DS-670

CAPSTONE: BIG DATA & DATA SCIENCE

WEEK-1 ASSIGNMENT

SAI HEMANTH KUMAR SANGEPOGU

GOPICHAND VEMULA

BHAVANI GODDINDLA

9th MARCH 2025

Defining the Subject: Provide an introduction and an abstract to your chosen topic.

The given topic for this capstone is cyber-apache kafka, elk stack and wireshark. The project is a cyber security project integrating tools such as Apache kafka, ELK stack and Wireshark. The idea behind this project is to understand network security and detect anomalies. In this project we are going to use Apache Kafka to build data pipelines to handle large volumes of data, Wireshark to capture and analyse the network traffic and ELK stack to manage the network traffic data and handle big datasets and complex queries.

By combining these tools, we aim to build a SIEM system (Security Information and event management system). SIEM systems are used to detect, analyse and respond to the anomalies or security threats of a network.

Project title: Enhanced Cyber Security Monitoring: A SIEM Approach with ELK Stack, Kafka and Wireshark

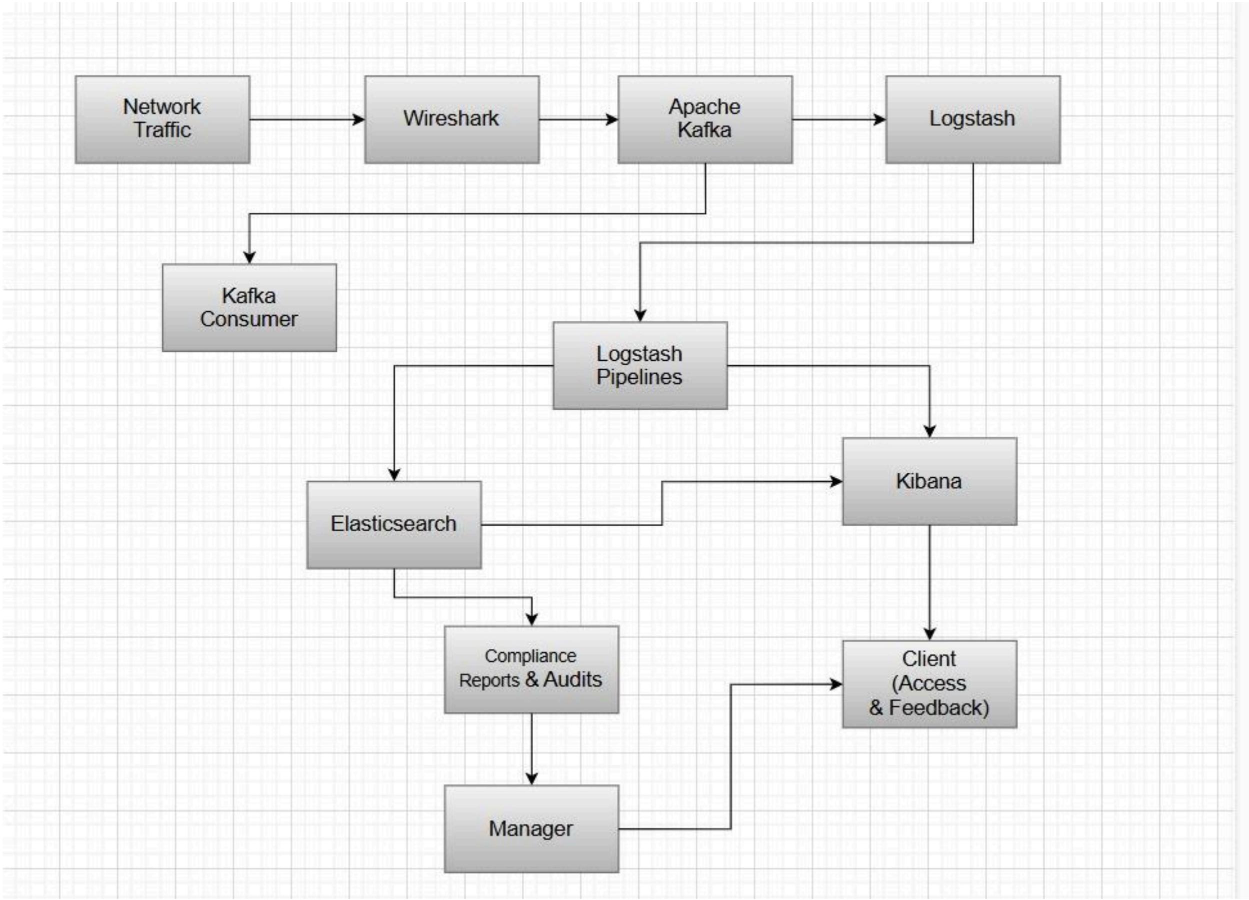
Business Need: Explain the business or societal need that your project addresses.

