International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1009-1014 www.ijaem.net ISSN: 2395-5252

An Assessment of Big Data Analytics for Cyber-threats Detection

N.Sivashanmugam, TNOU

Date of Submission: 25-08-2025 Date of Acceptance: 05-09-2025

ABSTRACT: Traditional infiltration detection systems separate the alarm and focus on low-level dangers. Due to several alerts obtained each day. Since human users should make a lot of effort, it is almost difficult to fully examine each alarm message. Analysis of historical data and looking for abnormalities that depart from the ideal detect discrepancy. The advantage is that it can identify unknown attacks. Many systems to detect discrepancies depend on data mining methods. However, these abilities rarely live with various forms of the attack and technologies that develop rapidly. However, keeping in mind human aspects in the identity of discrepancy, we get a chance to increase the current algorithm and provide better results. Snort is the actual industry standard and a reliable, proven system technology. In previous research, snort log data was not used to compare methods of detecting various discrepancy. Snort log data analysis software is already widely available; However these programs are purely visualization tools and do not use data mining techniques. Using Big Data Analytics, heteMSD is an outline to identify targeted cyber attacks. The name of the recommended framework is asymmetrical multi -level data. There should be a strong structure that can assist security analysts to reduce the blindness of data analysis from several data sources without reducing the level of digital safety assurance. A correlation engine can reduce the alert volume, while analyzing a log resource by grouping multiple warnings, which is a part of the ongoing attack. Alert threading is the word for this process. In the case of odd log resources, a correlation engine should be able to determine whether the report from many logs is related to the same incident.

Keywords: Intrusion Detection, SNORT, Hete MSD, Cyber-Security, Attack, Threat, Big Data

I. INTRODUCTION

All forms of computer network attacks are based on a set of common concepts that apply to physical security violations as well as computer safety violations. Can we identify the pattern of misuse using this information to identify the behavioral model of discrepancies? The question is that the purpose of this study is to solve. It is necessary to away from reinforced defense to a paradigm and a people's issue towards cyber security.

Big data analytics are applied to increase cyber security. Businesses can use large data and data analytics to determine what is "normal" and, depending on the results, to tighten cyber security standards. After researching an attack, Big Data Analytics enables businesses to see how an attacker can reach his system using data analytics and machine learning. Regular data analysis increases the ability to identify attacks before being identified manually. Computers can often do complex analysis often in real time, for guaranteeing the safety of your system. They search through many different data sources. From user actions and network events to server and application logs.

Big data analytics is rapidly expanding in cyber security networks as a result of business executives placing a high priority on the rapid and accurate identification of contemporary cyber security threats. Massively vast amounts of data may be quickly handled thanks to the use of big data analytics in cyber defense. In turn, this enables the early identification of weaknesses and abnormalities, significantly increasing overall resilience. The standards of corporate intelligence and cyber security networks are improved by big data analytics' statistical, machine learning, predictive modeling, and computing capabilities.

Historical data extracted from a wide range of sources, cyber security analysts and defense engineers can create statistical models or AI-based algorithms. Experts are able to swiftly identify weaknesses by setting up a baseline for typical activities. As a result, it can be claimed that using big data analytics in cyber security networks has enabled analysts to anticipate cyberattacks. Big data analysts are now able to identify deviations from the norm in order to anticipate impending assaults by combining technology and cutting-edge

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1009-1014 www.ijaem.net ISSN: 2395-5252

solutions like artificial learning, data mining, machine learning, natural language processing, and statistics.

Big data analytics offers automated monitoring and threat detection solutions that enable continuous environment monitoring and real-time detection.

It is important to provide credentials to authorized users as more organizations experience insider -hazards and careless cyber security issues. Automatic monitoring user fulfills this concern by gathering and evaluating behavior information. Automatic monitoring immediately creates an alarm in the event that any odd or potentially dangerous action is discovered. The most recent game changing invention, such as security information and incident management (SIM) system, originated with an intrusion detection system (IDS). In fact, Siem is still developing due to the unmatched power of machine learning, which helps to manage cyber safety analysts and cyber defense professionals, manage uncomfortable data sets more effectively and accelerates consolidation, correlation and insight.

