

**Academic Year:** 2021-2022

# 

## Abstract

Since the beginning of SOC around 16 years ago, its significance has expanded substantially, particularly in the last five years. This is mostly due to the critical need to prevent significant cyber disasters, which has resulted in implementing centralized security operations. Despite their popularity, extant scholarly study on the subject lacks a widely acknowledged point of view and concentrates on bits rather than the whole picture. These flaws exacerbate the situation of innovation. This research describes the fundamental concepts of SOC and the specifics of a practical solution that uses the Splunk tool for real-time cyber threat detection and remediation. We set up the lab on Splunk; Machine Learning Algorithms can detect anomalies from system logs. We were impressed with the accurate predictions of the Splunk machine learning linear regression algorithm.

## Keywords

*Cyber security; Security operation center; Cyber defence; threat identifications.*

## Acknowledgements

I might want to initially offer my genuine thanks to my manager, Dr. Issa Traore, for his persistent help and inspiration for me to seek after my examinations and exploration at the……. I'm incredibly grateful to Dr…., who gave me an open door as his research understudy. It could never have been feasible to direct this exploration without his nonstop consolation and great mentorship.

I might likewise want to thank Dr. for serving on my administrative council.

At long last, I owe a profound appreciation to my folks and companions for continuously being there and giving me persistent consolation throughout my long stretches of study.

## Table of Contents

Contents

[Abstract 2](#_Toc113623928)

[Keywords 2](#_Toc113623929)

[Acknowledgements 3](#_Toc113623930)

[Table of Contents 3](#_Toc113623931)

[Table of Figures 4](#_Toc113623932)

[List of Abbreviations 5](#_Toc113623933)

[CHAPTER 1: Introduction 7](#_Toc113623934)

[Types of SOC 10](#_Toc113623935)

[What are the various kinds of SOC reports? 10](#_Toc113623936)

[1.1 Aims 11](#_Toc113623937)

[1.2 Objectives 11](#_Toc113623938)

[1.3 Artefact Description 12](#_Toc113623939)

[1.4 Thesis Structure 12](#_Toc113623940)

[CHAPTER 2: Literature Review 14](#_Toc113623941)

[Predictive model with standard SIEM Applications. 22](#_Toc113623942)

[Gaps in literature 23](#_Toc113623943)

[CHAPTER 3: Research Methodology 24](#_Toc113623944)

[3.1 Research Methodology: 24](#_Toc113623945)

[3.2 How Log data flow in SOC 24](#_Toc113623946)

[3.4 ML used in Soc 29](#_Toc113623947)

[3.3 Project Management 31](#_Toc113623948)

[CHAPTER 4: Design / Development / Implementation 32](#_Toc113623949)

[4.1 Designing 32](#_Toc113623950)

[1. Splunk 32](#_Toc113623951)

[2. SIEM develops tools on machine learning techniques 33](#_Toc113623952)

[3. How ML Affects Security Operations Centers: 33](#_Toc113623953)

[2. Implementation 34](#_Toc113623954)

[Static implementation 34](#_Toc113623955)

[ML model Implementation using splunk 39](#_Toc113623956)

[CHAPTER 5: Results 45](#_Toc113623957)

[CHAPTER 6: Discussion / Conclusions 48](#_Toc113623958)

[6.1Discussion 48](#_Toc113623959)

[6.2 Conclusion 48](#_Toc113623960)

[References 49](#_Toc113623961)

[Appendices 55](#_Toc113623962)

[Appendix 1 55](#_Toc113623963)

## Table of sssssss

[Figure 1 organization to deploy soc 11](#_Toc113230781)

[Figure 2 Splunk 26](#_Toc113230782)

[Figure 3 dashboard creation 27](#_Toc113230783)

[Figure 4 dashboard type 28](#_Toc113230784)

[Figure 5 custom dashboard look 28](#_Toc113230785)

[Figure 6 search auth failed 29](#_Toc113230786)

[Figure 7 search internal audit log of the system 30](#_Toc113230787)

[Figure 8 extend features 30](#_Toc113230788)

[Figure 9 save the model in Soc dashboard 31](#_Toc113230789)

[Figure 10 dashboard of internal logs 33](#_Toc113230790)

[Figure 11 graph of internal logs 33](#_Toc113230791)

[Figure 12 Splunk Processing Language 33](#_Toc113230792)

[Figure 13 Splunk ML toolkit 34](#_Toc113230793)

[Figure 14 New experiment 34](#_Toc113230794)

[Figure 15 given title 34](#_Toc113230795)

[Figure 16 prediction search 34](#_Toc113230796)

[Figure 17 logistic regression model 35](#_Toc113230797)

[Figure 18 ML result 35](#_Toc113230798)

[Figure 19 ML SPL 36](#_Toc113230799)

[Figure 20 spl suspected login result. 36](#_Toc113230800)

[Figure 21 ML page 36](#_Toc113230801)

[Figure 22 Alert setting 37](#_Toc113230802)

[Figure 23 edit time scheduling for ML training 37](#_Toc113230803)

[Figure 24 Report schedule screenshots 39](#_Toc113230804)

[Figure 25visualizations 39](#_Toc113230805)

[Figure 26 statistics analysis 40](#_Toc113230806)

[Figure 27 ML login result 40](#_Toc113230807)

[Figure 28 Accuracy result 41](#_Toc113230808)

## List of Abbreviations

|  |  |
| --- | --- |
| LR | Linear regression |
| SIEM | Security Information and Event Management |
| VPN | Virtual private network |
| SOC | Secure operation center |
| ML | Machine learning |
| DDOS | Distributed Denial of Service |
| OS | Operating System |
| SPL | Splunk Processing Language. |
| UEBA | User and Entity Behaviour Analytic |
| IT | Information technology |
| SEM | Security Event Management |
| CIA | Confidentially integrity and availability |
| SIM | Security Information Management |

## CHAPTER 1: Introduction

Experts in SOC discover and triage time-sensitive security threats which require their total concentration. The commonness of false-positive alerts from myriad data sources is a critical concern. Reviewers working in SOC must study these benign alarms, low in influence but large in number, lowering SOC ability. These warnings need some examination, but not an analyst's inventive and costly mind. The use of machine learning models to forecast the maliciousness of such assaults and hence potentially identify false positives ahead of time might minimize analysts' efforts.

Analysts, however, cannot rely only on machine learning systems for such mission-critical activities, mainly because there are other options. There are frequently new sorts and sources of assault. Despite having a deployed, & to create an authentic ML model, mortal specialists must make and explain the ultimate decision and use the loop. (Al-Abassi, 2020).

The year 2017 was defined by the most severe cyber-attacks & more outstanding security breaches. For example, this week, personal credit score business Equifax announced that hackers gained access to up to 145.5 million user account credentials. The security breach began in July, as well as the information that stole included the identities, account numbers, vehicle license numbers, and banking information of around 5 million persons.

This breach may have resulted in millions of fines, but the price they will incur for the most crucial aspect of the business is consumer trust and brand, which they may have developed over time (Marcus, 2018).

Incidents like these have helped businesses recognize that perhaps the threat landscape has changed rapidly, with new issues arising daily. Organizations must shift their protection approach away from low-level majors and ad hoc responses and toward more comprehensive or resilient processes (Syed, 2019).

To recognize information compromises, companies must be aware of possible vulnerabilities and threats and be able to identify incidents at an early stage for faster response and recovery. There is no doubt that establishing a Security Operations Center is the most efficient approach to strategically supervise, analyze, and administer a company's defensive security strategy (Faulkner, 2007).

As technology progressed, so did cybersecurity risks and attack avenues. Malicious intent users began employing sophisticated tools and technologies for targeted assaults that may be launched quickly to collect large amounts of data and inflict more significant harm. Security techniques and technology are also growing to counter these assaults (Shahjee, 2022).

According to the definition, “A security operations center (SOC) can be described both as a group, constantly performing in modifications near the timepiece, and a building devoted to and managed to stem, witness, assess & reply to cybersecurity dangers & incidents, and to complete & evaluate regulatory observation.”

Cybersecurity professionals have a growing awareness that machine learning (ML) may facilitate the continuous evaluation of vulnerability data sources and ease some of the cybersecurity mentioned above difficulties arising from the shifting environment of cyber threats. Advances in machine learning-driven security solutions are projected to move the responsibility of handling enormous amounts of vulnerability data away from security specialists and onto some digital alternatives.

Because there are no one-size-fits-all ML-based solutions for automating cybersecurity activities, choosing an appropriate alternative becomes an onerous decision owing to possible cybersecurity hazards from using the wrong ML model. This analysis preaches the difficulty of selecting the best suitable ML model that automatically mitigates more sophisticated attacks. The aim is to improve cybersecurity specialists' situational understanding by providing them with up to date, diversified, & even national security & public.

False positives seem to be the most challenging hurdle for SOC analysts, accounting for more than half of all SOC analyst efforts. Integration of log sources and permitting out-of-the-box use cases, including rules of tracking solutions without sufficient relevance assessment, are important causes of misclassification. The problem for Security Experts was a lack of context to occurrences and threshold-based correlation criteria. Following the incorporation of AI/ML monitoring systems, threshold-based correlation rules are changed into machine learning models. The problem of content lacking was overcome by integrating Threat Data.

Detailing & maintaining security event playbooks/runbooks and maintaining up-to-date expertise are critical for any security solution requiring significant human labour. It is impossible to train resources on multiple technologies in a short period. Many isolated & isolated systems have been integrated into SOC tracking over time, so cybersecurity experts must switch between consoles to respond to issues.

The series of events reported by cybersecurity experts grew due to several point solutions. The leading causes for the security professional leaving the security operations position are repeated work or awake weariness.

Security Operations Centers are an incorporated unit giving checking capacities for the recognition, heightening, and recuperation of safety occurrences on a hierarchical and specialized level. When a security occurrence is recognized, the SOC means to contain the assault at the earliest opportunity, to restrict the possible harm, saving the association cash, information exfiltration, or reputational harm.

There are various sorts of SOCs as talked about in, which could be ordered by the administrations they give, their capacities, or development. Also, they can basically be ordered as

(1) In-house SOC, meaning the association constructs, furthermore, staffs the SOC for its association.

(2) Managed Security Service Provider (MSSP), where an association enlists an outsider, rethinking the danger checking, discovery, and reaction.

A few clients join these two methodologies, building their own SOC yet additionally employ a MSSP to inspire their range of abilities by doing joined observing. There are a few justifications for why an association would utilize one over the other. For instance, their financial plan, their absence of safety skill, or to keep away from the arrangement expenses of a SOC

SOCs comprise of perplexing cycles and innovation and affects different individuals from inside the SOC, association, client, and other outsiders.

## Types of SOC

Framework and Organization Controls (SOC) reports empower organizations to feel certain that specialist co-ops, or potential specialist organizations, are working morally and consistently. Nobody wants to listen to the term examination, yet SOC documents layout validity & dependability for a specialist co-op — an upper hand worth both the time and financial speculation.

SOC reports use free, outsider reviewers to look at different parts of an organization, for example,

* Security
* Accessibility
* Handling Integrity
* Classification
* Security
* Controls connected with monetary announcing
* Cyber security controls

## What are the various kinds of SOC reports?

SOC reports are administered by the American Institute of Certified Public Accountants (AICPA) and centre around offering confirmation that the controls administration associations set up to safeguard their clients' resources (information by and large) are successful. There are four fundamental sorts: SOC 1, SOC 2, SOC 3, and SOC for Cybersecurity, with subsets of each.

## SOC 1

The greatest distinction between a SOC 1 versus SOC 2 report is the focal point of assessment. A SOC 1 report centres around reevaluated administrations performed by administration associations pertinent to an organization's (client element) monetary revealing.

## SOC 2

A SOC 2 report is additionally an authentication report given by a free Certified Public Accounting (CPA) firm. Its centre tends to functional dangers of moving to outsiders outside monetary revealing. These reports depend on the Trust Services Criteria, which incorporate up to five classifications: security, accessibility, handling honesty, secrecy, and additionally protection.

## SOC 3

A SOC 3 report — previously known as a SysTrust or Web Trust — covers revealing comparative regions as the SOC 2 yet isn't as complete. It prohibits specific subtleties of the portrayal and the definite controls in general/aftereffects of testing. While a SOC 2 report confines clients, the advantage of a SOC 3 is that it is a general-use report making it an incredible apparatus to showcase.

## 1.1 Aims

This thesis aims to highlight the problems of thinking like hackers, identifying exploitable flaws in technology, process, and people, or if we are too preoccupied with dealing with billions of logs pushed into SOC. This is the time to begin upgrading our SOC tools with Artificial Intelligence and linear regression-based security technologies to automate and improve our organization's security posture.

## 1.2 Objectives

Below are a few research project objectives that the suggested technological artefact must meet.

* They must examine each record and classify it as hazardous or non-malicious. This task may be simpler by incorporating various tools to speed up the process.
* The goal of this study is to use techniques to identify dangers in SOC.
* Compare static implementation vs ML implementation using the Splunk tool.
* Alert generation and early reporting on possible attacks against the Critical Infrastructure.

## 1.3 Artefact Description

In the design of a Security Operations Center, numerous artefacts are offered (SOC). Regarding technology, it is vital to rely on source code to uncover risks and cut costs. Many devices and technologies must be used to create the SOC from a DiD (Defense in Depth) perspective. Based on industry experience, the technologies indicated below can be used to construct a suitable SOC to analyze risks and detect abnormalities to safeguard the company.

Because most attacks come from outside, appropriate network perimeter protections are crucial. We may reduce the cost of the product by using open-source technologies, and no support is necessary.

## 1.4 Thesis Structure

To meet the previously given research issues, this study is organised as follows.

In Chapter 2, The literature review highlights the research gaps we go on to address in later chapters.

Chapter 3 represents the methodology: We will describe the research Methodology, employed, and define and explain it.

Chapter 4, Design and Implementation: In this chapter we first design to collect logs from the system then we implement that log in our Splunk tool for further analysis.

In Chapter 5, we discuss and conclude the work.

Here is the details list of the Methodology table.

## CHAPTER 2: Literature Review

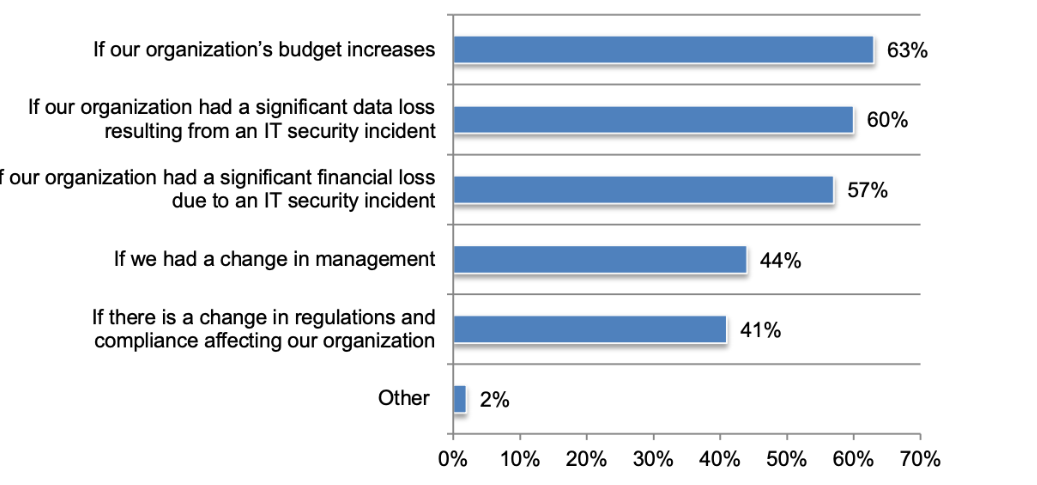
This section identifies and analyzes a few relevant articles on SOC behavior that we discovered throughout our literature study.

