Real-Time Cyber Incident Monitoring for Critical Information Infrastructure (CII) using Machine Learning and ELK Stack.

Dr. K. V. Deshpande
Department of Computer Science and Business System.
JSPM's Rajarshi Shahu college of Engineering Tathawade.
Pune, India
Kvdeshpande_comp@jspmrscoe.edu.in

Varad Pawar
Department of Computer Science and Business System.
JSPM's Rajarshi Shahu college of Engineering Tathawade.
Pune, India
vrpawar1910@gmail.com

Rutuja Bagad
Department of Computer Science and Business System.
JSPM's Rajarshi Shahu college of Engineering Tathawade.
Pune, India
rutujabagad17@gmail.com

Sanskruti Parkhe
Department of Computer Science and Business System.
JSPM's Rajarshi Shahu college of engineering Tathawade.
Pune, India
sanskrutiparkhe2011@gmail.com

Vaishnavi Thorat
Department of Computer Science and Business System.
JSPM's Rajarshi Shahu college of Engineering Tathawade.
Pune, India
vaishnavit582@gmail.com

***Abstract*— Monitoring and analyzing cyberattacks targeting critical infrastructure have become significant threats to both government agencies and organizations in this modern digital world. Such cyberattacks can potentially disrupt essential services and pose major threats to national security. There is, thus, a critical need to rapidly detect and respond to such threats. The purpose of this survey paper is to illustrate the challenges surrounding monitoring cyber threats and introduce the proposed solution - the real-time cyberattack monitoring tool.**

This tool does web scraping across various platforms along with using data from machine learning to store into a structured format through which it ensures easy accessibility by the cybersecurity team. With all this collected information, visualizing it into some interactive dashboard that would give visibility to attack trends. In addition to this, alerts are generated through patterns of frequent high-frequency attacks enabling the team members to respond sooner.

This solution is designed to improve the capability of organizations to safeguard their critical infrastructure by providing timely insights and effective response strategies. Though still in development, it has the potential to become an useful tool for addressing the growing challenge of cyber threats.

I. Introduction

Today's increasingly interconnected environment has left much of our critical systems vulnerable to cyber threats because of our heavy dependence on technology. Attacks against these essential services such as energy, transportation, and communication could severely jeopardize national security and the reliability of such services and can cause a massive economic damage as well as risks to public safety. Advanced techniques of cybercrime have increasingly seen a rising tide of incidents and hence require vigorous action from the government and the organization entrusted with safeguarding these infrastructures.

For instance, there is the recent ransomware attack on Colonial Pipeline in May 2021 that brought fuel supplies across the entire Eastern United States to a halt. This left an open chink in critical infrastructure. Again, the breach in SolarWinds exposed a severe flaw in the software supply chain, leaving a trail of damaged federal agencies and private firms. These have brought to light the urgent call for advanced monitoring systems to curtail the threat posed by cyberattacks.

Conventional cybersecurity strategies, which are often reactive and fragmented, are inadequate for handling the complexities of modern threats. Cybersecurity teams are overwhelmed by vast amounts of data, making it difficult to effectively identify and prioritize potential threats. Current systems often lack real-time insights and a cohesive understanding of the threat landscape, which hampers timely responses to emerging incidents.

This constant pace of technological innovation brings with it new vulnerabilities that cybercriminals exploit. Often, it outpaces the organizational defenses, so a proactive strategy is very important to identify threats and provide actionable insights for effective incident response.

This paper discusses the existing methods of cyber threat monitoring and proposes a new, highly efficient tool for real-time cyberattack monitoring. The tool applied here uses recent technologies like machine learning, web

scraping, and ELK Stack (Elasticsearch, Logstash, Kibana) in efforts to enhance data collection, storage, and visualization. By aggregating data from other sources online, the system categorizes information for access and analysis. It allows the use of an interactive dashboard where cybersecurity teams monitor, filter, and visualize incidents in real time. High-frequency attack patterns are detected and, accordingly, alert systems are triggered to inform teams about the need for prompt action. The proposed solution aims to bolster the abilities of organizations in protecting critical infrastructure by providing timely insight into effective incident response strategies..

