉ dừng lại ở việc cài đặt, em đã hiểu sâu về cơ chế vận hành của OpenSearch, Wazuh và đặc biệt là sự ưu việt của **Vector** (Rust-based) trong vai trò bộ thu thập trung tâm (Aggregator) thay thế cho các giải pháp truyền thống như Fluentd.
- **Kỹ năng xử lý dữ liệu (Data Engineering):** Nắm vững kỹ thuật Parsing và chuẩn hóa log theo chuẩn **Elastic Common Schema (ECS)**. Đã giải quyết thành công bài toán khó về xử lý log phi cấu trúc của **VMware ESXi** và log mạng, đồng thời tối ưu hóa Regex để ngăn chặn lỗi tràn bộ nhớ (Memory Leak).
- **Tối ưu hóa hệ thống:** Có khả năng tinh chỉnh cấu hình chuyên sâu (Performance Tuning), ví dụ: triển khai chế độ **Wazuh Archive Mode** kết hợp tắt module cảnh báo để giảm tải, hay chuyển đổi toàn bộ pipeline xử lý sang ngôn ngữ VRL (Vector Remap Language).
- **Tư duy hệ thống và mạng:** Hiểu và triển khai được các mô hình mạng phức tạp như thu thập log xuyên vùng **VLAN/DMZ** thông qua **Nginx Stream Proxy** và đảm bảo tính sẵn sàng cao (HA) bằng **Load Balancing**.

5.1.2. Về kỹ năng mềm:

- Rèn luyện khả năng tư duy giải quyết sự cố (Troubleshooting) thông qua việc debug các lỗi crash dịch vụ trên môi trường Production.
- Nâng cao kỹ năng làm việc nhóm và viết tài liệu kỹ thuật, thể hiện qua việc phối hợp với team Parsing và xây dựng kho tài liệu hướng dẫn triển khai tự động (Automation Scripts).

5.2. Thuận lợi và Khó khăn

5.2.1. Thuận lợi:

- Được Ban lãnh đạo và người hướng dẫn (Anh Phi) tin tưởng giao phó các bài toán thực tế, cho phép tiếp cận trực tiếp với dữ liệu thật (Production Data) thay vì chỉ mô phỏng trên Lab.
- Hạ tầng công nghệ của công ty hiện đại, tạo điều kiện thuận lợi để thử nghiệm các giải pháp mới như Vector hay OpenSearch Cluster.

5.2.2. Khó khăn và Thách thức:

- **Độ phức tạp của dữ liệu:** Dữ liệu log thực tế (đặc biệt là ESXi và các thiết bị mạng) có định dạng rất đa dạng và thiếu nhất quán, đòi hỏi tốn nhiều thời gian nghiên cứu Regex và VRL để chuẩn hóa.
- **Áp lực về hiệu năng:** Việc triển khai trên môi trường Production phát sinh các vấn đề về tài nguyên (RAM/CPU) mà môi trường Lab không gặp phải (ví dụ: Wazuh Manager bị crash), buộc phải thay đổi kiến trúc và chiến lược thu thập liên tục.
- **Rào cản mạng:** Việc thu thập log từ các phân vùng mạng bảo mật cao (DMZ, VLANs riêng biệt) gặp nhiều khó khăn trong cấu hình Firewall và Proxy.

5.3. Định hướng phát triển

Dựa trên nền tảng hệ thống đã xây dựng, các hướng phát triển tiếp theo để hoàn thiện giải pháp bao gồm:

- **Nâng cao khả năng phát hiện:** Viết thêm các bộ luật (Custom Rules) và Decoders chuyên sâu trên Wazuh để phát hiện các hành vi tấn công phức tạp (APT, Ransomware) dựa trên dữ liệu log đã chuẩn hóa.
- **Tự động hóa phản ứng :** Nghiên cứu tích hợp khả năng phản ứng tự động (Active Response) để hệ thống có thể tự chặn IP hoặc cô lập máy trạm ngay khi phát hiện tấn công.

- **Tích hợp AI/Machine Learning:** Sử dụng tính năng Anomaly Detection của OpenSearch để phát hiện các bất thường trong lưu lượng mạng mà các luật tĩnh không bắt được.
- **Mở rộng khả năng quan sát (Observability):** Tích hợp thêm Metrics và Tracing (sử dụng OpenTelemetry) vào hệ thống hiện tại để giám sát toàn diện sức khỏe của hạ tầng IT.

5.4. Kết luận chung

Kỳ thực tập tại Power of Passion Technology Solutions là một bước ngoặt quan trọng trong quá trình học tập của em. Từ những nghiên cứu lý thuyết ban đầu về mô hình **All-in-One**, em đã từng bước hiện thực hóa được một hệ thống giám sát tập trung có khả năng **chịu tải cao (High Availability)**, **tự động hóa (Automation)** và **xử lý dữ liệu thông minh**.

Quá trình chuyển đổi công nghệ từ Fluentd sang Vector, hay việc giải quyết bài toán log ảo hóa ESXi, không chỉ minh chứng cho khả năng áp dụng kiến thức vào thực tế mà còn thể hiện tư duy tối ưu hóa không ngừng. Những kinh nghiệm quý báu về vận hành hệ thống SIEM, xử lý sự cố quy mô lớn và làm việc trong môi trường chuyên nghiệp sẽ là hành trang vững chắc để em tự tin phát triển sự nghiệp trong lĩnh vực An toàn thông tin và Vận hành hệ thống (DevSecOps) trong tương lai.

TÀI LIỆU THAM KHẢO

Theo chuẩn IEEE:

Tài liệu trực tuyến (Online Documentation & Technical References)

[1] OpenSearch Project, "OpenSearch Documentation - Install and configure," *OpenSearch.org*, 2025. [Online]. Available: https://opensearch.org/docs/latest/install-and-configure/. [Accessed: Dec. 10, 2025].

[2] Wazuh Inc., "Wazuh Documentation - Architecture," *Wazuh.com*, 2025. [Online]. Available: https://documentation.wazuh.com/current/getting-started/architecture.html. [Accessed: Aug. 20, 2025].

[3] Datadog, "Vector Remap Language (VRL) Reference," *Vector.dev*, 2025. [Online]. Available: https://vector.dev/docs/reference/vrl/. [Accessed: Sep. 15, 2025].

[4] Elastic, "Elastic Common Schema (ECS) Reference - Version 8.11," *Elastic.co*, 2024. [Online]. Available: https://www.elastic.co/guide/en/ecs/current/index.html. [Accessed: Oct. 01, 2025].

[5] Docker Inc., "Docker Compose networking," *Docker Documentation*, 2025. [Online]. Available: https://docs.docker.com/compose/networking/. [Accessed: Aug. 12, 2025].

[6] F5 NGINX, "TCP and UDP Load Balancing with NGINX Stream Module," *NGINX Documentation*, 2025. [Online]. Available: https://docs.nginx.com/nginx/admin-guide/load-balancer/tcp-udp-load-balancer/. [Accessed: Dec. 08, 2025].

[7] Fluentd Project, "Fluentd vs. Vector Performance Benchmark," *Fluentd.org*, 2024. [Online]. Available: https://docs.fluentd.org/. [Accessed: Aug. 18, 2025].