In this (Demertzis, 2018), The author's study offers a unique intelligence-driven network forensic for the SOC that uses modest computing wealth and ability and relies only on sophisticated, highly autonomous intelligence methodologies. It is an efficient and dangerous Ensemble Classification analytics technology enabling traffic monitoring, malicious traffic decryption, & encoded traffic authentication.

This (Al-Abassi, 2020), Suggest a novel strategy for building positive sequence approximations of skewed datasets. The improved depictions are input through an ensemble dl attack detection model created expressly for an ICS setting. The proposed attack detection model employs Dnns or DT classifiers to identify information security from the novel representation. The proposed model's performance is assessed using a factor of 10 cross-validations using two real ICS datasets. The findings reveal that the suggested technique performs traditional classifications such as Random Forest, Deep Neural Network, and AdaBoost, along with currently existing systems in the research. The proposed method is a generic strategy that may be easily deployed in current ICS systems.

In this (Jindal, 2020), the Authors describe a novel strategy for SOC incident categorization & prioritizing based on discovering a selection of new ML features using graph visualization or graph monitoring.  They illustrate practically the value of adopting diagram characteristics in terms of better accuracy rate but use a real-world SOC collection using multiple ML classification approaches. They used three statistical algorithms for the investigation LR, XGBoost, and DNN. The investigational analysis showed that now the DNN, the highest performing clustering algorithm, has an area under curve values of 91 percent for the threshold feature set and 99 percent for the magnified feature package that includes diagram functionalities, representing an 8 percent progress in classiﬁer.

In this (Jacobs, 2013) SOCs are almost always a solution for companies that seek to handle compliance with threat monitoring. And some paradigms cover the technological parts of these operations, and there is presently no comprehensive framework that encompasses procedures, staffing, and technology. Furthermore, it would be beneficial for organizations and constituents contemplating developing, purchasing, or selling similar services to assess the efficacy and development of the products supplied.  They suggest a categorization and evaluation methodology for SOC operation in this study, considering both the skills or maturity of the services provided.



**Figure 1 organization to deploy soc**

In this (Agyepong E, 2020), SOCs are almost always a solution for companies that seek to handle regulation with threat monitoring. Although paradigms cover the technological parts of these operations, there is presently no conceptual method that encompasses procedures, staff, & technical. Furthermore, it would be beneficial for organizations & stakeholders contemplating developing, purchasing, or selling similar services to assess the efficacy and development of the products supplied.  They suggest a categorization or evaluation methodology for SOC operations in this study, considering both skills or maturity of the benefits rendered.

In this (Onwubiko, 2015) the author proposed a framework that includes log capture, evaluation, incident response, notification, management, and regular testing. A Cyber Battle Plan is also described, backed by the CSOC architecture. The well-known Protective Management Controls are overlaid on top of the approach (PMCs). Furthermore, the difficulties and advantages of running a Soc are examined.

In this (Agyepong, 2020), they did a detailed review of the tasks of a SOC as defined in many kinds of literature & completed analytical semi-structured discussions with several researchers, including SOC supervisors from numerous organizations. As critical success factors for measuring analysts' productivity, their study results indicate the integrity of. They anticipate that our findings will pique the attention of other defence researchers in analytical assessment approaches.

In this research (Kokulu, 2019), the authors performed 18 moderate discussions among SOC researchers and managers of several industrial sectors to best manage and highlight SOC challenges. Ey detected technical and non-technical difficulties in SOC through our study of the interview data. Furthermore, we discovered inherent conflicts involving SOC directors & their analysts, which, if not resolved, might jeopardize SOC quality and productivity. They compress these challenges into lessons that apply to both the future educational research industry and SOC administration. Authors feel that studies should have been directed on improving the productivity & efficacy of SOCs.

In this (Basyurt), the Authors performed 18 moderate discussions among SOC researchers and managers of several industrial sectors to best manage and highlight SOC challenges. Ey detected technical and non-technical difficulties in SOC through our study of the interview data. Furthermore, we discovered inherent conflicts involving SOC directors & their analysts, which, if not resolved, might jeopardize SOC quality and productivity. They compress these challenges into lessons that apply to both future educational research indusand try SOC administration. Authors feel that studies should have been directed on improving the productivity & efficacy of SOCs.

In this (Alahmadi, 2022), the authors performed a more profound, discovery-oriented participant observation with a security expert on system security limits, including the accuracy & integrity of their alerts. Their findings show that, notwithstanding all FPs, the vast majority are caused by innocuous impulses alarms described by genuine activity in the external structure, which experts may choose to avoid. To virtually assess the adequacy and efficacy of security solutions, suppliers and investigators must be able to distinguish between different sorts of FP. Warning verification is a time-consuming process that might lead to alarm exhaustion &, ultimately, desensitization.

In this (Beulah M, 2022), presents an ensemble learning classifier with an SVM and LR to wireless networking assaults. The pooling algorithm is used to combine models to achieve high efficiency, reduce misclassification, and improve overall recognition accuracy. This should detect the assault immediately through evaluating the data inside the KDD CUP sample, which seems to have 41 characteristics. An examination of the future framework reveals that when utilizing a  Spyder IDE, the ensemble methods EL  obtained a prediction performance of 99.2 percent and a confidence interval of 0.8 percent.

In this (Ding, 2022), the authors present information on secure connection posts for power generation that can monitor and analyze security conditions in electrical networks and issue alerts to companies when the system detects suspicious threats. To be more specific, the suggested framework continually collects public data linked to electromagnetic safety and scans for power control technology available on the Internet to comprehend the global security architecture of power systems. They subsequently install an extendable honeypot technology in the real world to detect the security and intelligence state of powin er real-time data time. Furthermore, they employ a data structure to combine cyber risks in and around the power system to capture the holistic security consciousness of transmission lines. Investigation findings demonstrate demonstrated their framework can identify the efficacy of the control in real-time and successfully protect it from infiltration and incursion.

(Bryant, 2017) made use of the Bryant Kill Chain method for intrusion detection. A virtual Window 7 workstation was used to simulate a corporate environment, and pen testing was done using the Kali Linux default toolset on 26 security test cases picked by the researcher. Default LogRhythm SIEM and the Deconstructed Kill chain method was used for evaluation. Their framework was able to detect 25 out of their 26 test cases (achieving a 96% detection rate), while the default LogRhythm detected 7 out of 26 (achieving g 26.9% detection rate). The test cases were limited to attacks coming from the remote shell (rsh) and malware installations. Also, real-time use behaviour (server-clients) was not included in the virtual environment simulation.

Feng et al. (2017) presented a user-centric machine learning system that leverages big data of various security logs, alert information, and analyst insights to the identification of the risky user. The system provides a complete framework and solution to risky user detection for enterprise security operation centers. Feng and the team also demonstrated that the learning system is able to learn more insights from the data with highly unbalanced and limited labels, even with simple machine learning algorithms. They used Multi-Layer Neural Network (MNN) with two hidden layers, Random Forest (RF) with 100 Gini Split trees, Support Vector Machine (SVM) with radial basis function kernel, and Logistic Regression (LR). Their model though intelligent, achieved a low detection accuracy (Feng et al., 2017).

Demertzis et al. (2018) proposed a novel network forensics framework for security operating centers that rely solely on fully adaptive computational intelligence approaches. The model is a useful and accurate ensemble machine learning forensics tool that uses low utilization of computing power and resources to analyze the network flow instantly to identify encrypted or Malware traffic. The model employs an ensemble architecture that combines Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forest (RF), and k-Nearest Neighbors (k-NN) to investigate malicious activities from data flows in real-time. Their method appears to have the same or a slightly lesser performance across all datasets, and the prediction is less accurate (Demertzis et al. 2018).

(Rahul, Vinayakumar, Soman, & Poornachandran, 2018) the paper utilized Deep Neural Networks (DNNs) to predict the attacks on Network Intrusion Detection System (N-IDS). They applied a DNN with 0.1 rates of learning and is ran for 1000 number of epochs, and KDDCup-'99' dataset has been used for training and benchmarking the network. They regularized the network with a Dropout of (0.01). Their DNN 3-layer network outperformed all the other classical machine learning algorithms on the KDDCup-99 dataset. It is so because of the ability of their DNNs to extract data and features with higher abstraction, and the non-linearity of the networks adds up to the advantage when compared with the other algorithms. The neurons are trained with a bygone benchmarking dataset, as discussed several times in their paper. This comes as a pitfall for their methodology. They propose that additional studies regarding the flexibility of these DNNs in adversarial environments are required.

Rahul et al. (2018) developed a framework of the generative adversarial networks, IDSGAN, to generate the adversarial attacks, which can deceive and evade the intrusion detection system. IDSGAN leverages a generator to transform original malicious traffic into adversarial malicious traffic examples. A discriminator classifies traffic examples and simulates the black-box detection system. More significantly, they only modify part of the attacks' nonfunctional features to guarantee the validity of the intrusion. IDSGAN is a novel framework of generative adversarial networks based on Wasserstein GAN, consisting of the generator, the discriminator, and the black-box IDS. IDSGAN shows its adequate capacity in generating adversarial malicious traffic examples of different attacks, leading the detection rates of various black-box IDS models to decrease to nearly 0. IDSGAN focuses more on evasive and intrusive attacks on the network and not on intrusion prevention. The algorithm was evaluated on an obsolete dataset (KDD).

Azizi et al. (2018) proposed a framework for lateral movement detection based on pattern risk scoring. Their proposed framework can be integrated into the existing network security devices such as next-generation firewall, Advanced Persistent Threat (APT), or Security Information and Event Management (SIEM) to improve the lateral movement detection. Their research developed a rule-set for user scoring to detect lateral movement. Their process flow was not tested on any security data to evaluate its performance.

(Kwak, 2018) proposed a novel visualization system for finding out network-based Underneath attacks (VISNU), which can help security experts of the CSOC to analyze security events more effectively. The VISNU classifies the security events according to each organization and displays them based on both real-time and accumulated information, such as the appearance patterns and their history. Their algorithm developed represented security occurrences in cubes, and the height of accumulated cubes and colors was used as an indicator for a security threat. They developed the formula height of the cubes. Their experimental results demonstrated that it is beneficial for finding out abnormal activities from the security events and provides a better understanding and insights for analyzing them. Their method is limited to some abnormalities and needs improvement.

(Oprea, 2018)They resolved the issue of recognizing pernicious movement in big business organizations and focusing on the identified exercises as per their gamble. They planned a framework called MADE utilizing AI applied to information removed from security logs. MADE uses a broad arrangement of elements for malevolent venture correspondence and utilizes managed learning originally for prioritization, as opposed to recognition, of big business pernicious exercises. The calculation is an irregular backwoods model utilizing the main 40 most elevated positioned highlights and 500 trees. North of one month, MADE effectively focuses on the most hazardous spaces reached by big business, accomplishing an accuracy of 97% in 100 identified areas at a low bogus positive rate. We likewise exhibit MADE's capacity to distinguish new vindictive exercises (18 out of 100) ignored by cutting-edge security advances. For very subtle goes after like Advanced Persistent Threats (APTs), it is plausible for aggressors to create their pernicious associations with stay unnoticed cautiously. MADE has not been tried in distinguishing HTTPS vindictive correspondence with a restricted arrangement of elements. Ill-disposed assaults, because of element significance in their model, could take advantage of the system execution.

Demertzis et al. (2019) Presented an original insight-driven mental registering SOC that depends solely on a moderate, completely programmed methodology. The SOC carries out the Lambda AI design that can dissect a combination of the bunch and streaming information, utilizing an Extreme Learning Machine brain network with Gaussian Radial Basis Function part (ELM/GRBFk) for the group information investigation and a Self-Adjusting Memory k-Nearest Neighbors classifier (SAM/k-NN) to look at designs from continuous streams. The technique performs inadequately with colossal volumes of ongoing information. Likewise, the model presents inactivity (Demertzis et al., 2019).

(Chen, 2019) used three ML approaches (TF-IDF, ET, & Fisher's LDA) are used for automating malware screening & pattern identification. To test their methodologies, seven well-known malware assaults (WannaCry, DBGer, Cerber, Defray, GandCrab, Locky, & nRansom) were employed. Logs for training and pattern extraction was gotten from Cuckoo sandbox. Their investigations revealed that the TF-IDF showcase ranking produces the most identiﬁcation of malware detection characteristics, however the ET randomized trees) the technique is the most day when & resistant to input volatility. The sole virtual environment used (Cuckoo with offline windows environment) represents a real-world enterprise environment. Enterprise environment variables such as data volume, and online server-client interaction, were not considered. Also, the research was mainly focused on ransomware, a variation of Malware, which is just one of the many security challenges tackled by an Intelligent SIEM.

(Zhang, 2019) used Gaussian mixture models (GMMs) to detect XSS through web requests and web responses representing the integration of the dual-stage model. Word2Vec model was used for feature engineering on normal and XSS payloads on web traffic datasets (response and requests). Then trained using GMMs, and their model was evaluated using ROC (Receiver operating characteristic). Their model successfully used both response and requests for analytics and prediction of Cross-Site Scripting (XSS) events instead of single-stage most current research use. The evaluation metrics showed that their dual-stage model was more effective than a single model. The model was not evaluated on real-time traffic data and real-time XSS payloads. 2. Their model was not integrated into a SIEM for real-time alarm generation and prevention mechanism.

(Majeed, 2019) introduces a revolutionary graphical system that enables security analysts to grasp a comprehensive picture of SIEM rule executing & alert scenarios that might occur in preparation depending on the closest condition graphically and in real-time. Aside from actual rule monitoring, it also allows security analysts to investigate the reasons behind alarms in an orderly and effective manner using online accounts. The created rule-set was connected in order to create a graphic representation of the reasons behind protection warnings. Their suggested solution allows security analysts to examine the status of several rules in real-time, analyze susceptible rules, and undertake discovery using login details. In other phrases, a security researcher may profit from concentrating just on the constraints that may cause an alarm. Their system is currently in the early stages of development.

(Andrade, 2018) The research investigated the use of Big Data as a supplemental method for investigating security incidents in a real-world cyber incident response centre (CSIRT) context.

Security Operation Centers are entrusted with gathering and examining digital danger information from various sources to impart advance notice messages and arrangements. These errands are broad and asset consuming, which makes supporting methodologies important to specialists. Be that as it may, to execute such methodologies, data about the difficulties these specialists face while it is important to play out these undertakings. Author (McLeod, 2022) consequently led semi-organized master meetings to distinguish these difficulties. Thusly, important experiences into these difficulties in light of master information are procured, which consequently could be utilized to foster mechanized ways to deal with help specialists and address these difficulties.

The overlay of specialized arrangements onto digital territory is affected by conventional security models (e.g., layered protection). Notwithstanding, basic layering of symmetrical protection innovations doesn't guarantee an effective digital guard. For instance, zero-day exploits and private key trade-off are instances of weaknesses that rout solid specialized arrangements. Digital investigation is hence key to furnishing network protectors with a reasonable organizing of how a framework is safeguarded at each period of an assault cycle. What's more, an organized execution of guarded strategies (e.g., misdirection), addressed by secure cycles and innovation arrangements, gives a comprehensive way to deal with architecting a Security Operations Center (SOC) that is intended to safeguard current endeavour processing systems (Couretas, 2022).