All firms, no matter how big or low it is, can use Big data analytics solutions to improve their cyber security. Businesses of all sizes can now change traditional safety equipment with large data analytics in cyber defense, which are attempts to remove cyber security challenges thanks to machine learning and AI, real-time detections, automatic monitoring and data intelligence.

SNORT There are universal principles that underlie all types of computer network attacks; these principles apply to both computer security breaches and physical security breaches. Can this information be used to spot behavioral models of anomalies where we can spot patterns of abuse? This calls for a paradigm change to see cybersecurity as a people problem rather than a matter of fortified defense.

Network intrusions are a new societal issue brought on by the development of computer technology. Today's internet has a growing selection of tools and techniques that can be used to break into and attack private networks. Network intrusions are growing more common and more significant, making them a truly sensitive matter at all levels, including the government, small businesses, and personal life.

To safeguard networks and computers, there is a high requirement for efficient tools and detecting techniques. There are several Intrusion Detection Systems (IDS) for networks that have been created and are usable (Rehman 2003). There are generally two types of IDS: abuse detection systems and anomaly detection systems. The majority of commercial systems adopt mistreatment tactics that recognise common sorts of incursions. These are additionally known as signature-based invasions.

One such well-known signature-based detection system that has seen extensive application is SNORT. Researchers are now concentrating on using data mining and social network algorithms to analyze the alert records. These articles describe several intrusion detection techniques that the authors have developed. They also reveal hidden patterns that are not apparent from simply analyzing system communications.

One of the most widely used IDS systems worldwide is SNORT. More than 400, 000 registered users are utilizing SNORT to secure their systems, claims the website www.snort. org. SNORT is a robust and established system technology and is the de facto industry standard. SNORT log data was not used in earlier studies to evaluate different anomaly detection techniques. There are various software programmes already available to analyze SNORT log data, however these programmes are merely visualization tools and do not contain any data mining methods.

Understanding IDS's setup and the data in the alarm log it produces is one of the main challenges of employing IDS alerts. Sometimes the information about log variables in IDS user manuals is sparse or spread out over several chapters. Although SNORT has a very thorough user manual, it might be challenging to locate all the log variable information because it is dispersed over many chapters.

Parsing the log data into a format that may be used is a significant problem to solve. SNORT creates alarm log data in text format, which makes it difficult to utilise as a database for analysis. Reading the text data into a database is the initial step in parsing these text alerts. This work is difficult in a number of ways. The text data is semi structured, to start. Each alert has a different set of data pieces or an alert structure. Therefore, when we import the whole data into a database, we must extract all of the missing variables and account for them for other entries. Second, each warning or data record is divided into numerous lines rather than being contained on a single line.

So, in order to read all the components of the same record into a single data record, we identify patterns. Third, the names of the data variables are part of the records and may be the same across alerts. As a result, we must be aware of them and transpose the variable names into the header data rather than the data lines. Fourth, there

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1009-1014 www.ijaem.net ISSN: 2395-5252

is a lot of data. Therefore, we handle them using database approaches. An attacker will select a target in a cyberattack before starting an attack.

An attacker will need to attempt more than once and switch up their tactics till they are successful or give up, before the attack succeeds or fails. This describes the nature of cyberattacks, which are essentially all of the same kind. This presumption eliminates assaults by insiders because they are already familiar with the system and do not need to scan it or try several unauthorized access techniques.

So, in order to read all the components of the same record into a single data record, we identify patterns. Third, the names of the data variables are part of the records and may be the same across alerts. As a result, we must be aware of them and transpose the variable names into the header data rather than the data lines. Fourth, there is a lot of data. Therefore, we handle them using database approaches.

The sensors of the network must be positioned inside the network to gather data that is intended to uncover such assaults in order to detect insider or expert attacks when there are no scanning or probing stages. We see these assaults as having been effective in obtaining the required access after the first hurdle. In this study, the more specific characterization of assaults on the obtaining access stage was investigated. The destructive phase, in which successful attackers attempt to either take the information or disrupt the network, was not our main emphasis.

