Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/58

An ELK Stack Method with Machine Learning Algorithm for Alerting Traffic anomaly

Dibyahash Bordoloi[1], Surendra Shukla[2]

[1]Head of the Department, Department of Computer Science & Engineering,Graphic Era Hill University, Dehradun, Uttarakhand India, 248002
[2]Department of Computer Science & Engineering,Graphic Era Deemed to be University, Dehradun, Uttarakhand India, 248002

ABSTRACT

A network's logs are still not required to be flawless in perpetuity; this is not a requirement. It is inevitable for the behaviour of data traffic to occasionally deviate from what is expected, when this occurs, the behaviour of traffic is referred to as being anomalous. Troublesome traffic can arise for a variety of reasons, including external attacks, the transfer of obsolete data, or even the act of serving people for networking businesses. This becomes a significantly more difficult challenge to overcome when the extent of the network is on a grander scale. Either the anomaly detection systems that are currently in place were trained with outdated datasets, or they lack the capacity to handle large loads in an effective manner. Therefore, there is a requirement for a scalable solution that can provide security to a network by detecting anomalies within the network itself and notifying with a fast reply whenever an anomaly takes place by gaining knowledge from the network's previous behaviour. In this paper, an end-to-end solution for the effective introduction of a machine learning-based anomaly detection method into an enterprise environment is presented. This solution starts with the collection of log data and continues all the way through to the generation of alerts. In order to put this into action, a framework is put in place that consists of Logstash, Kibana and Elasticsearch, as well as additionally, machine learning is a component of it.

Keywords: anomaly detection, machine learning, Alert

INTRODUCTION

The current climate has created the conditions for everyone to observe a paradigm shift, which can be seen in the rise in internet use and the digitalisation of things that were previously done offline. Because of the rise in the number of users and the volume of traffic on networks, there is a pressing requirement for such a raise in the level of security provided by these networks. A distributed denial of service (DDOS) attack, an out-of-date variant of a client's operating system, or even poor network performance can all disrupt the normal operations of an organisation. When an organisation has grown to the point where it serves millions of customers through its network, it becomes necessary to search for an anomaly detection solution that is both scalable and protected and has the ability to produce the desired outcomes. In a setting where the safety of a network infrastructure is paramount,

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/58

the passage of time is of the utmost importance, and therefore, the system in question must be able to provide a prompt response.

Companies just goes to demonstrate how significantly the proliferation of digital engagements has had an impact on the subject of cybersecurity, and it has been predicted that this trend will only continue to rise from here on out. In addition to this, it places an emphasis on the requirement for improved security systems that are able to anticipate security flaws rather than having to solve these problems as they appear. Threats in the network infrastructure, such as either hardware or software faults, make it possible for third parties to take advantage of those vulnerabilities. A persistent awareness of new problems that are likely to emerge is required, followed by the determination of an activity that will target those problems. In the last ten years, machine learning has become increasingly popular in the field of intrusion prevention systems [1] where it has been used to improve detection performance as well as adaptability. In addition, supervised learning is superior to unsupervised learning when it comes to dealing to data that contains known attacks. However, supervised learning is superior when it comes to handling unexpected types of data. Elasticsearch is becoming widely attractive as a storage system, and large organisations such as GitHub, among others, are adopting it to manage their ever-increasing amounts of data and the storage of it [2]. Mining contemporary data sets also turns out to be a promising route for researchers to investigate. Its speed in real-time data analysis, which is made possible by its one-of-a-kind data sharding and standardisation process, is another factor that contributes to its widespread adoption and popularity [3]. The purpose of this paper is to suggest a prompt and expandable anomaly based system that utilises machine learning in conjunction with Kibana, Elasticsearch and Logstash also known collectively as Elastic stack. This system will analyse network logs, identify anomalies, and generate alerts all in real time. The administrator will then be able to take appropriate action in response to the alerts. This solution is geared specifically toward production environments, which typically generate significant amounts of network log data. The sections of the paper organisezd as follows: section II discusses related tasks in this domain; section III gives information on the dataset that was utilised; section IV discusses the performance measures that were selected for analysis; and section IV describes the methodology and components of the ELK infrastructure that was utilised. In section V, some perspectives and results of the system are presented. In the VI section of the paper, the conclusion is presented.