## Predictive model with standard SIEM Applications.

The IT division gets many cautions produced by the interruption identification and avoidance frameworks (IDPs). SIEM was at first evolved to manage countless alarms.

It initially developed from the log of the executive's discipline and has extended its degree to incorporate a mix of safety occasion the board (SEM) and security data the board (SIM). With this mix, SEIM gives an ongoing investigation of safety cautions created by the application and organization equipment (Khaleghi, 2022).

By and large, SEIM is utilized to distinguish security issues, log security information, and for consistency announcing purposes. Payment Card Industry Data Security Standard (PCI DSS) drove SIEM reception in enormous ventures, and gradually more modest associations took their steps.

Today, SIEM arrangements have developed extensively to incorporate high-level strategies like User Behavior Analytics (UBA), Deep Packet Inspection and Security Orchestration, Automation, and Response (SOAR). UBA utilises calculations in light of AI and chips away at a prescient model. It has supplanted the standard based calculations to generally increment productivity and help the association recognise possible danger.

Overseen SIEM arrangements have arisen for private companies, where SMBs with restricted assets can likewise profit from high-level AI abilities, social examination, and different administrations (Eskelinen, 2022).

Amid cyberattacks, huge organizations have assets to relieve the harm, through SMBs, by and large, don't have something very similar available to them. Overseen SIEM for independent companies assists them with safeguarding their resources during such circumstances.

## Gaps in literature

There is a lack of literature on security operations centers. Only a few studies on security operations centers have been published recently. Most security operations center material is based on best practices and security vendor blogs and presentations. There are few research publications on security operations centers using machine learning.

.

## CHAPTER 3: Research Methodology

## 3.1 Research Methodology:

In this, we discuss the quantitative research strategy thesis adopted to provide organizations with a secure environment by analyzing logs in the security operation center using the Splunk tool.

## 3.2 How Log data flow in SOC

Logs to be considered for mining, analytics, and generating a machine learning model for an Intelligent Security Operations Center. Explanation

1. **Logs from Security Controls**

* IDS
* Endpoint Security (Antivirus, antimalware)
* Data Loss Prevention
* VPN Concentrators
* Web filters
* Honeypots
* Firewalls

1. **Logs from Network Infrastructure**

* Routers
* Switches
* Domain Controllers
* Wireless Access Points
* Application Servers
* Databases
* Intranet Applications

1. **Non-log Infrastructure Information**

* Configuration
* Locations
* Owners
* Network Maps
* Vulnerability Reports
* Software Inventory

1. **Non-log Business Information**

* Business Process Mappings
* Points of Contact
* Partner Information

1. **Logs from the Internet Infrastructure Format**

The TCP is a communication system that sits among an application software or the IP at a moderate level. TCP has become one of SIEM's primary data transmission modes. (Thompson, 2015; 2020; SIEM Architecture: Technology, Process, and Data | Exabeam)

A transmission Control Protocol is a system that allows clients and servers to connect transmit streams of data. Transmission Control Protocol ensures that data is delivered and that bits are sent in the same sequence as they were sent.

TCP connection logs may be broken down into three categories:

1. Essential aspects of individual TCP connections
2. Content features inside a reference given by domain knowledge
3. Traffic features computed over a two-second time window

The following tables 1–3 detail the characteristics of these classifications:

Table 1: Basic features of individual TCP connections.

|  |  |  |
| --- | --- | --- |
| *feature name* | *Description* | *type* |
| duration | length (number of seconds) of the connection | continuous |
| protocol\_type | type of the protocol, e.g., TCP, UDP. | discrete |
| service | network service on the destination, e.g., HTTP, telnet | discrete |
| src\_bytes | number of data bytes from source to destination | continuous |
| dst\_bytes | number of data bytes from destination to source | continuous |
| flag | normal or error status of the connection | discrete |
| land | 1 if connection is from/to the same host/port; 0 otherwise | discrete |
| wrong\_fragment | number of "wrong'" fragments | continuous |
| urgent | number of urgent packets | continuous |

Table 2: Content features within a connection suggested by domain knowledge.

|  |  |  |
| --- | --- | --- |
| *feature name* | *description* | *type* |
| hot | number of "hot'" indicators | continuous |
| num\_failed\_logins | number of failed login attempts | continuous |
| logged\_in | 1 if successfully logged in; 0 otherwise | discrete |
| num\_compromised | number of "compromised" conditions | continuous |
| root\_shell | 1 if root shell is obtained; 0 otherwise | discrete |
| su\_attempted | 1 if “su root” command attempted; 0 otherwise | discrete |
| num\_root | number of "root" accesses | continuous |
| num\_file\_creations | number of file creation operations | continuous |
| num\_shells | number of shell prompts | continuous |
| num\_access\_files | number of operations on access control files | continuous |
| num\_outbound\_cmds | number of outbound commands in an FTP session | continuous |
| is\_hot\_login | 1 if the login belongs to the "hot" list; 0 otherwise | discrete |
| is\_guest\_login | 1 if the login is a “guest” login; 0 otherwise | discrete |

Table 3: Traffic features computed using a two-second time window.

|  |  |  |
| --- | --- | --- |
| *feature name* | *Description* | *type* |
| Count | number of connections to the same host as the current connection in the past two seconds | continuous |
|  | *Note: The following features refer to these same-host connections.* |  |
| serror\_rate | % of connections that have "SYN" errors | continuous |
| rerror\_rate | % of connections that have "REJ" errors | continuous |
| same\_srv\_rate | % of connections to the same service | continuous |
| diff\_srv\_rate | % of connections to different services | Continuous |
| srv\_count | number of connections to the same service as the current connection in the past two seconds | Continuous |
|  | *Note: The following features refer to these same-service connections.* |  |
| srv\_serror\_rate | % of connections that have “SYN” errors | Continuous |
| srv\_rerror\_rate | % of connections that have “REJ” errors | Continuous |
| srv\_diff\_host\_rate | % of connections to different hosts | continuous |

Also, below log format from PowerShell in windows server accessing contents of a CSV file in a database server:

*AccessLogParser.ps1 - HostName jboss\_host123 -source ".\access\_log.2016-05-08.log" -target "./access\_log\_analysis.csv" -fields"3:Timestamp;7:Url;10:HttpCode;11:PageSize;12:TimeTaken" -group "Url;HttpCode" -aggregate "HttpCode:Count;PageSize:Sum;TimeTaken:Percentile\_90;TimeTaken:Average" -calculate "Concurrency=(([HttpCode:Count]/300)\*([TimeTaken:Average]/1000))" -time "[dd/MMM/yyyy:HH:mm:ss" -Interval 5 -split "*

## 3.4 ML used in Soc

Machine learning can assist examiners with tracking down basic security occasions, expanding their efficiency and results. ML can distinguish designs, construes ends, ties information from different sources, and can diminish the number of cautions without relying upon the expert's insight to recognize the occasion. ML-driven SOCs can forestall dangers through endpoint security, identify and examine dangers through broadened identification and reaction (XDR), and influence security coordination capacities to carry out mechanized playbooks for quicker occurrence reactions. For identifying clever assaults, the ML-driven SOC utilizes XDR elements to gain from the information in the information lake (a brought together wellspring of log information) to get experiences into the identified action. This assists the SOC with laying out an acknowledged standard way of behaving and separating it from pernicious ones. The ML models should be prepared to order information to accomplish the ideal results. The information expected to prepare the ML model is of two sorts. Marked information and Unlabelled information. Marked information implies labelled information, generally connected with an earlier comprehension. It could have more than one characteristic worth, helping the ML calculation sort and order them. This information is utilized for directed learning models to determine numerical models and empower the framework to order information. Be that as it may, this learning model can function admirably just with limited information classes. Unlabeled information has no names labelled on them. This is the kind of information that ML models need to characterize.

Regulated ML utilizes marked informational collections to go with informed choices. The information can be dubious documents or malware contaminated records, which can assist the framework with realizing the distinction between tainted records and ones that are not. This type of learning requires a consistent feed of information to handle developing dangers and functions admirably when the quantity of informational collections is limited. Unaided AI is utilized when the quantity of informational collections isn't characterized. This model develops math models from arranged informational indexes to track down designs in input information to infer yield and perceive any oddities in information.

ML upgrades the framework's capacity to comprehend and foresee yield given various calculations and information with practically no programming. This assists a SOC with identifying any irregularities and signatureless dangers consequently. It assists them with being proactive via mechanizing the recognition and reaction of episodes that the investigator might have missed in a load of alarms. Nonetheless, ML can't supplant a security investigator.

ML can augment SOC results through compelling danger location, weakness in the board, and empowering quality security proposals. An ML-driven SOC, alongside the experience of experts, can be an intense power in distinguishing zero-day takes advantage of malware and ransomware with high accuracy to guarantee better degrees of security. ML's capacities to persistently and powerfully gain from ordinary gauge ways of behaving, network traffic designs, and framework utilization across the hierarchical conditions make it significantly more dependable to relieve assault dangers. ML can consequently investigate factors and informational indexes to rapidly anticipate and distinguish dangers, weaknesses, and shortcomings before even an assailant can take advantage of them. By successful ingestion of danger knowledge, ML can improvize danger reaction by executing the right playbook, decreasing the examination season of investigators by giving them the right perspective to act, and limiting the interim to distinguish and answer. It can give ideas on distinguishing dangers and how to relieve them later on. ML allows us to adopt a more smart security strategy by consolidating various information types and cognitive thinking to draw deductions and empower better security forecasts and choices.

As the digital transformation of society continues, the importance and need for a focus on security in these systems become more and more poignant. As incorporating security in a security operation center is a difficult problem, we aim to address the detection of threats in SoC. By leveraging the ability for ensemble learning to respond to novel scenarios, a more adaptable and powerful heuristic detection engine for the use in detecting the threat in SoC could be possible, compared to a deterministic or another Machine learning (for example, using Bayesian networks) detection engine.

## 3.3 Project Management

In this research agile project management technique is chosen to handle the complete research project so that it may be adequately managed and within the stated deadline. It aided in the management of the entire project following the specifications. The entire project is divided into several stages and sub-phases using this technique, and an appropriate timetable is created for assigning proper time and resources to each work. The agile strategy will assist us in conducting the best possible research with the funds available to achieve each project's success.

## CHAPTER 4: Design / Development / Implementation

In this chapter we first design to collect logs from the system then we implement that log in our Splunk tool for further static and ML analysis.

## 4.1 Designing

An experiment was designed to confirm thartefactct. The experiment was carried out in a service control and system architecture setting. The effective demonstration of comprehensive cyber powers in that environment is defined as executing the task and validating our artefact.

## 1. Splunk

Splunk Cloud Platform provides you with the usefulness of Splunk Enterprise for gathering, looking, observing, revealing, and examining all of your ongoing and verifiable machine information utilizing a cloud administration that Splunk conveys halfway and consistently to its huge assortment of cloud-based clients, going from Fortune 100 organizations to little and medium-sized organizations. Splunk Cloud Platform offers the honour-winning highlights of Splunk Corporate as a cloud-based help. Splunk controls and updates the Splunk Cloud Platform administration indistinguishably, guaranteeing that all Splunk Cloud Platform clients approach the latest highlights and abilities.

Splunk programming might supplant or enhance a SIEM in a SOC. Splunk programming might be used \ completely independent from a current SIEM or quickly move a portion of its information into a SIEM. The current SIEM is used for two situations while connections, alarming, and episode work process Splunk programming is utilized fundamentally for serious occurrence examinations/criminology and progressed investigation.

The current SIEM takes care of information into Splunk programming, and the Splunk stage plays out all utilization cases. This situation typically exists when the SIEM authorities are on hundreds or thousands of hosts and eliminating them, or it is hard to supplant them.

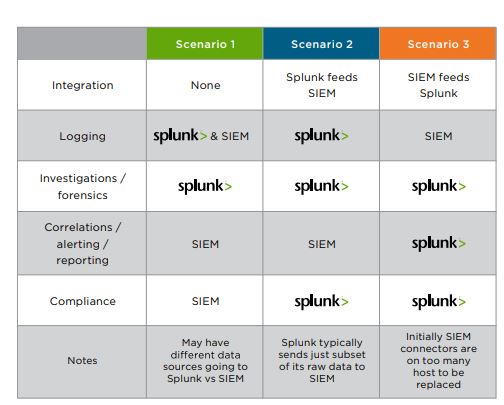
****

Figure 2 Splunk

## 2. SIEM develop tools on machine learning techniques

Many experts still associate SOC with SIEM tools & security management. To some extent, this is valid since the SIEM tool serves as the SOC's main engine, collecting logs from linked log types, processing them according to established rules, and issuing warnings.

Since the SIEM tool's debut, security engineers/analysts have implemented various rules to monitor and correlate the log data pushed into the SIEM tool. These approaches are introduced one by one by the needs of the environment. The volume of log data that the SIEM system must analyze makes it hard to construct a rule that can spot abnormalities.

The next generation of SIEM solutions will include ML techniques, merging rules or algorithms with data to create experience and understanding, and thoughtful analysis to deliver predicted reactive outcomes.

## 3. How ML Affects Security Operations Centers:

The SOC monitors the natural cyber security threat environment 24 hours a day, 365 days a year, and where any attack vectors may be prevented and neutralized.

As bright as they seem, SOCs have one major flaw: the IT security personnel who maintain this fortress are sometimes overworked and exhausted from keeping track of everything, especially regular daily alerts.

This problem is exacerbated by the current shortage of cybersecurity workers in the employment market. This is underscored in a recent survey, which found that the average cost of security breaches increased by 6.4 percent, from $3.62 million to $3.86 million.

These data portray a concerning picture for the SOC and highlight a tendency known as "Alert Fatigue." according to the survey, 30% of IT security teams ignored most of their alerts, while 4% turned off their notifications entirely. Similarly, 56 percent of IT security personnel dismissed any form of attention based on their previous experience with false positives.