II. Web Scrapping

Web scraping is an essential part of our project that provides structured collection channels through which we can gather real-time data on online media such as any type or brand's cybersecurity blog, public bulletin board system news articles forums, or social networks The essential significance of this form of data-gathering method is that it supplies vast amounts of chaotic information about events caused by cyber-attacks, letting us look for patterns and emerging trends in computer security incidents. Unlike traditional methods of data acquisition, which may require human intervention and so take time, web scraping is automated. This means that it works continuously and independently. This automated method for gathering data provides us with recent insights. Subsequent work in data preprocessing, classification, and analysis cornerstone to the successful realization of our project depends on this phase taking place as soon as is humanly possible.

The use of web scraping with tools such as Scrapy and Beautiful Soup streamlines data acquisition, not only saving time but also cutting costs. These tools' adaptability allows us to specifically target fields in the scraping process.

III. Machine Learning

The project is built on a fundamental component of AI that is Machine Learning (ML), which allows the system to learn based on history, find out patterns and predict cyber threats for future. In effect, ML allows organizations to automate things so complex or voluminous that otherwise it would be extremely difficult for us humans do manually particularly in the ever growing and shifting world of cybersecurity.

Role of ML in Project:

Identifying Anomalies:

Machine learning really excels in identifying anomalies areas where behavior deviates from the norm. Whenever there is some unusual activity, such as a surge in traffic attacks or an unauthorized access attempt; ML Becomes the first line of defense and flags these types of activities making it easy for quick decision pipelines to stop any attack before further damage.

Classifying Attack Types:

ML is used to categorize these incidents by their specific characteristics with regards to various cyberattack types, such as malware, phishing and DDoS. This allows security advance teams to prioritize alerts more easily, optimize responses and deal with the great threats first.

Predicting Future Threats:

ML models trained using historical data can predict potential threats by identifying patterns indicative of a future offense. Such forethought enables the security team to get ready and protect itself from catastrophe beforehand.

Batch Processing:

One of the biggest problems in cybersecurity is dealing with huge amounts of data- logs, network traffic dumps and user activity. ML quickly automates data processing by correctly recognizing high-level intelligence, allowing security teams to detect and respond to threats faster.

Machine learning tools:

BoW (Bag of Words): It is responsible to handling text data from logs and incident reports.

Random Forest: Classification of types of incidents with high accuracy.

K-Means Clustering: Anomaly traffic patterns and potential threats detection.

NLP (Natural Language Processing): Used to analyze unstructured text like logs, and extract valuable information.

IV. ELK Stack

ELK: The ELK Stack (Elasticsearch, Logstash and Kibana) in this project is an indispensable component of real-time cyber incident monitoring; empowering robust capabilities to process large amounts of data concerning ongoing threats within Indian cyberspace.

1. Elasticsearch:

The central search engine and storage Elasticsearch where structured cyber incident data is stored for quick analysis Figs 1,2 (Taifa et al. Distributed architecture for scale and high availability to process large amounts of real time event data It indexes data so you can do fast searches, use live threat monitoring.

2. Logstash:

Logstash is the powerhouse that moves raw data through your system, acting as a data pipeline workhorse. It scrubs and transforms the data, preparing it for Elasticsearch analysis and machine learning model interpretation, ensuring the dataset is ready for both processes.

3. Kibana:

Kibana: a visualization tool that makes interactive dashboards from data. It can be used by security teams to track patterns, monitor incidents in real time and create alerts if there is any potential threat.

Why ELK Can Use in Cybersecurity:

Our project heavily relies on the ELK Stack for monitoring and real-time response to cyber threats. Every piece of the stack has a precise role to play that together creates an efficient system configuration allowing for huge amounts of data management and querying.