RELATED WORK

Every location with a network also has the potential for a security breach in that location. There are a lot of potential dangers in the modern world, the most common of which are infections caused by viruses, malware and worms. Not only that, but in addition to the fact that security systems are always being updated in order to deal with these threats, the threats themselves are always being updated in order to avoid being identified on the network.

It is mentioned in [10] that there are some domains of networks, such as the defence and security and hospital data, that must not be disrupted under any circumstances; however, even though this is the case, these domains continue to be at risk. The absence of adequate cyber security facilities, in addition to the rapid advancement of technological capabilities, highlights the requirement for improved lines of defence. In this instance, it was suggested that the data be encrypted using

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/58

symmetric and asymmetric encryption to protect its confidentiality, that backups be created to guarantee its integrity, and that the most recent software be used to guarantee its availability. Encryption is a good first step forward into protecting sensitive information, but we require a solution that can protect all types of data and is not as computationally intensive as encryption.

Xiaokui Shu investigated a solution [4] that included digital signatures, adaptive warning strength, connections among alerts, as well as a zero trust strategic plan. However, it was discovered that digital certificates are hard to use on client computer systems because those computer systems may have programmes running that were not provided by the identified seller, and it is difficult to process a large number of security alerts. In addition, the Zero Trust plan called for an enormous amount of computing capability to be utilised in order to monitor traffic. This is because the Zero Trust method attempts to defend against both outsider and insider attacks. Because of this, it was difficult to scale up in extremely large networks. Systematic Literature Reviews, also known as SLRs, have already been carried out in the past upon that data mining techniques that have been utilised in intrusion detection systems [5]. The findings of those reviews came to the conclusion that the process of data extraction was not an easy one. Real-time surveillance has revealed a variety of difficulties in research as a result of the growing quantity of networks every year. In addition, the majority of the existing intrusion detection systems make use of obsolete datasets such as KDD'99, which has been established in 1999, a time when nearly half of the dangers that exist presently did not exist. A comparison of adaptive graph-based optimization approaches and one-class support vector machines (SVM) has been done for the purpose of identifying anomalous activities occurring on a network [6]. However, due to space limitations, the method was only employed with two datasets, and it was only applied in two distinct contexts; as a result, the performance could not be generalised. Even taking into account the increased complexity of the approach, there is room for advancement. The most difficult problem that content-based data leak prevention and detection (DLPD) systems for businesses have to deal with is scalability. This means that these systems are incapable of process high quantities of network logs in a timely manner [7]. A great number of papers have been written about anomaly systems that are able to detect anomalies from training and test datasets that have been obtained and saved. However, network traffic is stored as a time series, and these solutions do not have the capacity to perform time project or provide real-time alerts when they detect anomalies. They were trained and tested with stale and unpredictably accurate data, so their performance in real-world scenarios with actual data was poor. In addition, they did not perform well. Techniques of supervised learning suggested in some papers would not function well when applied to unpredictable data; consequently, unsupervised machine learning must be investigated. In conclusion, the solution must be scalable in order it to be suitable for use in businesses.

As a result, the paper makes a proposal for an anomaly based system that makes use of unsupervised learning to model normal behaviour and predict unexpected problems, has the ability to deal with large amounts of data, and can generate alerts based on a flow of records that is produced in real time.

DATASET

Anomaly detection method may still be vulnerable to becoming out of date despite the fact that the

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/58

datasets they use 10 years of span. Thus, here we employ an organization's actual, real-time network logs. The usable data is comprised of six months' worth of network logs beginning in January 2021 and continuing to date, with new data being added as and when it is created. The logs were gathered from a setting where applications are extensively tested, so there are also examples of failure that could be leveraged. The proposed solution is intended to scale, which means that it should be able to handle huge amounts of network logs. This necessitates the availability of corresponding storage space. There are roughly 20 GB worth of data generated over the course of 6 months. Packet capture files for one-minute windows in datasets like DARPA2009 are typically 1 GB in size [8]. Therefore, it is preferable to utilise enterprise data consisting of only the essential fields. Memory requirements for machine learning infrastructure, known as the trained model, can range from a few hundred kilobytes to several gigabytes when dealing with complex problems. As a result, the design memory limit (the maximum amount of memory that can be allocated as part of a given model) is 1 GB by default. If the user so chooses and sufficient memory is available, the cap can be increased.