## 2. Implementation

### Static implementation

In this, first, we create a dashboard

Create a dashboard

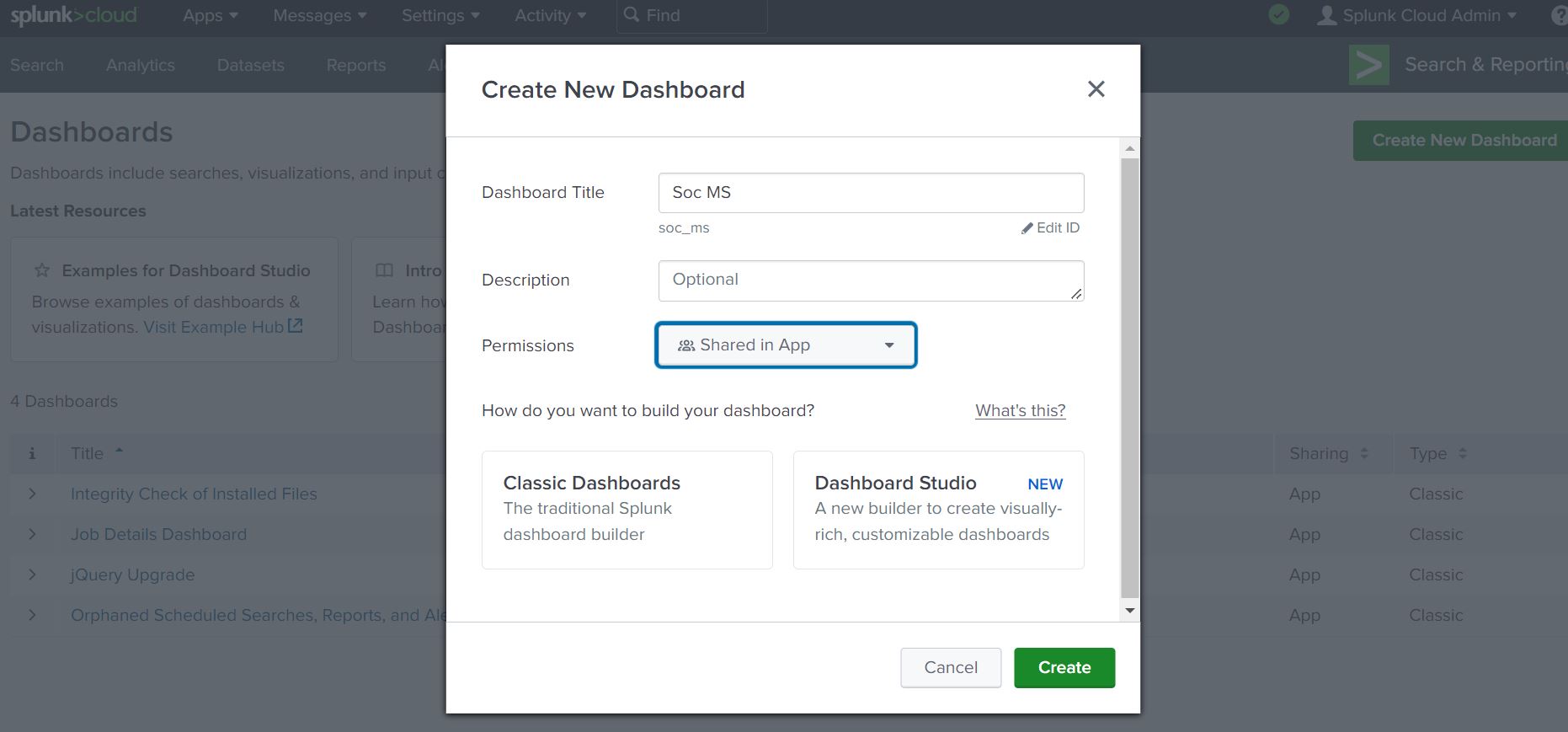


Figure 3 dashboard creation

Now, we must select a dashboard-type.

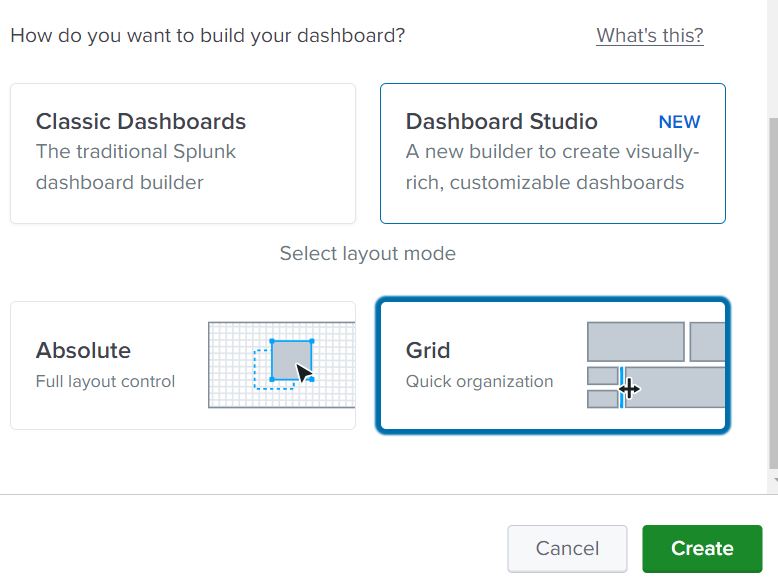


Figure 4 dashboard type

Now press the button to create.

Our dashboard is created

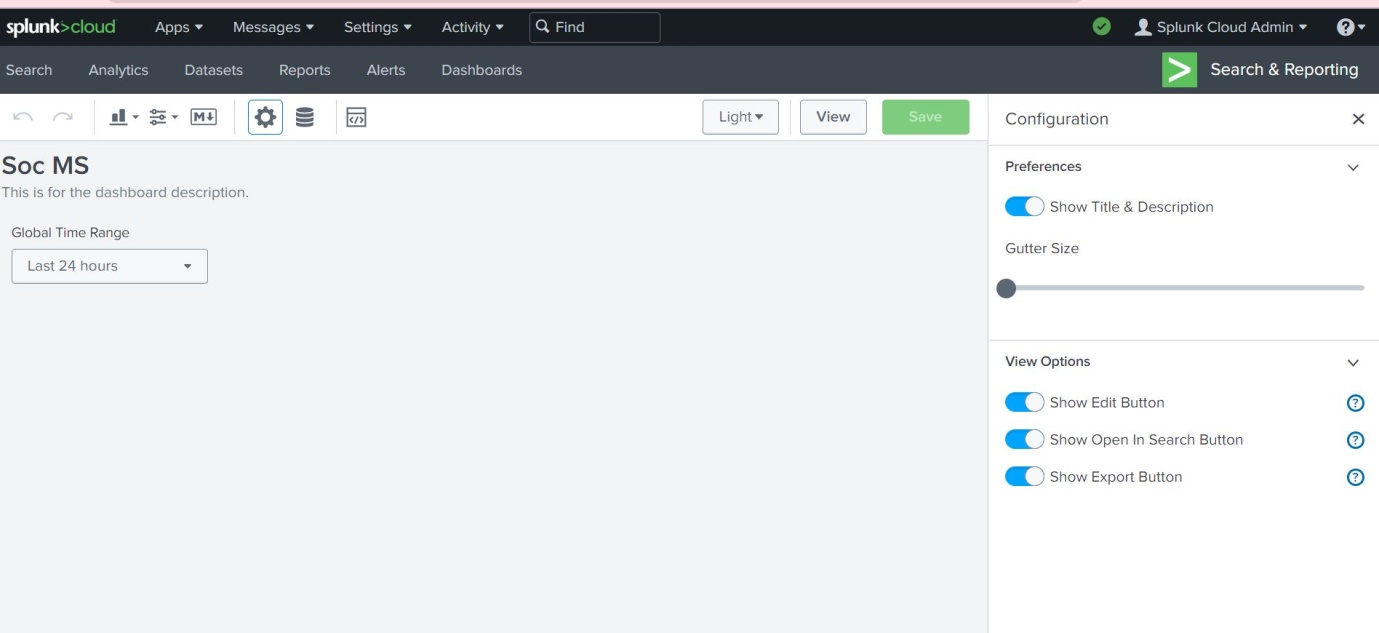


Figure 5 custom dashboard look

Let’s begin by creating the dashboard itself and then generate the panels:

1. Go to the search bar in the Destinations app
2. Run this search command:

sourcetype\*"auth"failed

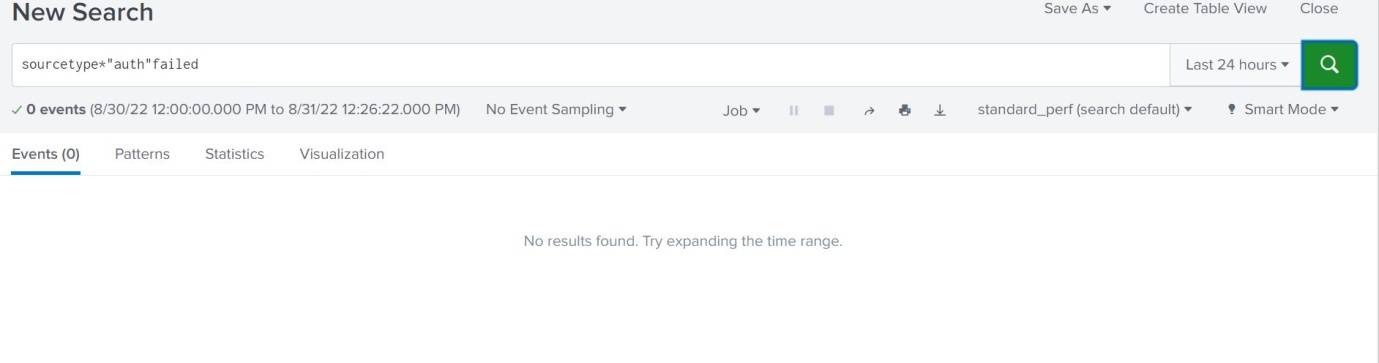


Figure 6 search auth failed

| from datamodel:"internal\_audit\_logs.Audit"

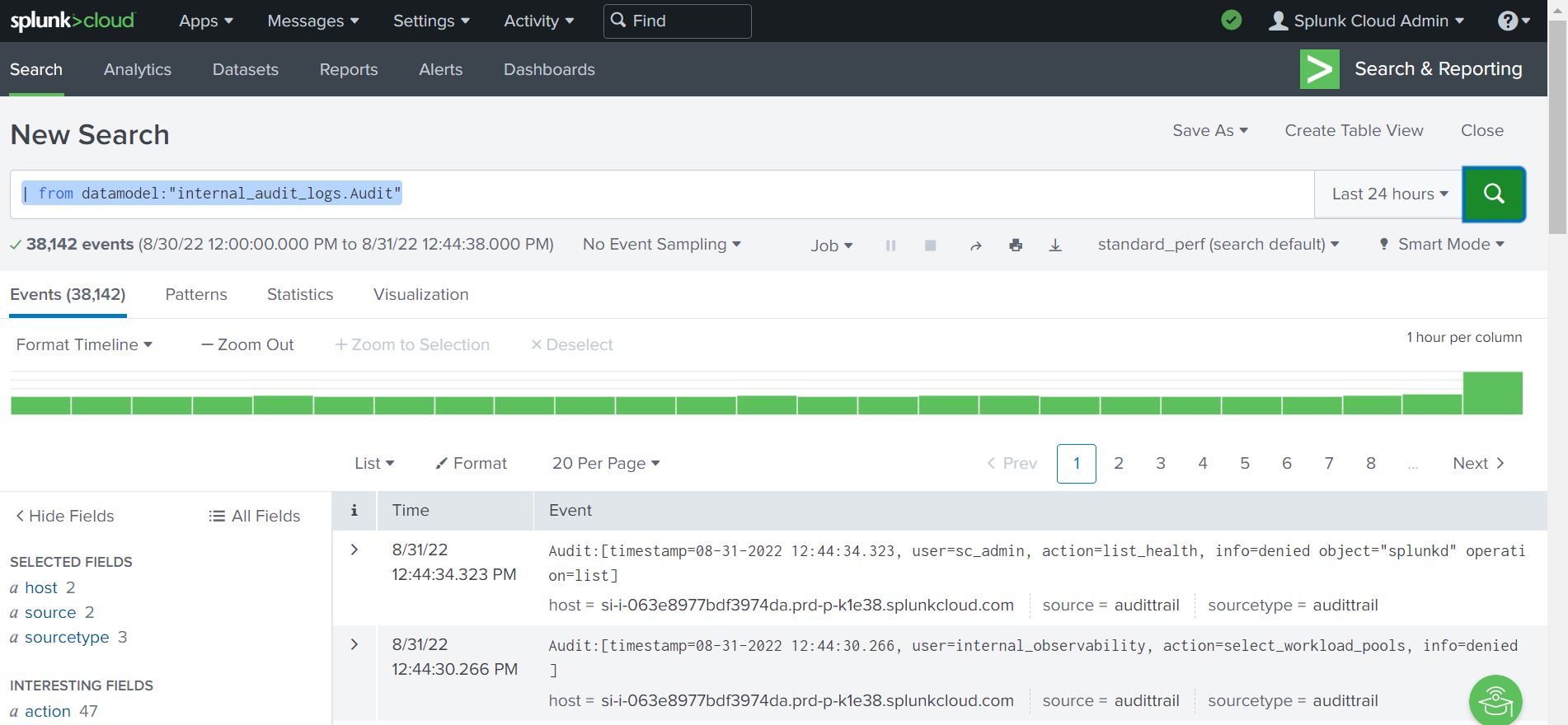


Figure 7 search internal audit log of the system

By simply clicking on the blue marked hyperlinks, you can see the SPL operating behind the scenes. Please view the screenshot below.

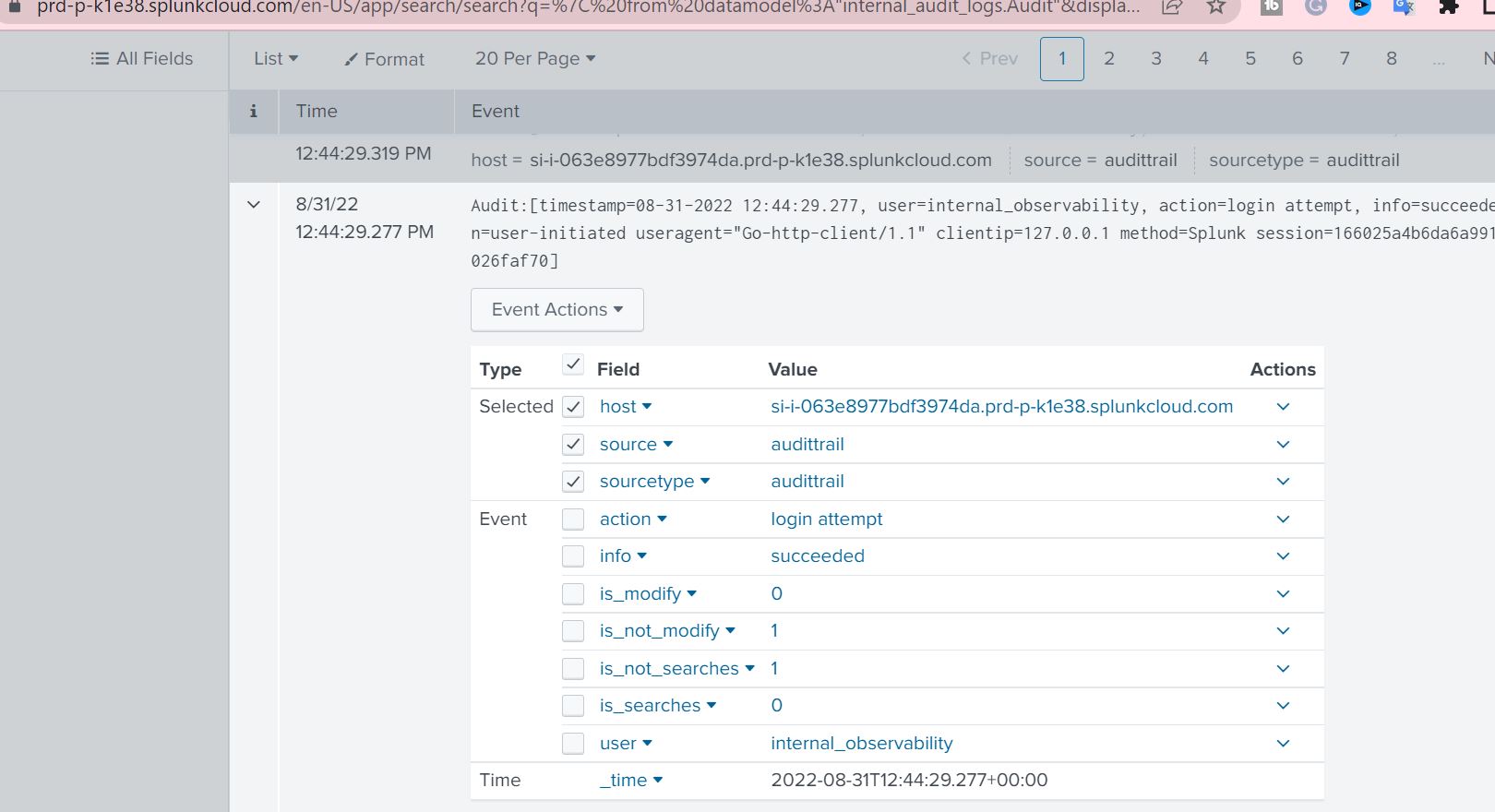


Figure 8 extend the features

(index=\* OR index=\_\*) (index=\_audit) host="si-i-063e8977bdf3974da.prd-p-k1e38.splunkcloud.com" | eval is\_searches=if(searchmatch("(action=search NOT dmauditsearch)"),1,0), is\_not\_searches=1-is\_searches, is\_modify=if(searchmatch("(action=edit\_user OR action=edit\_roles OR action=update)"),1,0), is\_not\_modify=1-is\_modify | fields "\_time" "host" "source" "sourcetype" "action" "info" "user" "is\_searches" "is\_not\_searches" "is\_modify" "is\_not\_modify"

We save the model by tapping the "Save" button to apply the model to new lines.