In most cases, a cyberattacker won't be successful on their first attempt. To access the target, the attacker will use a variety of techniques. The target address is often unique or scarce when an attack occurs. Attackers will keep going for the special targets whether they succeed or fail. A target IP address is being attacked if it receives access from a disproportionately large number of distinct IP addresses in a brief period of time.

Massive efforts made in a short period of time to learn the password are one of the most popular attack strategies. Therefore, if a single or a small number of IP Sources are sending a lot more traffic to an IP Target than usual, the IP Target is under assault.

Attackers often do not carry out slow and persistent attacks manually once they have been launched. They'll create robots or software to automate these processes. These automated programmes will have characteristics in common that might be used as hints when designing models to find such assaults. The length of time between such strikes is one of their characteristics. These attacks pass for regular site visits, yet they happen at precisely the same intervals over a longer duration.

Host detection and network detection are two categories under which intrusion detection technology is currently being researched. The malicious code is represented by the host-based intrusion detection approach, which analyzes process behavior to discover the payload. In order to recognise unidentified network intrusions, network-based intrusion detection primarily monitors network flow. A more thorough understanding of security is provided by the combination of threat intelligence with the big data analysis technique of enormous logs and traffic data.

1) Heterogeneous multi source security data is challenging and has a greater variety of expression semantics. For diverse data sources, there are several attack detection techniques. However, it is not quite evident how the research questions combine. Academic research is deficient in structured discussion.

2) Network attack-related abnormalities cannot be found fast and effectively using current approaches. Data correlation is still a challenging topic. The advancement of targeted cyberattacks detection is still hampered by the data association approach in heterogeneous multisource.

3) The automation and intelligence of the current detecting technologies are unsatisfactory. Secure data semantic information is not well conveyed. Attack recognition still primarily relies on manual analysis.

The suggested methodology makes use of correlation analysis for attack investigation, which effectively handles the detection of targeted cyberattacks in a huge data setting. This concept serves as the foundation for a suggested innerlayer and cross-layer analytical strategy for targeted cyberattacks. On the basis of this, follow-up researchers can do more study.

HeteMSD is a Big Data Analytics Framework for Targeted Cyber-Attacks Detection that Uses Heterogeneous Multisource Data is the name of the suggested framework. In order to minimize the blindness of data analysis from diverse data sources without lowering the degree of digital security assurances, there needs to be an effective framework that can help security analysts. There are still some notable differences between the study on targeted cyber-attacks detection based on heterogeneous multisource data, and both practical and theoretical efforts for this purpose are still in the exploratory stage. The relevance of our

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1009-1014 www.ijaem.net ISSN: 2395-5252

suggested approach lies in its goal to resolve the tensions between sophisticated network attack defense and heavy diverse data loads. The following are the primary contributions of our study.

Targeted cyber-attacks are a subset of devoted assaults that are directed at a particular person, business, or organization with the goal of achieving a specific goal, such as stealing confidential information from a back-end database or disrupting system functionality. A successful targeted cyberattack method often consists of acquiring information, infecting targets, exploiting systems, stealing data, and keeping control. These actions are all crucial components of focused cyberattacks. Targeted cyberattacks require the achievement of all the aforementioned phases in order to be implemented. Attackers take more time selecting their targets, looking for security holes, and creating unique malware. Targeted cyber-attacks are usually implemented by professionals instead of simply using attack tools. Another idea, known as an "advanced persistent threat, " must be brought up when discussing targeted cyber-attacks. A targeted cyberattack may be thought of as a subclass of APT. APTs are typically targeted cyberattacks using more sophisticated attack techniques. APT is implemented using a variety of different attack routes. It has been around for a while without being found in a real network setting.

Antivirus software and HIDS are the representations of the host-based detection technique. Malicious programmes are often found by monitoring system calls, network access, file operations, process formation, and memory modifications. Malicious programmes may be identified by static and dynamic analysis, and APT assaults can be stopped..