There are many SIEM systems and the popular ones are splunk, IBM QRadar, ArcSight, Microsoft Sentinel and AllenVault(AT&T). The SIEM system is used to help organizations to detect,analyse and respond to the anomalies and also ensure the security policies of the organization are being enforced. With the help of this system we can identify potential weaknesses and take preventive measures.

Problem Statement

We are in a digital age where many organisations are prone to cyber security threats such as data breaches and other malicious activities. It is important to monitor and manage these threats as it is necessary for the organisations to implement security measures to protect their confidential data. So we aim to develop a security information and event management system(siem system) which helps us in better understanding of the network and security and we will be able to detect, analyze and respond to the anomalies and take preemptive measures by identifying the potential weaknesses.

High-Level Architecture:



Minimum Viable Product (MVP):

Overview

To create a Minimum Viable Product (MVP) for the Enhanced Cyber Security Monitoring project, firstly we should focus on developing the core functionalities that are essential for real-time security monitoring, threat detection, and incident response.

The MVP should include the following key components:

1. Data Collection
2. Data Processing
3. Data Storage
4. Data Analysis and Visualization
5. Real-Time Alerts and Incident Response

Components of the MVP

Data Collection

Objective: Collect log data from various sources

Sources:

- System Logs (Windows Event Logs, Syslog)
- Security Logs (Firewalls, IDS/IPS, Antivirus Solutions)
- Application Logs (Web Servers, Databases, Cloud Services)
- Endpoint Logs (Desktops, Laptops, Mobile Devices)
- Network Logs (Routers, Switches, Wireless Access Points)

Tools:

- Logstash: Ingest logs from multiple sources.
- Filebeat: Lightweight shipper to collect and forward logs.

Data Processing

Objective: Filter and transform the collected data to make it suitable for analysis.

Tools:

- Logstash: Use filters to parse and transform log data.
- Kafka: Stream processing and ingestion.

Data Storage

Objective: Store the processed data in a scalable and efficient manner for quick retrieval and analysis.

Tools:

- Elasticsearch: Store and index the processed log data.

Data Analysis and Visualization

Objective: Analyze the stored data and provide visual representations of security events and trends.

Tools:

- Kibana: Create dashboards and visualizations for monitoring and analyzing log data.

Real-Time Alerts and Incident Response

Objective: Detect potential security threats in real-time and initiate incident response workflows.

Tools:

- Kibana: Set up alerting rules and notifications based on specific conditions.
- Custom Scripts: Automate incident response actions.

Workflow of the MVP

The MVP workflow begins with Data Ingestion, where logs are collected from various sources using tools like Logstash and Filebeat. These tools are essential for analysing the logs from multiple systems, including servers, applications, and network devices. To handle the volume and velocity of incoming data, Kafka is used for real-time data streaming and ingestion, ensuring that data is efficiently processed and transmitted to the subsequent stages.

Then we start with the Data Processing, the collected logs are filtered, transformed, and enriched using Logstash pipelines. This step ensures that the data is clean, structured, and enhanced which makes it suitable for analysis. Logstash configurations are tailored to handle the specific log formats and requirements of the data sources, which helps with efficient indexing and searching in the later stages.

In the Data Storage phase, the processed data is stored in Elasticsearch indices. Elasticsearch provides a robust and scalable solution for storing large volumes of log data, with powerful indexing capabilities that help with quick retrieval and analysis. Proper indexing is crucial to ensure that the data can be accessed for real-time analysis and querying.

The next step is Data Analysis and Visualization, where Kibana is employed to create interactive dashboards and visualizations. Kibana's interface allows the security analysts to monitor security events and trends in real-time, providing valuable insights to the organization's security. Visualizations help in identifying patterns, anomalies, and potential threats, enabling proactive security measures.

Finally, the Real-Time Alerts and Incident Response phase involves in setting up alerting rules in Kibana based on predefined conditions such as failed login attempts or detected anomalies. These alerts are configured to notify security analysts of potential threats, enabling immediate action. Additionally, custom scripts are developed to automate incident response actions, such as isolating compromised systems, thereby improving the efficiency and effectiveness of the security operations.

By following this workflow, the MVP of the project ensures security monitoring, threat detection, and incident response, laying a solid foundation for further enhancements and scalability.

Implementation Timeline (8 Weeks)

Week 1-2: Project Planning and Setup

- Define objectives and scope
- Set up development environment

Week 3-4: Data Collection and Processing

- Configure data collection from various sources
- Set up Logstash pipelines for data processing

Week 5-6: Data Storage and Analysis

- Store processed data in Elasticsearch
- Create initial dashboards in Kibana

Week 7-8: Real-Time Alerts and Incident Response

- Set up alerting rules and notifications
- Develop custom scripts for incident response

Final testing and deployment

By focusing on these core components and functionalities, the MVP of the project will provide a solid foundation for real-time security monitoring, threat detection, and incident response. As we gather feedback and iterate, we can enhance and expand the system to include additional features and capabilities.