Button at the upper right-hand corner. When saved, the model shows in the rundown of saved models.

Tests like those portrayed here.

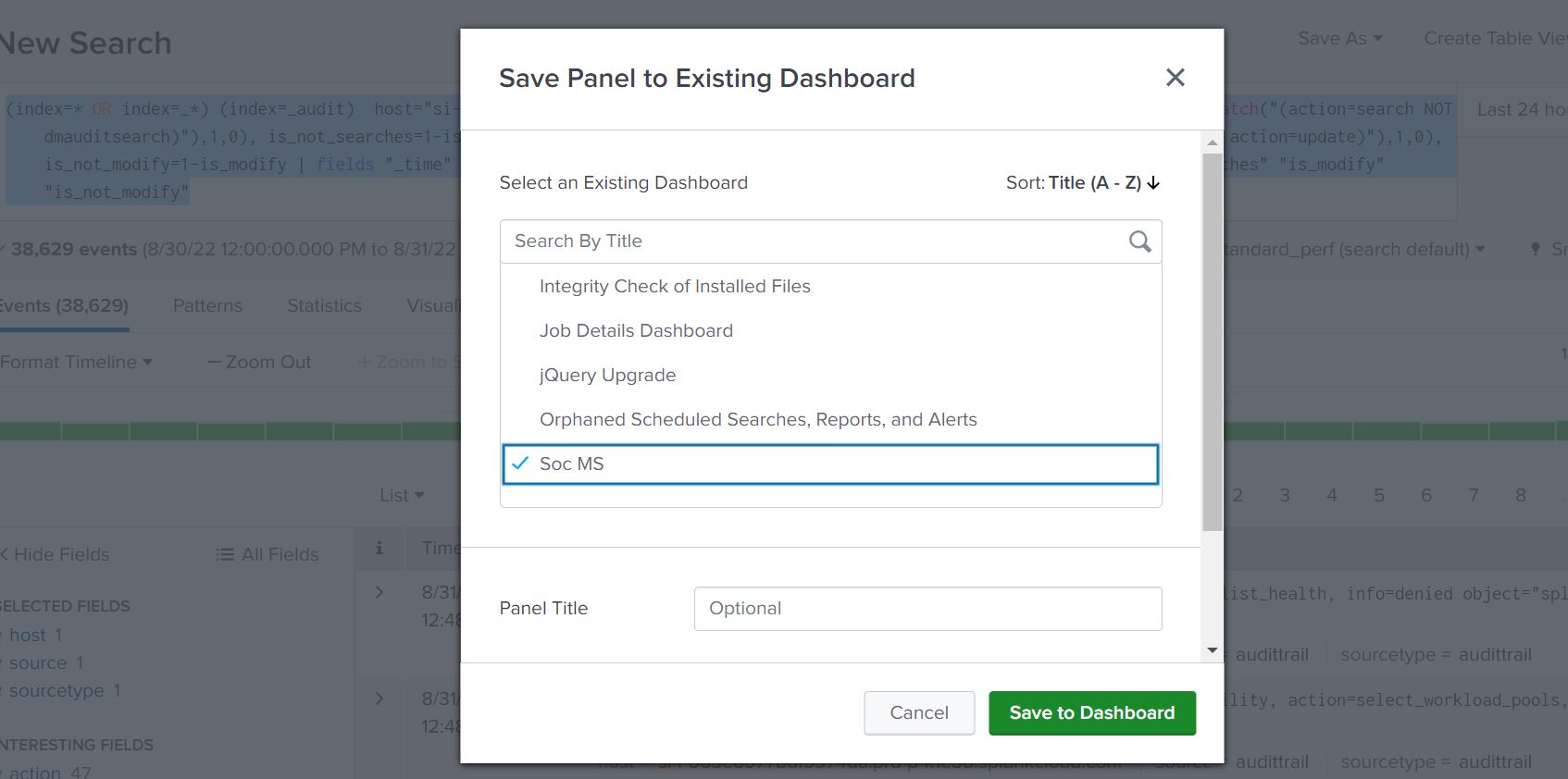


Figure 9 save the model in the Soc dashboard

We utilize this saved experiment model to give cautions on the approaching information. By tapping on the regulate button or choosing the "create alert" choice, we empower ourselves to send a caution in case of any suspicious logins on the server. Notwithstanding, in this task, we wanted an illustration of Phantom being utilized as a framework.

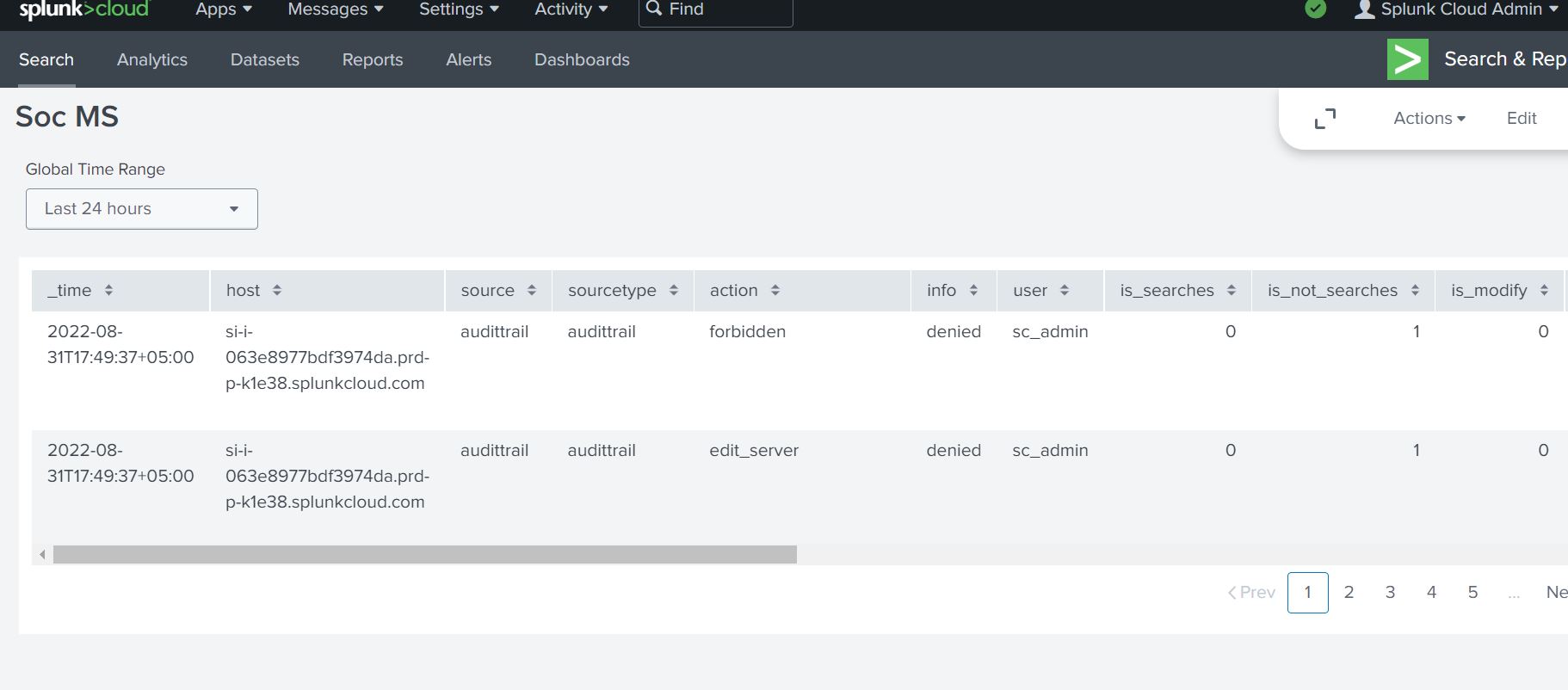


Figure 10 dashboard of internal logs

Now, we used graph to check the logs of the system in the dashboard.

Which easy to analyze for anyone.

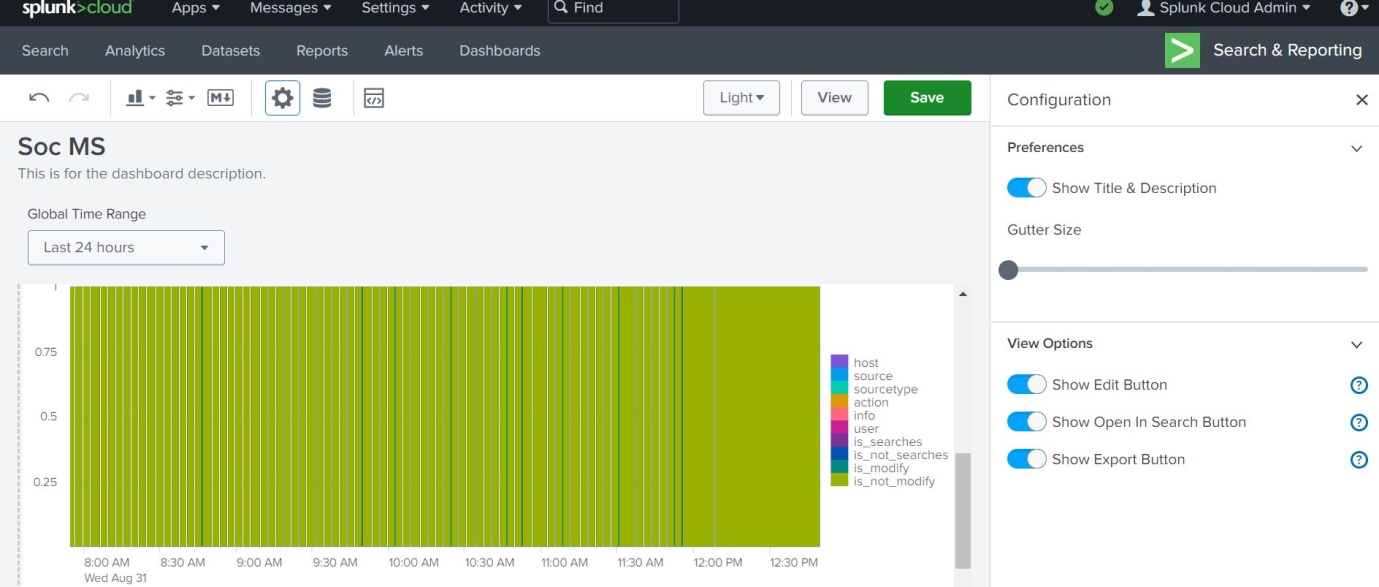


Figure 11 graph of internal logs

## ML model Implementation using Splunk

We transferred many years of logs from our system into Splunk cloud-based and established a basic use case named "Suspicious logins" on the server. To select and prepare given data, Splunk Processing Language was utilised. The computer's login information is shown below.

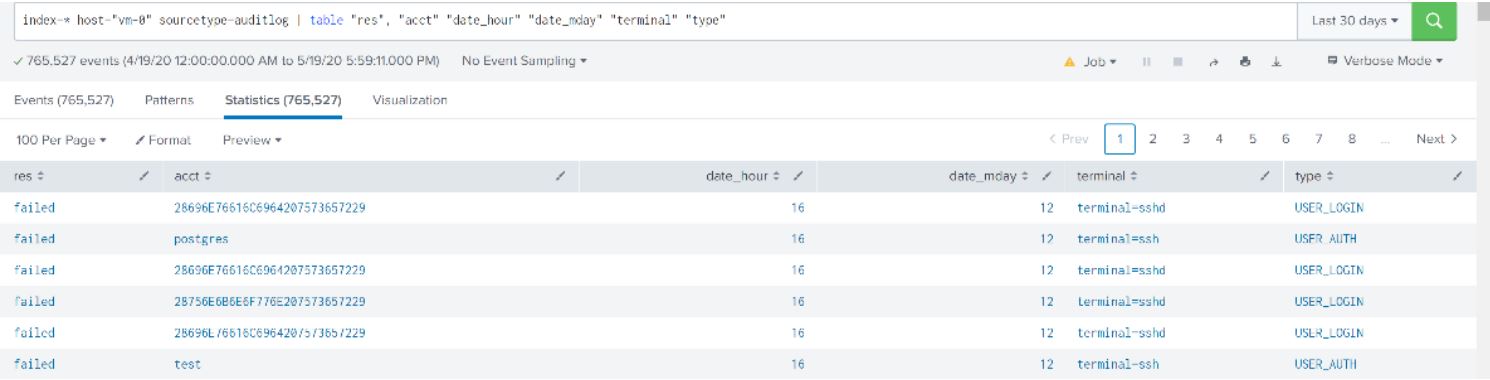


Figure 12 Splunk Processing Language

Splunk includes several features that may be utilized for a variety of applications. On the Splunk cloud, we installed the "Splunk ML Toolkit.

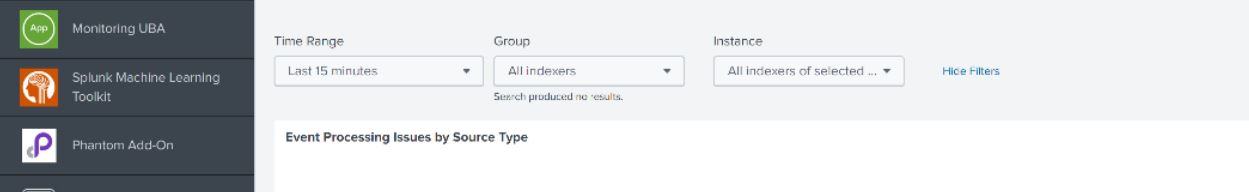


Figure 13 Splunk ML toolkit

"Splunk ML" was utilized to develop a linear regression after loading the programme.