Network-based detection technique: Malware's pattern of command and control channel exhibits certain regularity (e. g., attack payload signature, network communication sequence characteristics, and created domain name). Network traffic produced by targeted cyberattacks is distinct from that produced in typical office settings. As a result, using the network detection approach, it is possible to determine the attack load of the attacking process.

Multisource data fusion can be used to achieve the correlation between events. Results of anomalous detection indicate event correlation for raw input data. Through complementarity, data fusion may lessen data duplication and cooperative information gathering. Since the formats of the original heterogeneous data are inconsistent, the characteristics must be extracted. The comprehensiveness connection of data from several sources is used to provide the overall anomalous outcomes. The primary approach of heterogeneous event fusion used nowadays is to create a global feature vector by extracting features from various security data. A worldwide abnormality that cannot be conveyed by a single data source might be better reflected by the interaction between many data sources. In comparison to single source data, multisource heterogeneous data requires a distinct approach to anomaly detection. There are three different types of multisource heterogeneous data anomaly detection methods currently available. The global anomaly is determined following an analysis of the results of the application of anomaly detection algorithms to various data sources. Second, several data sets are combined to create a single data source that has the same data pattern. In this approach, the classic singlesource data anomaly detection problem is created out of the multisource data anomaly detection problem. Thirdly, more data sources are used in the anomaly detection process to reinforce and support the findings.

Alert-Intrusion Correlation

Three types of security mechanisms offered by computer protection - naturally, are designed to protect a system - authority and auditing. To protect the systems against the attack, these three processes are necessary. However, an additional layer of protection requires if the concept and execution of these techniques is flawed.

ID has been suggested to add another line of defense. With the massive use of both commercially sponsored and open source components, the technique of detecting infiltration is becoming more and more popular in the workplace network. It has some flaws, however, including the trend for vigilant floods, relevant issues brought by attacks that are likely to produce many related alerts, false alerts and scalability. The correlation is suggested as a solution to these flaws. However, it is not clear that the choice of a safety administrator is required to avoid potential additional benefits from a variety of devices, in which the report applies to the same or separate events. This problem inspired the researcher to examine the relationship between the infiltration detection sensors and alerts generated by diverse logs.

A multi-step procedure called alert correlation takes as input alerts from one or more IDS and outputs a high-level description of the malicious behavior on the network. Data must be

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 08 Aug. 2025, pp: 1009-1014 www.ijaem.net ISSN: 2395-5252

gathered from a variety of sources (such as firewall, web server logs, IDS from various manufacturers, and so on) in order to obtain effective identification. The most evident advantage of correlation of warnings generated by diverse log resources is the reduction in the number of alerts that a security officer must deal with.

By organizing several warnings that are a part of an ongoing assault, a correlation engine can minimize alert volume while only evaluating a single log resource. This technique is known as alert threading. A correlation engine should be able to tell whether reports from different log resources pertain to the same occurrence in the situation of heterogeneous log resources. The degree to which one or more characteristics or measurements on the same set of elements exhibit a propensity to fluctuate simultaneously is known as correlation. Correlation can improve detection abilities and provide a fuller picture of threats that a single sensor or device may only be able to monitor in part without losing the security-relevant data. Correlation can also take use of the supplementary coverage provided by various log resources.

Reports from various log resources using various analysis methods may support one another and so increase the confidence in the detection.

An alert will be compared to all other alert threads that have comparable properties or features (such as source IP address, destination IP address, or ports), and if a match is found, alerts with a high degree of feature similarity will be correlated; if none are, a new thread will be formed. To group alerts that are a part of the same ongoing assault, lowlevel events are aggregated in the first phase utilizing the ideas of attack threads. If there is an attribute overlap, the alerts are clustered, which means that only attributes that are present in both alerts are taken into account. The goal of this phase is to offer a higher-level perspective of the security condition of the system by combining warnings that indicate various attack actions.

Analysis

These programs can be roughly divided into two categories: discrepancy-based identity and misuse or signature-based identity. With a signature-based identity, the installed attack signature pattern is saved and used to identify new attacks with similar signature. The important defect is that it is unable to identify new attacks with unknown signatures. The snort mainly focuses on loan wolf attacks unlike group attacks. While our algorithms identify potential risks based on the support patterns of the attackers, it classifies visits as potential hazards based on travel types and behaviors. The human variable is taken into consideration by identifying these patterns. The snort can only analyze a trip at a time, while our models can collect data from several time periods to get more accurate references. The information collected by our models can be added to a real - time intrusion detection system, so that additional can improve the ability to identify the potential threats.