METHODOLOGY

The system as a whole relies heavily on each individual part. obtaining the expected outcomes from anomaly detection. Rhythms for Elasticsearch is indeed the storage component, as well as Kibana provides visualizations for investigating trends and shift patterns in the data by means of streaming data; Logstash is used for formatting and processing the data prior to its transmission. Machine learning's role in Elastic stack is to foresee and prevent problems. Anomaly detection is the practice of monitoring for and responding to unusual occurrences. Fig. the system's flow is shown in Fig. 1. Disturbing behaviour from the system using unsupervised learning to detect patterns in time series data. This methodology incorporates correlation analysis, time series decomposition, Bayesian distribution modelling and clustering.

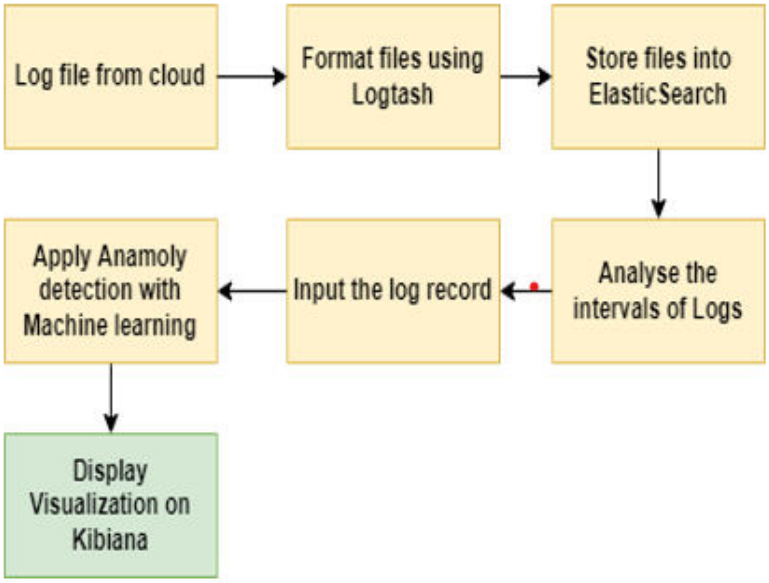

Fig 1: Workflow Diagram

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/58

A. LOGSTASH

Data is gathered by Logstash, which in this case consists of network logs, besides listening to a given path and then sending the collected data to an end destination, which in this case is indeed an Elasticsearch cluster. Logstash additionally formats the data, which up until now have only been values, by transforming them into key-value pairs. Because of this, Logstash builds a pipeline to transport the data from its starting point to its final destination, while also applying the necessary formatting to the data along the way.

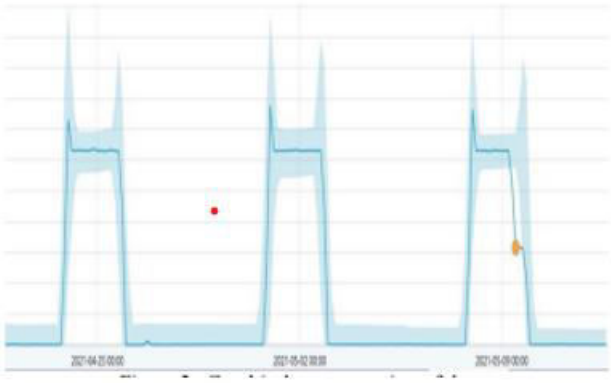

Fig 2 : Graphical representation of Data

B. ELASTICSEARCH

The stack's memory is represented by Elasticsearch. It's made up of a group of nodes about which network logs are partitioned and stored in a formatted fashion, with replicas set up as needed. Here, the logs undergo daily indexing, with new indexes being created for every date. The indices could be queried to obtain this information, and it can then be aggregated into a larger dataset for anomaly detection based on the time interval of interest.

C. KIBANA

As the primary means of communication between the user and the anomaly detection method, Kibana acts as the interface. The data, the normal behaviour, and the outliers are all graphically represented. It can create graphical images, such as heatmaps in an anomaly explorer or illustrations representations in a single metric viewer. The available options make it possible to create and configure a dashboard that can show all relevant data in a digestible format. Co-ordinate maps, gauges, line graphs, and pie charts are just a few examples of [9] these. Kibana also includes a Graph feature, where access points and corners can be used to display the connections between entities of your choosing.