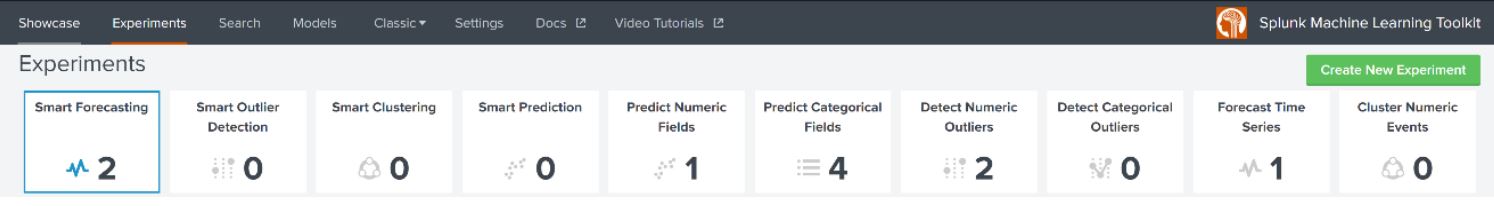


Figure 14 New experiment

Now we select new experiment from the experiment

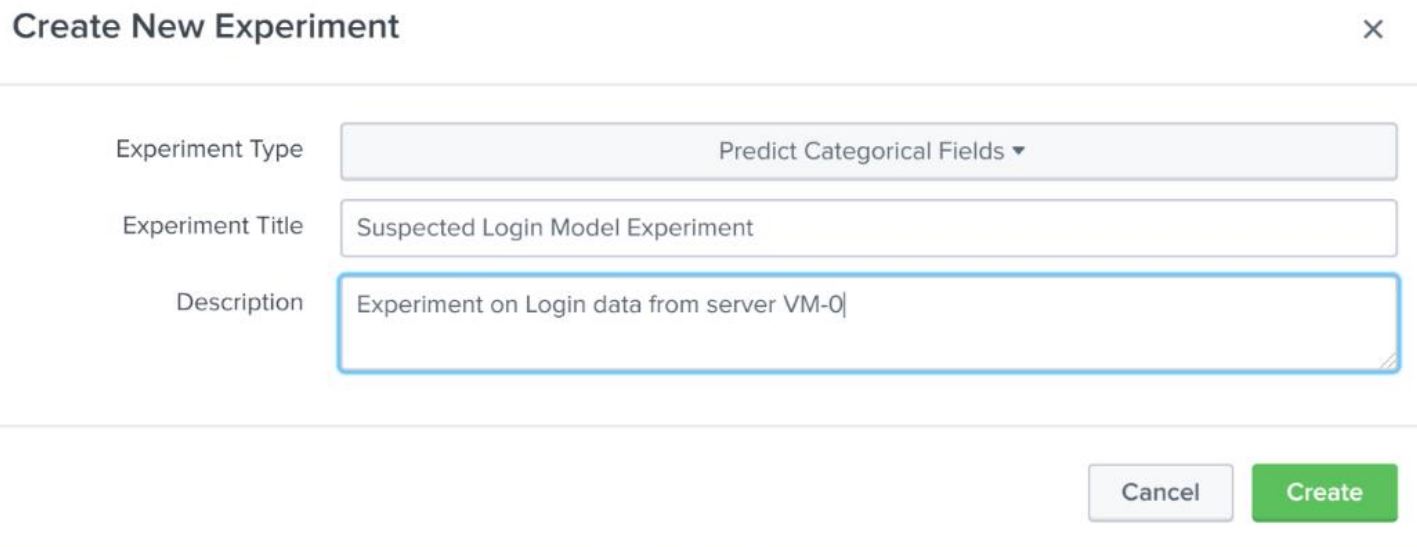


Figure 15 given title

After giving the title click create.

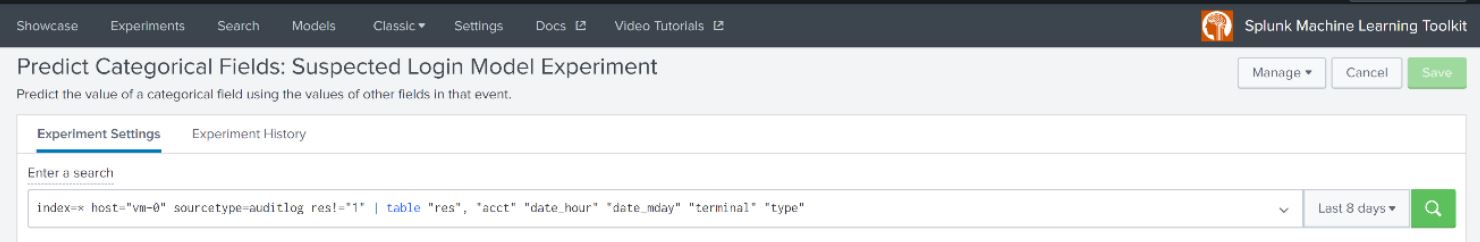


Figure 16 prediction search

"Splunk ML" was used to generate a regression model when the tool was loaded.

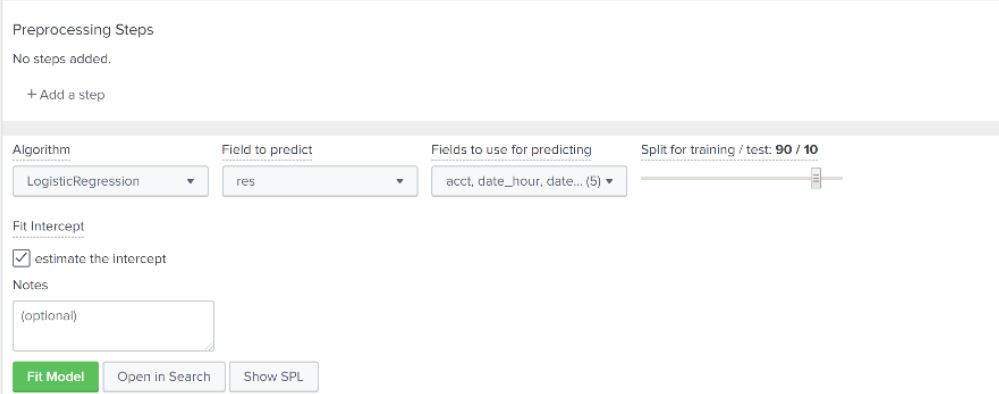


Figure 17 logistic regression model

Here we select a logistic regression model from multiple models, whereas ‘res’ is our final resultant output used in prediction.

Now we select the feature engineering for the model under fields used for predicting. Finally, we choose the training/testing field.

When all fields are completed, we select "Fit Model," & our predictions are displayed.

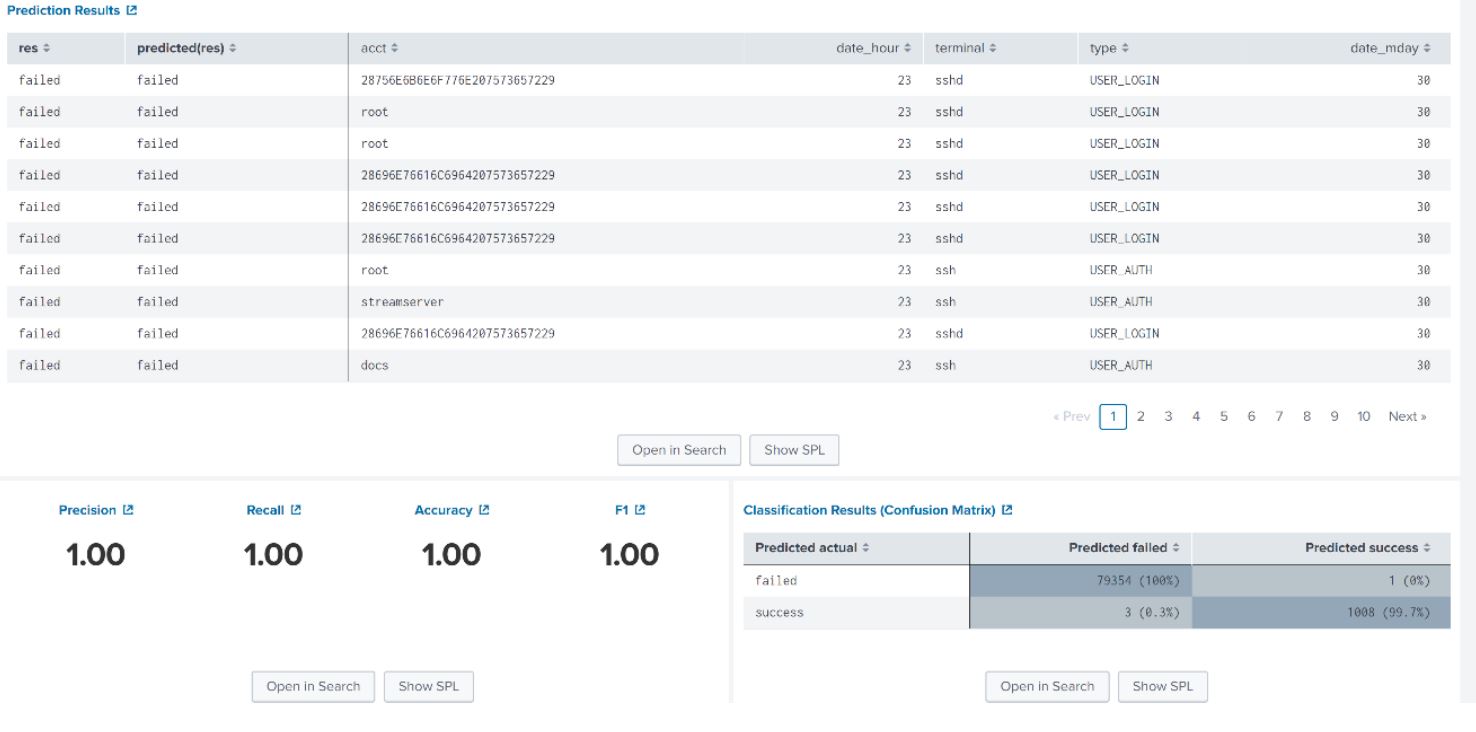


Figure 18 ML result

we can check SPL operating behind the settings by right-clicking on the blue underlined hyperlinks.

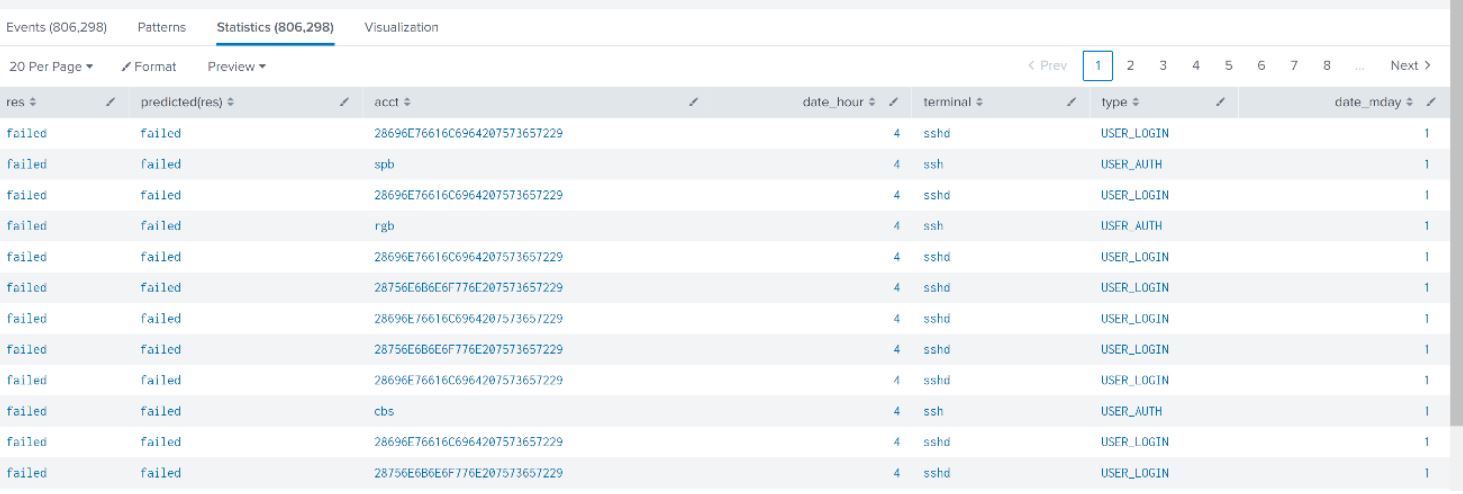


Figure 19 ML SPL

As we can see, we get a suspected login results.

We use SPL and use this query.



Figure 20 spl suspected login result.

We save the model so that it may be used in future rows by pressing the "Save" icon in the top right corner. The model appears in the list of completed studies once it has been saved.



Figure 21 ML page

Based on the incoming data, this saved experiment model is utilised to create alerts. We may send an alert in the event of suspicious logins on the computer by clicking the management button and selecting the "create alert" option. However, in order to demonstrate system usage, we needed an automated action to be performed in this experiment. We have the freedom to set the training timetable by selecting "Edit Training Timetable.

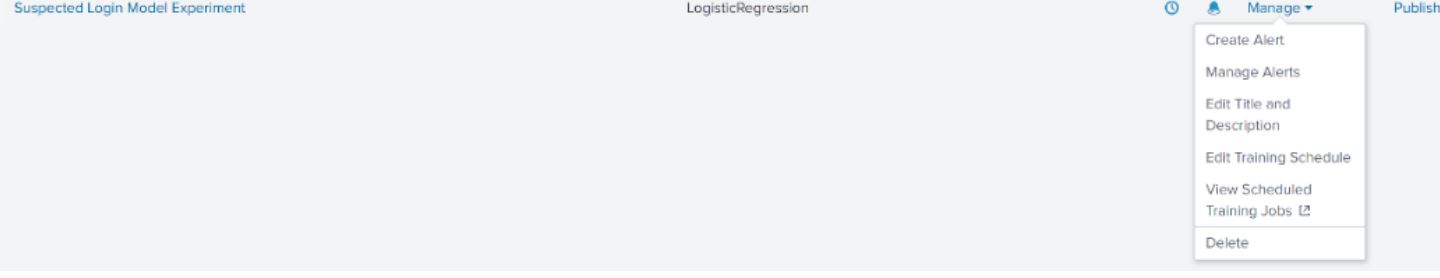


Figure 22 Alert setting

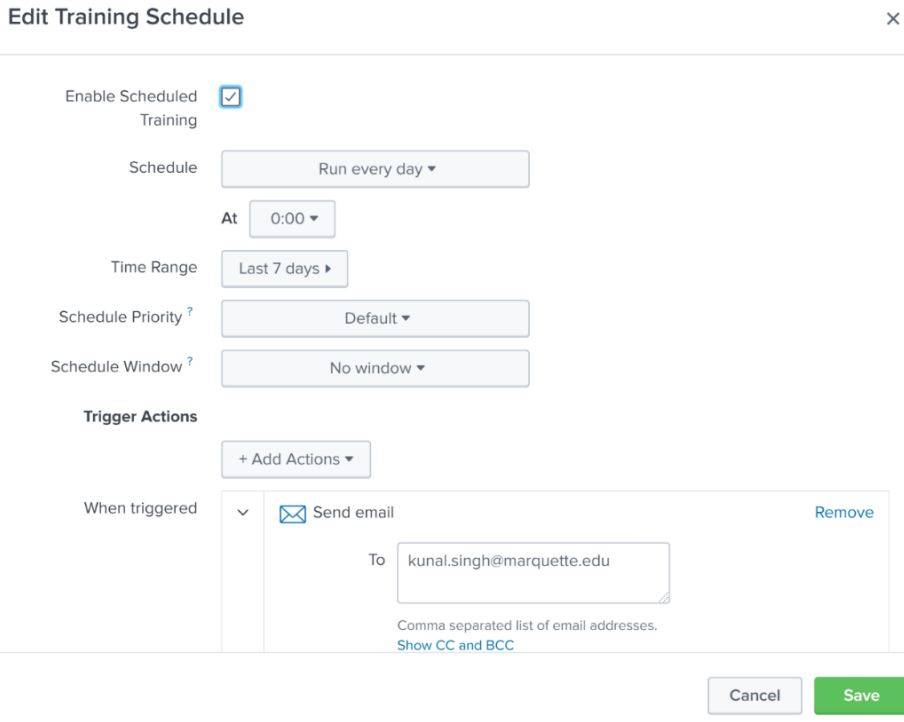


Figure 23 edit time scheduling for ML training

We had the option to construct the model and ensure the forecasts were right; we were ready to produce warnings and train the model on a set schedule.