References

Ahmed, A., Asim, M., Ullah, I., Zainulabidin, & Ateya, A. A. (2024). An optimized ensemble model with advanced feature selection for network intrusion detection. *PeerJ Computer Science*, *10*, e2472. https://doi.org/10.7717/peerj-cs.2472

Ariffin, M. A. M., Darus, M. Y., Haron, H., Kurniawan, A., Muliono, Y., & Pardomuan, C. R. (2022). Deployment of Honeypot and SIEM Tools for Cyber Security Education Model in UITM. *International Journal of Emerging Technologies in Learning*, *17*(20), 149–172. https://doi.org/10.3991/ijet.v17i20.32901

Calderon, G., del Campo, G., Saavedra, E., & Santamaría, A. (2023). Monitoring Framework for the Performance Evaluation of an IoT Platform with Elasticsearch and Apache Kafka. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-023-10409-2

Coppolino, L., Sgaglione, L., D'antonio, S., Magliulo, M., Romano, L., & Pacelli, R. (2022). Risk Assessment Driven Use of Advanced SIEM Technology for Cyber Protection of Critical e-Health Processes. *SN Computer Science*, *3*(1). https://doi.org/10.1007/s42979-021-00858-4

Jain, G., & Anubha. (2021). Application of SNORT and Wireshark in Network Traffic Analysis. *IOP Conference Series: Materials Science and Engineering*, *1119*(1), 012007. https://doi.org/10.1088/1757-899x/1119/1/012007

Khan, J., Elfakharany, R., Saleem, H., Pathan, M., Shahzad, E., Dhou, S., & Aloul, F. (2025). Can Machine Learning Enhance Intrusion Detection to Safeguard Smart City Networks from Multi-Step Cyberattacks? *Smart Cities*, *8*(1). https://doi.org/10.3390/smartcities8010013

Lakkad, A. K., Bhadaniya, R. D., Shah, V. N., & Lavanya, K. (2021). Complex events processing on live news events using apache kafka and clustering techniques. *International Journal of Intelligent Information Technologies*, *17*(1), 39–52. https://doi.org/10.4018/IJIIT.2021010103

Liu, J. C., Yang, C. T., Chan, Y. W., Kristiani, E., & Jiang, W. J. (2021). Cyberattack detection model using deep learning in a network log system with data visualization. *Journal of Supercomputing*, *77*(10), 10984–11003. https://doi.org/10.1007/s11227-021-03715-6

López Velásquez, J. M., Martínez Monterrubio, S. M., Sánchez Crespo, L. E., & Garcia Rosado, D. (2023). Systematic review of SIEM technology: SIEM-SC birth. *International Journal of Information Security*, *22*(3), 691–711. https://doi.org/10.1007/s10207-022-00657-9

Patil, N. V., Krishna, C. R., & Kumar, K. (2022). KS-DDoS: Kafka streams-based classification approach for DDoS attacks. *Journal of Supercomputing*, *78*(6), 8946–8976. https://doi.org/10.1007/s11227-021-04241-1

Poat, M. D., Lauret, J., & Fedele, D. (2023). Flexible visualization of a 3rd party Intrusion Prevention (Security) tool: A use case with the ELK stack. *Journal of Physics: Conference Series*, *2438*(1). https://doi.org/10.1088/1742-6596/2438/1/012040

Santos, V. F., Albuquerque, C., Passos, D., Quincozes, S. E., & Mossé, D. (2023). Assessing Machine Learning Techniques for Intrusion Detection in Cyber-Physical Systems. *Energies*, *16*(16). https://doi.org/10.3390/en16166058

Shameem, S., Venkatesh, K., Shaik, L., T N D, M., Harsha, S., & Lopes, B. R. (2024). Estimating Malware Impact on Network Traffic Analysis by Using Wireshark. In *J. Electrical Systems* (Vol. 20, Issue 7).

Sharma, A., Rani, S., & Driss, M. (2024). Hybrid evolutionary machine learning model for advanced intrusion detection architecture for cyber threat identification. *PLoS ONE*, *19*(9 September). https://doi.org/10.1371/journal.pone.0308206

Blum, D. (2020). Rational Cybersecurity for Business: The Security Leaders' Guide to Business Alignment. In *Rational Cybersecurity for Business: The Security Leaders' Guide to Business Alignment*. Apress Media LLC. https://doi.org/10.1007/978-1-4842-5952-8