A. *Figures and Tables*

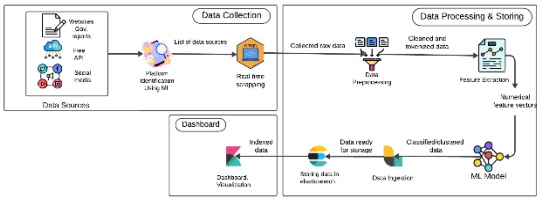

Fig. 1. Workflow DIagram

V. Comparative Analysis between Traditional and Our Approach

Feature	Traditional Approach (MongoDB / SQL)	Latest Approach (ELK Stack)
Data Storage and Structure	**MongoDB**: Schema-less, ideal for flexible, unstructured data. Horizontal scaling but can be inefficient for complex queries. **SQL**: Relational and schema-defined, powerful querying but faces challenges with horizontal scaling in large, distributed systems.	**Elasticsearch**: Distributed, indexed format optimized for large-scale, real-time querying. Handles structured and unstructured data efficiently using JSON-based documents.
Data Ingestion and Preprocessing	Data requires cleansing, structuring, and transformation before ingestion. MongoDB allows semi-processed data, but SQL demands structured data, making preprocessing rigid and time-intensive.	**Logstash (or Beats)**: Customizable ingestion and real-time transformation, reducing preprocessing needs. Faster pipeline from data ingestion to insights, especially in real-time applications.
Querying and Analytics	**MongoDB**: Flexible queries for unstructured data but may slow down with complex searches. **SQL**: Optimized for structured data but struggles with flexible, nested data structures.	**Elasticsearch**: Designed for high-speed, large-volume analytics, with robust full-text and advanced search capabilities via REST APIs and Query DSL. Real-time analytics optimized.
Visualization	Third-party tools (e.g., Tableau, Power BI) required for visualization, often leading to integration complexity and delayed insights.	**Kibana**: Native visualization tool with real-time dashboards, anomaly alerts, and streamlined integration for faster, actionable insights without additional configuration.
Scalability and Performance	**MongoDB**: Horizontally scalable but may face performance lags with large datasets and complex queries. **SQL**: Limited scalability; can encounter bottlenecks in distributed systems.	**Elasticsearch**: Optimized for horizontal scaling and distributed querying, supporting vast datasets with low-latency, real-time responses suitable for high-demand applications.

VI. Methodology

Our methodology is primarily divided into three main phases: Data Collection, Data Preprocessing

and Storing, and Data Visualization. Every phase is critical for the proper collection, analysis, and representation of cyberattack data.

1. Data Collection

The initial setup involves data gathering through web scraping. Cybersecurity discourse is largely distributed across online news, blogs, and forums; from these open-text sources, critical information can be semi-automatically extracted [5].

Web Scraping Development: A web scraper is developed to automatically gather content about cyberattacks using Python libraries like Scrapy or Beautiful Soup from cybersecurity blogs, news articles, forums, and social media channels [10]. Structured data (such as titles, timestamps, and geographic locations) and unstructured data (full-text articles, posts, and tweets) are targeted for collection [5].

Mapping: Web scrapers are employed to continuously collect data, followed by the Data Pipeline Setup with Logstash to automate ingestion into the ELK stack [8]. Logstash is used to structure and route the scraped data in real-time to the appropriate storage.

2. Data Preprocessing and Storing

Data Cleaning Missing values are resolved; outliers are identified, and noise is eliminated for the model to be accurate [7]. Mean imputation and interpolation help maintain data quality [3]. Missing Data Handling: Imputation or partial records are deleted while handling missing or incomplete fields in order to preserve data integrity [2].Feature Extraction and Classification: Cleaned data is used to extract event features like attack type, time, and region, which are classified into categories (malware, phishing, DDoS, etc.). using algorithms like logistic regression and decision trees [2]. This preps the data for analysis and visualization [1]. Classified data is kept in Elasticsearch, a structured and unstructured data system. It provides high availability and fault tolerance and deals with large quantities efficiently [8]. Preprocessed data flows through Logstash to Elasticsearch, structuring it for further analysis [8].