## CHAPTER 5: Results

In this chapter, I will test the results performed and check the graph and logs of internal system from static analysis and ML, which I implemented in the previous chapter.

Company firewalls secured Splunk, and we did not have access to the firewall or network logs at the time of this experiment. We did our experiment using the system logs. We set up Splunk forwarder on each auth log, system log, etc.

Our experiment focused on the framework's External Remote Services' HTTP. We transmitted login audit data from the machines to Splunk cloud to identify incoming threats via an external remote service. We used Splunk to provide notifications when an unsuccessful login attempt was made on our servers. We found a brute force assault on our servers while examining audit logs with SPL (Splunk processing language).

We established a scheduled report every 24 hours for the unsuccessful logins and forwarded it to our \semail addresses.

Figures show screenshots of emails we got due to automatic notifications.

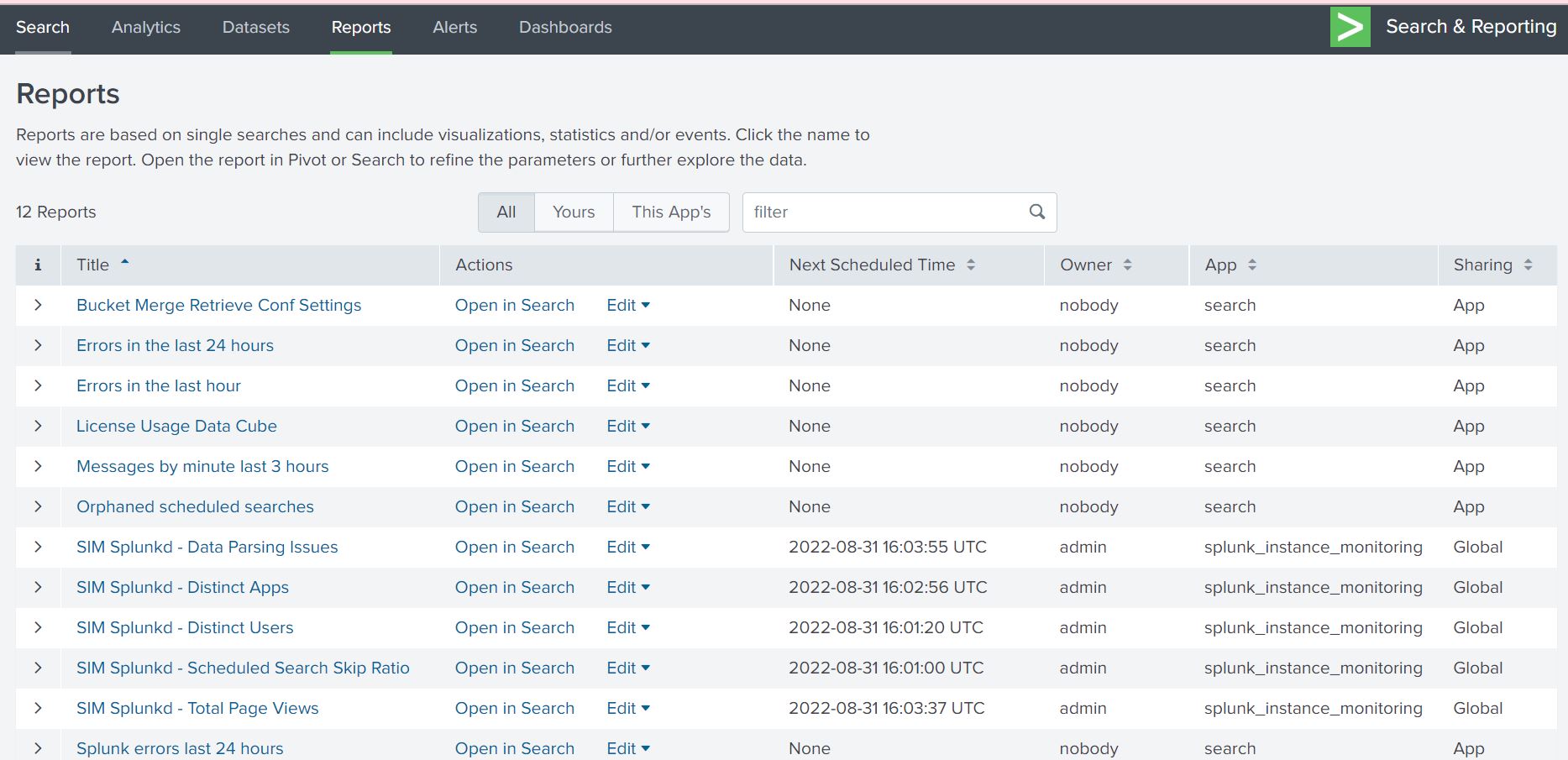


Figure 24 Report schedule screenshots

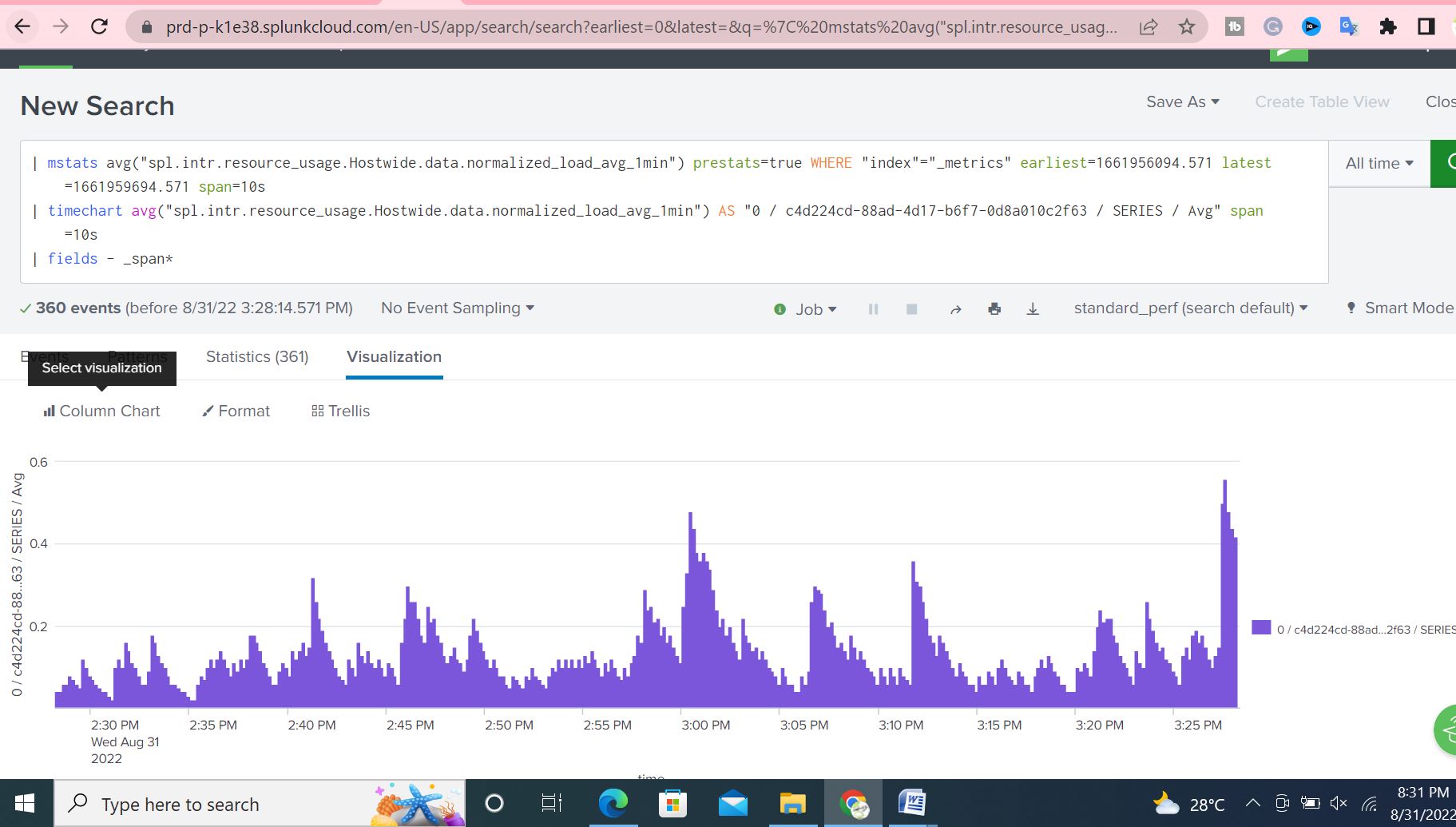


Figure 25visualizations

Splunk provides several apps that can give capabilities, but we used the "Splunk cloud based for our trial." SPL was used to preprocess the log data. Show in the figure depicts the SPL used for data preprocessing. We could choose the needed columns for our regression experiment using the SPL table command. Following that, we detect anomalous server activity automatically.

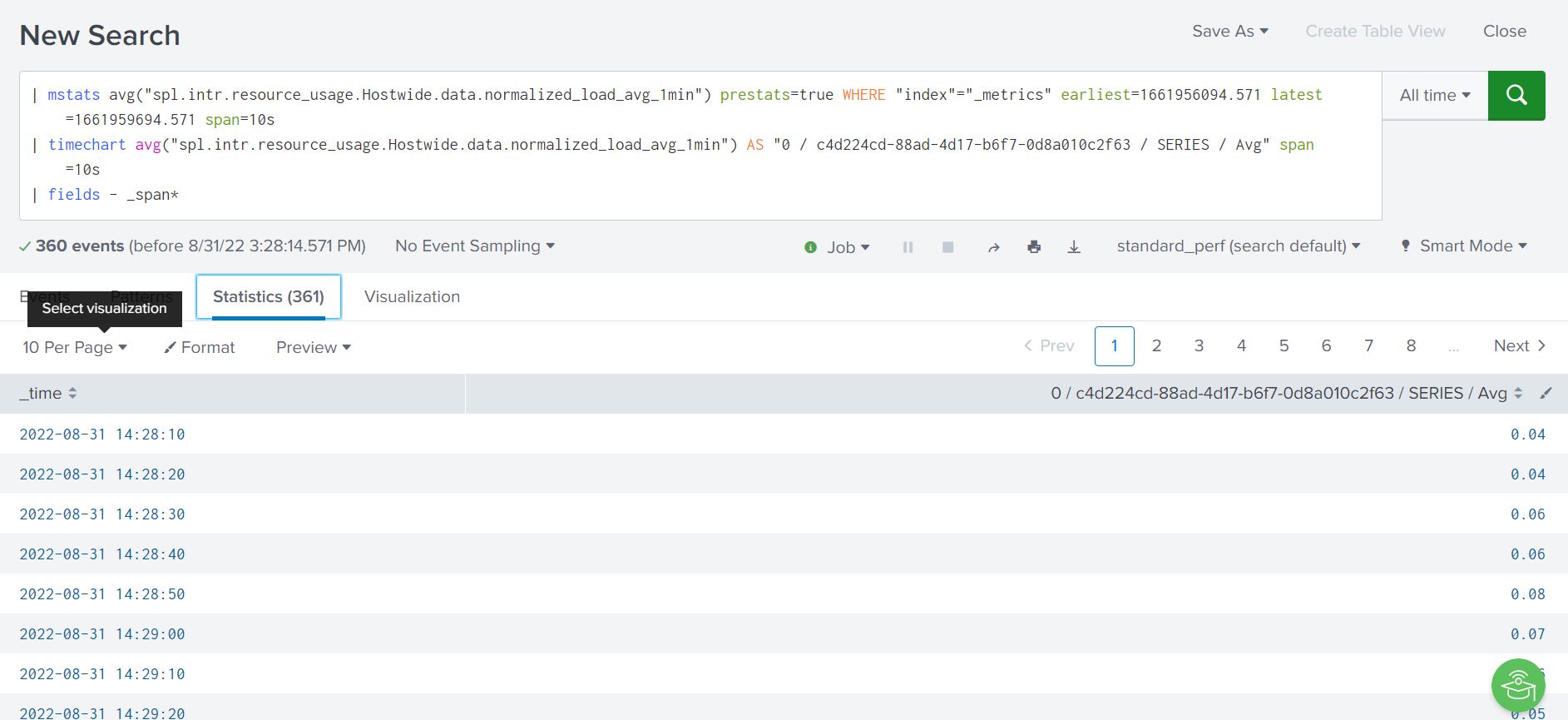


Figure 26 statistics analysis

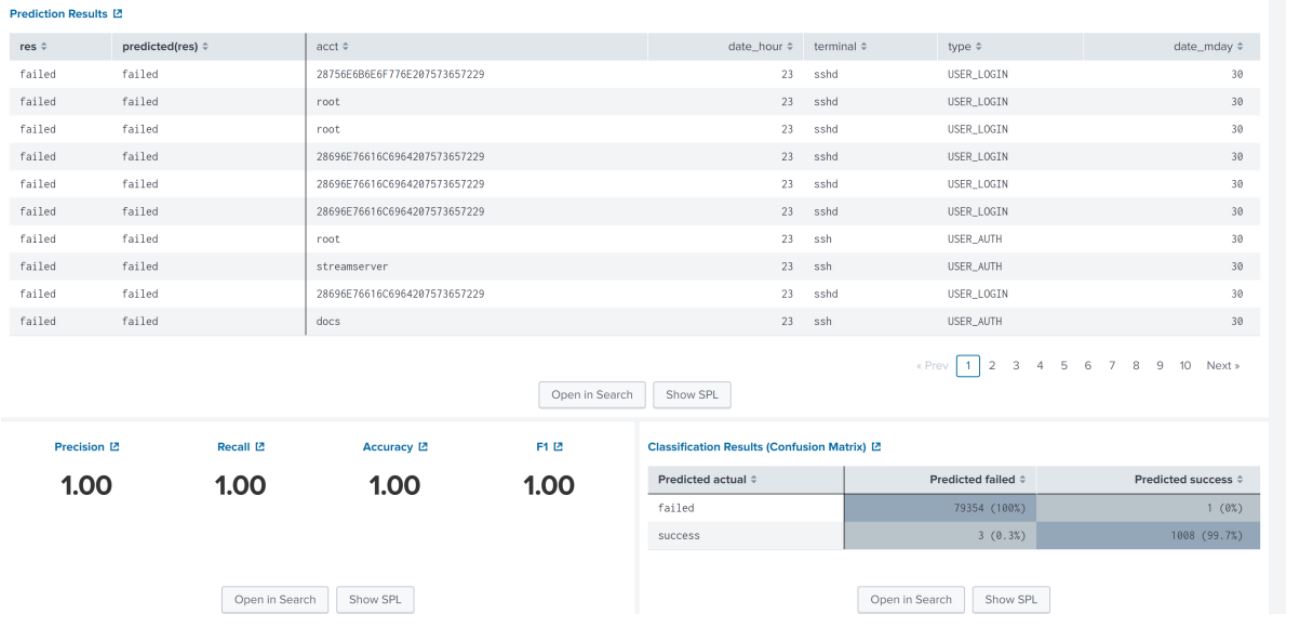
We also constructed an MLmodel (UEBA) to naturally distinguish uncommon conduct on the servers. Splunk has an assortment of applications that can give UEBA abilities, yet we worked with the "Splunk AI tool stash" to play out our analyze. We utilized SPL to preprocess the log information. Below is a screen shot shows the SPL utilized for preprocessing the information. By utilizing the SPL table order, we could choose the required segments for our relapse try



Figure 27 ML login result

LR was run utilizing the Splunk MLtool compartment on the pre-processed information to foresee the following login probability on the servers. Beneath shown the outcomes from the LR model.