ANALYSIS AND RESULT

A. VISUALIZATION OF DATA

Kibana's method predicts the typical value to fall within a shaded region. That's the normal way of doing things. An anomaly is noted if the observed value deviates significantly from this pattern. To achieve this, it computes the likelihood that a given observation is an outlier and displays the result on a scale from one to one hundred for straightforward interpretation. The anomaly score uses four color-coded categories to indicate increasing degrees of severity: warning (scores of 30 and above),

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/58

minor (scores of 70 and above), major (scores of 80 and above), and critical (scores of 0 and above).

FILTERING DATA

Occasionally, it is undesirable for a certain metric value to be included in the anomaly detection process. To do this, there are filtration lists that can be used to eliminate any information that matches the value systems on the list, and bespoke regulations can be implemented to forego model latest update and results altogether.

B. ELK COMPARED TO SPLUNK

Splunk, like Elastic stack, is a log analysis tool, but it takes a slightly different approach. According to Google, the ELK stack is currently more popular than Splunk and is only expected to grow in popularity as new features are added. This transition with the majority of the change occurring between 2014 and 2016. Similarly, retrieving one billion records using Elastic's recommended setup took one minute and fourteen seconds, while using Splunk's recommended setup took approximately one minute and twenty-two seconds, and therefore effectiveness of the Elastic stack is slightly better.

C. PERFORMANCE ANALYSIS

Due to the observation of each partition of logs, inter jobs are found to be the most space-intensive of the 3 machine learning tasks discussed. Kibana requests and their respective response times are graphed in Fig. 3.The median response time for two separate client requests made to Kibana was 1 millisecond. Kibana's responsiveness and overall health both increase when its load is optimised. Seeing as Kibana's interface is where most users will be using the system, it is a crucial factor to think about.

Fig 3: Kibiana Performance

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/58

CONCLUSION

The paper suggests a viable option for deploying a machine learning-based system for anomaly detection in business setting. Companies with a lot of client network traffic will benefit greatly from this solution. Since the proposed solution is highly scalable and can handle huge amounts of log data, it will serve as a benchmark for upcoming anomaly detection methods.

REFERENCE

1. M. Zamani, "*Machine Learning Techniques for Intrusion Detection*", arXiv:**1312.2177, 2013**.
2. O. Kononenko, O. Baysal, R. Holmes and M. W. Godfrey, "*Mining modern repositories with Elasticsearch*", In the Proceedings of the 2014 Conference on Mining Software Repositories, **India**, pp. **328–331, 2014**
3. N. Shah, D. Willick and V. Mago, "*A framework for social media data analytics using Elasticsearch and Kibana*", Wireless Networks, Vol. **24**, Issue.**8**, pp. **1-9, 2018.**
4. X. Shu, K. Tian, A. Ciambrone and D. Yao, "Breaking the Target: An Analysis of Target Data Breach and Lessons Learned", arXiv:1701.04940, 2017.
5. F. Salo, M. Injadat, A. B. Nassif, A. Shami and A. Essex, "Data Mining Techniques in Intrusion Detection Systems: A Systematic Literature Review", in IEEE Access, Vol. 6, pp. 56046-56058, 2018.
6. S. D. Bhattacharjee, Y. Junsong, Z. Jiaqi, Y. Tan, "ContextAware Graph-Based Analysis for Detecting Anomalous Activities", In the Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), China, pp. 1021- 1026, 2017.
7. L. Cheng, F. Liu and D.Yao, "Enterprise data breach: causes, challenges, prevention, and future directions", WIREs: Data Mining and Knowledge Discovery, Vol. 7, Issue.5, 2017.
8. N. Moustafa and J. Slay, "Creating Novel Features to Anomaly Network Detection Using DARPA-2009 Data set", In the Proceedings of the 2015 14th European Conference on Cyber Warfare and Security ECCWS-2015, UK, pp. 204-212, 2015.
9. P. P. Bavaskar, O. Kemker and A. Sinha, "A SURVEY ON: "LOG ANALYSIS WITH ELK STACK TOOL", International Journal of Research and Analytical Reviews (IJRAR), Vol. 6, Issue.4, pp. 965-968, 2019.
10. S. J. Son and Y. Kwon, "Performance of ELK stack and commercial system in security log analysis", ", In the Proceedings of the 2017 IEEE 13th Malaysia International Conference on Communications (MICC), Malaysia, pp. 187- 190, 2017