3. Data Visualization

Data Visualization: In the post-cleaning step, the data will be saved into readable insights and transformed into those insights to make cybersecurity professionals understand better. Kibana for Real-Time Dashboards: Kibana is part of the ELK stack and is utilized to design interactive dashboards for cybersecurity data.It can process large datasets and represent them in the form of bar graphs, pie charts, heat maps, and geographical maps because it is integrated with Elasticsearch. In this way, data analysis becomes much easier and faster. Geographical Mapping of Cyberattacks: Kibana uses the IP2LOCATION database to map the location of attacks. This way, teams are able to monitor global attack patterns, which clearly shows where the attacks are coming from and where they are targeted [3].

Kibana for Real-Time Cyberattack Monitoring: Kibana's real-time dashboard updates are key to identifying emerging threats in real-time. It also can send notifications to teams so that they can react promptly in cases of high priority incidents [4]. Trend Analysis: With Kibana's query system, teams can monitor trends because it can illustrate rising or declining attack rates with time. This is a way to keep ahead of emerging threats through early identification of patterns [1]. Attack Timeline: Teams can understand when and how cyberattacks occur across systems or regions by visualizing the sequence of attacks, providing valuable context for responding effectively [6]. Localized Alert System: Kibana can be configured to alert teams if there are repeated attacks in the same area within a short period. This is helpful in identifying and stopping coordinated attacks [4]. Conclusion: This is an easy three-step process on how to construct a cybersecurity framework by collecting data, preprocessing, and then visualizing. By combining ELK Stack, web scraping, and machine learning, the system can be able to offer real-time insights and alerts for teams to react faster and generally improve their defenses in cybersecurity.

VII. Conclusion:

In today’s digital world, cyber threats are becoming a bigger concern, especially for governments and organizations that depend on critical infrastructure. To address this, we’ve built a real-time monitoring and visualization tool that brings together web scraping and machine learning. This system constantly gathers data from various online sources, processes it efficiently, and stores it—all in real time.

With the Kibana dashboards, security teams can quickly detect active patterns of attacks. Therefore, teams are able to react before it is too late. The system has an in-built alert feature, which allows the teams to know whenever a new risk develops and sends out alerts.

All in one, this would make threat detection faster and lead to quicker response times, better protecting the operation of these vital systems against disruptions. Organisations can thereby stay ahead of evolving cyber threats as well as remain resilient in critical operations.

VIII. Literature Survey

Lughbi, H., Mars, M. & Almotairi K.(2024) NLP-based Cyber attack Tweets Classification and Visualization Real-Time Dashboard The tool makes use of NLP to provide better situational actionability in order to proactively deter cybersecurity threats by offering interactive data visualizations which can easily be integrated with the integration capabilities present due to its usage of natural language processing for real-time threat detection.

To address these limitations, Ghasiya and Okamura (2022) proposed a hybrid approach that integrates information extraction with sentiment analysis to analyze cybersecurity news articles. This paper utilizes sentiment analysis to measure public opinions on cybersecurity issues and applies TF-IDF for term identification as well as word embeddings (W2V)to recognize the context. This makes the study supportive of hybrid NLP to understand public sentiment, a further useful approach for strengthening cyber security frameworks.

Model and methods Iorga et al. [3] introduced an OSINT-based model with machine learning for determining cybersecurity vulnerabilities in the case of news sources Given the successful performance of a fully tuned BERT model that accurately alerted on threats, this study shows promise as an early threat detection mechanism using NLP and machine learning — something in line with further exploiting AI for preemptive cybersecurity means.