Accuracy is to what extent achievement expectations were right. Our LR model had the option to anticipate 100 percent TP (genuine positive or Awareness or Recall) ineffective login and 99.7 % of TN (genuine negative or explicitness) fruitful logins. In our case, it seemed as though the model had the option to anticipate 99.711%.

Figure 28 Accuracy result

## CHAPTER 6: Discussion / Conclusions

## 6.1Discussion

The preceding experiment successfully postulated threat/anomalies conditions for detecting log packets in a restricted time frame. Testing tests based on system log files demonstrated that machine learning could efficiently identify malicious threats. Once the hypothesis has been transformed into actual logical code, the threat-finding tool may be reused. To learn from training data, linear regression requires numerous rounds.

We successfully set up a SIEM and developed Splunk cloud-based models using audited internal log data during this procedure. We also finished the Splunk configuration and demonstrated how to send an automatic response using the Splunk dashboard on static analysis and ML analysis.

The limits of my investigation are as follows:

1. The following Linear regression on the Splunk technique will want a massive training dataset.

2. In the initial experiment, data is labeled depending on circumstances; this may be time-consuming because it must be repeated for each batch of training data.

3. Because there is a chance that the data will be skewed toward a particular outcome, noise removal may be needed.

## 6.2 Conclusion

In this research, we set up the lab on Splunk; Machine Learning Algorithms can detect anomalies from system logs. We were impressed with the accurate predictions of the Splunk machine learning LR algorithm. Splunk SPL made it easy to create an alert report of the abnormal logins on the forwarded servers.

We were successful in developing automated notifications for failed system attempts. Based on ML linear regression algorithm analysis, we were able to successfully issue notifications of any inappropriate activity and execute an automatic remedy process.

## References

(n.d.). Retrieved from https://tools.kali.org/password-attacks/hydra

Abd Majid, M. a. (2021). "Model for successful development and implementation of Cyber Security Operations Centre (SOC). *PloS one 16, no. 11*.

admin. (2021). *80-of-soho-routers-contain-vulnerabilities*. Retrieved from https://www.infosecurity-magazine.com/news/80-of-soho-routers-contain-vulnerabilities/

admin. (2021). *wi-fi-routers-vulnerable-to-default-logins-attack*. Retrieved from https://cdn.comparitech.com/wp-content/uploads/2021/07/uSADF-wi-fi-routers-vulnerable-to-default-logins-attack-1.webp

admin. (2022). Retrieved from https://www.kali.org/

admin. (2022). *brute-force-attack*. Retrieved from https://phoenixnap.com/blog/brute-force-attack

Agarwal, A. K. (2021). Honey Encryption: Fortification Beyond the Brute-Force Impediment. *In Advances in Mechanical Engineering, pp. 673-681. Springer, Singapore*.

Agyepong E, C. Y. (2020). Challenges and performance metrics for security operations center analysts: a systematic review.". *Journal of Cyber Security Technology 4, no. 3 (2020): 125-152.*

Agyepong, E. Y. (2020). Towards a framework for measuring the performance of a security operations center analyst. *In 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*.

Al-Abassi. (2020). An ensemble deep learning-based cyber-attack detection in industrial control system. *IEEE Access 8*.

Al-Abassi. (2020). An ensemble deep learning-based cyber-attack detection in industrial control system. *IEEE Access 8*.

Alahmadi, B. A. (2022). 99% False Positives: A Qualitative Study of SOC Analysts’ Perspectives on Security Alarms. *In Proceedings of the 31st USENIX Security Symposium (USENIX Security), Boston, MA, USA,*.

Andrade, R. a. (2018). Enhancing intelligence SOC with big data tools. *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE*.

Arun Kumar, S. R. (2021). A survey on graphical authentication system resisting shoulder surfing attack. *In Advances in Artificial Intelligence and Data Engineering, pp. 761-770. Springer*.

Aung, W. P. (2020). Developing and analysis of cyber security models for security operation center in myanmar. *In 2020 IEEE Conference on Computer Applications (ICCA)*.

Aweya, J. (2021). IP router architectures: an overview. *International Journal of Communication Systems 14.5 (2001): 447-475.*

Basit, A. M. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems 76,*.

Basyurt, A. S.-A. (n.d.). Help Wanted-Challenges in Data Collection, Analysis and Communication of Cyber Threats in Security Operation Centers. *2022*.

Bazli, B. M. (2017). he dark side of I2P, a forensic analysis case study. *Systems Science & Control Engineering 5, no. 1* .

Bellovin, S. M. (1992). Encrypted key exchange: Password-based protocols secure against dictionary attacks.

Beulah M, P. M. (2022). Detection of DDoS Attack Using Ensemble Machine Learning Techniques. *InSoft Computing for Security Applications* .

BISCHOFF, P. (2021). *One in 16 home wi-fi routers tested vulnerable to default password attacks: report*. Retrieved from https://www.comparitech.com/blog/information-security/default-password-routers-study/

box, v. (n.d.). Retrieved from https://www.virtualbox.org/

Bryant, B. D. (2017). A novel kill-chain framework for remote security log analysis with SIEM software. *computers & security 67* .

Chen, Q. S. (2019). Automated ransomware behavior analysis: Pattern extraction and early detection. *In International Conference on Science of Cyber Security, pp. 199-214. Springer, Cham*.

Cloud-Based, M. D. (2019). Hou, Jiahui and Qian, Jianwei and Wang, Yu and Li, Xiang-Yang and Du, Haohua and Chen, Linlin. *2019 IEEE/ACM 27th International Symposium on Quality of Service*.

Conti, M. N. (2016). A survey of man in the middle attacks. *IEEE Communications Surveys & Tutorials 18.3* .

Couretas, J. M. (2022). Cyber Security and Defense for Analysis and Targeting. *In An Introduction to Cyber Analysis and Targeting,*.

Dave, K. T. (2013). Brute-force Attack ‘Seeking but Distressing. *Int. J. Innov. Eng. Technol. Brute-force 2.3*.

Davis, M. a. (2018). SOHO Router Security.

Demertzis. (2018). The next generation cognitive security operations center: network flow forensics using cybersecurity intelligence. *Big Data and Cognitive Computing 2*.

Demertzis, K. P. (2018). he next generation cognitive security operations center: network flow forensics using cybersecurity intelligence. *Big Data and Cognitive Computing 2, no. 4*.

Ding, J. C. (2022). A Data-Driven Based Security Situational Awareness Framework for Power Systems. *Journal of Signal Processing Systems*.

Duna, Y. T. (2021). Grasp on Next Generation Security Operation Centre (NGSOC): Comparative Study. *Int. J. Nonlinear Anal. Appl 12, no. 2*.

Dwivedi, S. M. (2022). Defense against distributed DoS attack detection by using intelligent evolutionary algorithm. *International Journal of Computers and Applications 44.3* .

economics, P. m. (2006). Brian Monahan. *economic of securing network*.

Eskelinen, T. (2022). Development of Open-Source SIEM and Security Operation Centre in A Company.

Faulkner, B. (2007). Hacking into data breach notification laws. *Fla. L. Rev. 59*.

Golightly, L. V. (2018). Towards Ethical Hacking-The Performance of Hacking a Router and Lessons Learned.

Golightly, L. V. (2021). Towards Ethical Hacking—The Performance of Hacking a Router. *In Information Security Technologies for Controlling Pandemics, pp. 435-461. Springer, Cham*.

Gupta BB, C. P. (2022). Smart defense against distributed Denial of service attack in IoT networks using supervised learning classifiers. *Computers & Electrical Engineering. 2022 Mar 1;98:107726.*

Happa, J. I.-R. (2021). Assessing a decision support tool for soc analysts. *Digital Threats: Research and Practice*.

Haslum, K. A. (2007). Dips: A framework for distributed intrusion prediction and prevention using hidden markov models and online fuzzy risk assessment. *In Third International Symposium on Information Assurance and Security, pp. 183-190. IEEE*.

Huang, Z. Q. (2020). A survey on machine learning against hardware trojan attacks: Recent advances and challenges. *IEEE Access 8* .

Idika, N. a. (2007). A survey of malware detection techniques. *Purdue University 48.2*.

Jacobs, P. A. (2013). Classification of security operation centers. *In Information Security for South Africa, pp. 1-7. IEEE*.

Jindal, N. (2020). Automated event prioritization for security operation center using graph-based features and deep learning. *PhD diss.*

Khaleghi, M. M. (2022). Comprehensive Comparison of Security Measurement Models. *Journal of Applied Security Research*.

Knight, S. W. (1998). Virtual router redundancy protocol .

Kokulu, F. B.-J. (2019). Matched and mismatched socs: A qualitative study on security operations center issues. *In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*.

Kook, J. G. (2016). Hacking Countermeasures for Wireless Internet Service. *Journal of Service Research and Studies 6, no. 3*.

Kumar, S. B. (2021). A Semantic Machine Learning Algorithm for Cyber Threat Detection and Monitoring Security. *In 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 1963-1967. IEEE,*.

Kwak, M.-J. H.-K. (2018). Rhizosphere microbiome structure alters to enable wilt resistance in tomato. *Nature biotechnology 36, no1*.

Kwon, S.-H. a.-W. (2012). Hacking and security of encrypted access points in wireless network. *Journal of information and communication convergence engineering 10, no. 2*.

Lewis, J. A. (2022). Assessing the risks of cyber terrorism, cyber war and other cyber threats. *Washington, DC: Center for Strategic & International Studies,*.

Mahamadu, A.-M. L. (2013). Challenges to BIM-cloud integration: Implication of security issues on secure collaboration. *2013 IEEE 5th International Conference on Cloud Computing Technology and Science. Vol. 2. IEEE*.

Majeed, A. F. (2019). Near-miss situation based visual analysis of SIEM rules for real-time network security monitoring. *Journal of Ambient Intelligence and Humanized Computing 10, no. 4*.

Mallik, A. (2019). Man-in-the-middle-attack: Understanding in simple words. *Cyberspace: Jurnal Pendidikan Teknologi Informasi 2.2*.

Marcus, D. J. (2018). The Data Breach Dilemma: Proactive Solutions for Protecting Consumers' . *Personal Information." Duke LJ 68* .

McLeod, A. a. (2022). Information security policy non-compliance: Can capitulation theory explain user behaviors? *Computers & Security 112 (2022):* .

Nicula, S. a.-D. (2021). Technical and Economical Evaluation of IOT Attacks and their Corresponding Vulnerabilities. *Informatica Economica 25, no. 1 (2021): 31-41.*

Oh, C. J. (2021). A Survey on TLS-Encrypted Malware Network Traffic Analysis Applicable to Security Operations Centers. *Applied Sciences 12, no. 1*.

Onwubiko, C. (2015). Cyber security operations centre: Security monitoring for protecting business and supporting cyber defense strategy. *2015 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA).* .

Oprea, A. Z. (2018). Made: Security analytics for enterprise threat detection. *In Proceedings of the 34th Annual Computer Security Applications Conference, pp. 124-136.* .

Osman, A. I.-S. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *in Shams Engineering Journal 12, no. 2*.

Porta, I. (2020). *Brute force attack with Hydra and Kali Linux*. Retrieved from https://gtrekter.medium.com/brute-force-attack-with-hydra-and-kali-linux-3c4ede55d119

Rashidi, B. C. (2017). Android resource usage risk assessment using hidden Markov model and online learning. *Computers & Security 65*.

Resul, D. A. (2020). Analysis of cyber-attacks in IoT-based critical infrastructures. *International Journal of Information Security Science 8.4*.

Sagi, O. a. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8, no. 4* .

Scholten, C. P. (2019). Hacking the router: characterizing attacks targeting low-cost routers using a honeypot router. *Bachelor's thesis, University of Twente,* .

Sendi, A. S. (2012). Real time intrusion prediction based on optimized alerts with hidden markov model. *Journal of networks 7, no. 2* .

Shahjee, D. a. (2022). "Integrated Network and Security Operation Center: A Systematic Analysis (February 2022). *IEEE*.

Shmelkin, R. L. (2021). Generating Master Faces for Dictionary Attacks with a Network-Assisted Latent Space Evolution. *In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pp. 01-08.*

Simma, K. C. (2019). Wi-Fi router network-based occupancy estimation to optimize HVAC energy consumption. *In Proceedings of the CSCE Annual Conference, pp. 1-10. 2019*.

Singh, P. a. (2021). Attack and intrusion detection in cloud computing using an ensemble learning approach. *International Journal of Information Technology 13, no.2*.

Sinha, M. S. (2021). Sniffer: A machine learning approach for DoS attack localization in NoC-based SoCs. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems 11, no. 2*.

Sopan, A. M. (2019). Machine Learning for Security Operation Center.

Stiawan, D. M. (2019). Investigating brute force attack patterns in IoT network. *Journal of Electrical and Computer Engineering 2019*.

Stojmenovic, I. a. (2014). The fog computing paradigm: Scenarios and security issues. *In 2014 federated conference on computer science and information systems, pp. 1-8. IEEE*.

Susukailo, V. I. (2020). Analysis of the attack vectors used by threat actors during the pandemic. *In 2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT), vol. 2, pp. 261-264*.

Syed, R. (2019). Enterprise reputation threats on social media: A case of data breach framing. *The Journal of Strategic Information Systems 28,*.

Tama, B. A. (2015). Performance analysis of multiple classifier system in DoS attack detection. *Tama, Bayu Adhi, and Kyung Hyune Rhee. "Performance analysis of multiple classifier system in DoS attack detection." International Workshop on Information Security Applications. Springer, Cham*.

Threats, U. a. (2015). Venkatesh Jaganathan ,1 Priyesh Cherurveettil ,1 and Premapriya Muthu Sivashanmugam1. *The Scientific World Journal, vol.* .

Tianqi Chen, T. H. (2017). xgboost: eXtreme Gradient Boosting.

Veroni, E. C. (2022). A large-scale analysis of Wi-Fi passwords. *Journal of Information Security and Applications 67* .

Zhang, J. Y.-T. (2019). "Cross-site scripting (XSS) detection integrating evidences in multiple stages. *In Proceedings of the 52nd Hawaii International Conference on System Sciences.*

Zulkifli, M. A. (2018). Live forensics method for analysis denial of service (DOS) attack on routerboard. *Int. J. Comput. Appl 180, no. 35* .

## Appendices

## Appendix 1

| mstats avg("spl.intr.resource\_usage.Hostwide.data.normalized\_load\_avg\_1min") prestats=true WHERE "index"="\_metrics" earliest=1661956094.571 latest=1661959694.571 span=10s

| time chart avg("spl.intr.resource\_usage.Hostwide.data.normalized\_load\_avg\_1min") AS "0 / c4d224cd-88ad-4d17-b6f7-0d8a010c2f63 / SERIES / Avg" span=10s

| fields - \_span\*