For the alert-strings correlation, all alert correlation mechanisms should be used in the alert correlation process according to capacity requirements. Known and unknown attacks, as well as the method of identifying multi-step attacks, should be made even better, as these capacity criteria will help solve the problem of a significant number of false alerts.

II. CONCLUSION

To analyze alert records, researchers are now focusing on employing data mining and social network algorithms. These publication underlines many of the methods of detecting the original intrusion of authors. They also appear to hidden patterns that will not be seen only by looking at the system communication.

SNORT is one of the most commonly used IDS systems globally. According to the website www.snort. ORG, Snort is being used by more than 400, 000 registered users to protect the system. Snort is the actual industry standard and a reliable, proven system technology. In previous research, snort log data was not used to compare methods of detecting various discrepancy. To analyze Snort log data, there are currently several software programs available, although these programs are only visualization tools and are actually no data mining methods.

III. FUTURE RESEARCH

We will need to use research from many areas, such as new techniques, data mining techniques, attacker psychology, specific user behavior, etc. to identify the attacks more accurately and build a strong identity system. A strategy will never be enough to handle all landscapes. Despite the complexity of the infiltration pattern and the unexpectedness of human behavior, we will work to identify the attacks quickly and cheaply.

HeTemsd is still required to allow for the integration of the human expert and requires the characteristics, which besides being a simple observer, the framework has suggested the necessary expertise to improve the initial analysis. There is a need to detect discrepancy based on

multisource data fusion. The proposed expansion of correlation approach will be the subject of further study. Last but at least, the security check is expected to focus a huge focus on the argument of safety knowledge in the near future.

It is necessary to study further on infiltration alert correlation technique to detect unknown attacks using an anomaly and misuse detection approach.

REFERENCES

[1]. Y. Li, W. Dai, J. Bai, X. Gan, J. Wang, and X. Wang, ―An intelligence-driven security-aware defense mechanism for advanced persistent threats, ‖ IEEE Transactions on Information Forensics and Security, vol.14, no.3, pp.646–661, 2019.

[2]. J. Navarro, A. Deruyver, and P. Parrend, ―A systematic survey on multi-step attack detection, ‖ Computers & Security, vol.76, pp.214–249, 2018.

[3]. Y. Liu, M. Zhang, D. Li et al., ―Towards a timely causality analysis for enterprise security, ‖ in Proceedings of the Network and Distributed System Security Symposium, 2018.

[4]. R. Zuech, T. M. Khoshgoftaar, and R. Wald, ―Intrusion detection and Big Heterogeneous Data: a Survey, ‖ Journal of Big Data, vol.2, no.1, pp.1–41, 2015.

[5]. J. Navarro, V. Legrand, S. Lagraa et al., ―HuMa: A multi-layer framework for threat analysis in a heterogeneous log environment, ‖ Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface, vol.10723, pp.144–159, 2018.

[6]. P. Bhatt, E. T. Yano, and P. Gustavsson, ―Towards a framework to detect multi-stage advanced persistent threats attacks, ‖ in Proceedings of the 8th IEEE International Symposium on Service Oriented System Engineering, SOSE 2014, pp.390–395, IEEE, UK, April 2014.

[7]. Kim, S. J., & Hong, S. (2011). Study on the development of early warning model for cyber attack. In 2011 International Conference on Information Science and Applications (ICISA) (pp.1–8). IEEE.

[8]. Namayanja, J. M., & Janeja, V. P. (2013). Discovery of persistent threat structures through temporal and geo-spatial characterization in evolving networks. In IEEE Intelligence and Security Informatics (ISI).

[9]. Youssef, A., & Emam, A. (2012). Network intrusion detection using data mining and network behaviour analysis. International Journal of Computer Science & Information Technology, 3.6, 87–98.