Security related work for the ELK Stack: An approach to harden its resilience against cyberattacks' published by Vadhil, F.A., Salihi, M.L. and Nanne, P.F (2019). These security fixes serve as critical knowledge for your project intelligent use of ELK in data visualization and threat analysis, promoting log management systems while keeping them efficient through identifying ways to secure the ELK Stack.

One such project that we can refer is the one developed by Chaudhari, S., Maurya, V., Singh, V., Tomar,S., Rajan,A.. & Rawat A.(2020) which implemented a real-time log and traffic monitoring capturing network data to identify any abnormalities or potential threats. Real-time visualizations provided by the system help security teams quickly analyze traffic patterns, which enhances cybersecurity response efficiency and corresponds with real-time monitoring focus in your project.

Baykara, M., Gurturk, U. and Das R. (2018) A comprehensive analysis of real time cyber attack monitoring tools Those findings provide insight into the benefits these tools bring to cyber security frameworks, particularly for dynamic threat detection environments.

IX. References

1. Lughbi, Huda, Mourad Mars, and Khaled Almotairi. "A Novel NLP-Driven Dashboard for Interactive CyberAttacks Tweet Classification and Visualization." Information (2078-2489) 15, no. 3 (2024).
2. Gonaygunta, Hari. "Machine learning algorithms for detection of cyber threats using logistic regression." *Department of Information Technology, University of the Cumberlands* (2023).
3. Deshpande, K. V., Shubham Asbe, Akanksha Lugade, Yash More, Dipali Bhalerao, and Anuradha Partudkar. "Learning Analytics Powered Teacher Facing Dashboard to Visualize, Analyze Students' Academic Performance and give Key DL (Deep Learning) Supported Key Recommendations for Performance Improvement." In *2023 International Conference for Advancement in Technology (ICONAT)*, pp. 1-8. IEEE, 2023.
4. Ghasiya, Piyush, and Koji Okamura. "A hybrid approach to analyze cybersecurity news articles by utilizing information extraction & sentiment analysis methods." International journal of semantic computing 16, no. 01 (2022)
5. Staves, Alexander, Tom Anderson, Harry Balderstone, Benjamin Green, Antonios Gouglidis, and David Hutchison. "A cyber incident response and recovery framework to support operators of industrial control systems." International Journal of Critical Infrastructure Protection 37 (2022): 100505.
6. Khder, Moaiad Ahmad. "Web scraping or web crawling: State of art, techniques, approaches and application." International Journal of Advances in Soft Computing & Its Applications 13, no. 3 (2021).
7. Iorga, Denis, Dragos Corlătescu, Octavian Grigorescu, Cristian Săndescu, Mihai Dascălu, and Razvan Rughiniș. "Early detection of vulnerabilities from news websites using machine learning models." In 2020 19th RoEduNet Conference: Networking in Education and Research (RoEduNet), pp. 1-6. IEEE, 2020.
8. Chaudhari, Swati, Vinod Maurya, V. Singh, S. Tomar, Alpana Rajan, and A. Rawat."Real time logs and traffic monitoring, analysis and visualization setup for IT security enhancement." In 5th International Conference on Next Generation Computing Technologies (NGCT-2019). 2020.
9. Vadhil, Fatimetou Abdou, Mohamed Lemine Salihi, and Mohamedade Farouk Nanne. "Toward a Secure ELK Stack." International Journal of Computer Science and Information Security (IJCSIS) 17, no. 7 (2019): 139-143.
10. Baykara, Muhammet, Ugur Gurturk, and Resul Das. "An overview of monitoring tools for real-time cyber-attacks." In 2018 6th International Symposium on Digital Forensic and Security (ISDFS), pp. 1-6. IEEE, 2018.
11. Lotfi, Chaimaa, Swetha Srinivasan, Myriam Ertz, and Imen Latrous. "Web Scraping Techniques and Applications: A Literature."