UNIVERSITY OF THE WEST of SCOTLAND

UWS

What are some of the most common attacks performed by an advanced adversary within a network, and what does this traffic tell us?

Using elastic search to identify patterns within network traffic

Michelle Pantelouris
B00389651

Thesis Project for the partial fulfilment
of the requirements for the Master Degree
in Information Technology/E-Business Management /
Advanced Computer Systems Development (delete one)

University of the West of Scotland
School of Computing
3RD July 2023

Table of Contents

Acknowledgements

The author of this project would like to thank the following people for their help and support throughout:

Graham Parsonage for giving me guidance on what he was expecting from my project.

Shaun Whorton who mentored me through this whole project.

1.0 Introduction

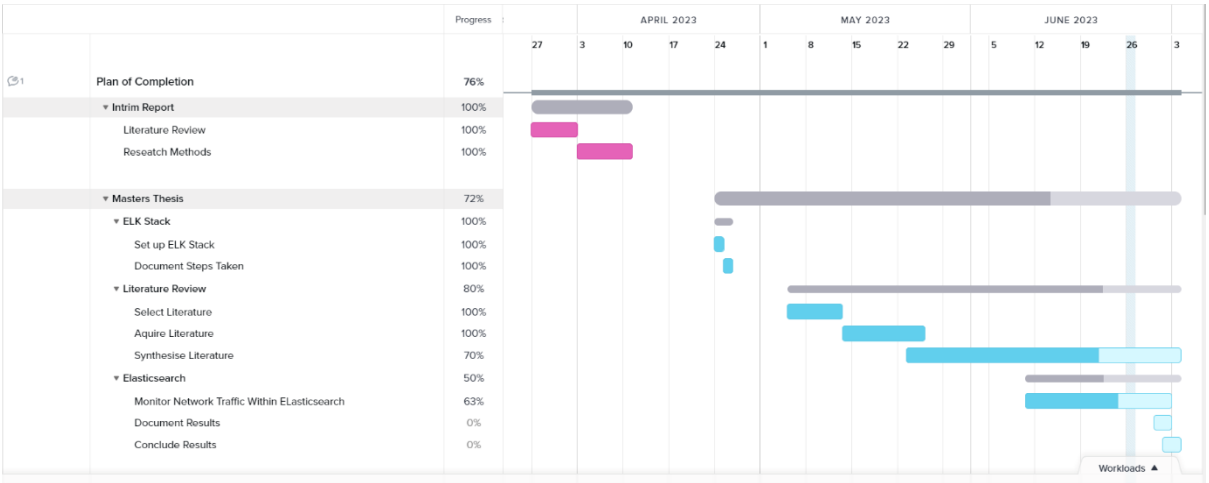
With cyber-attacks increasing each year as stated by Forbes who say the cost of cybercrime is predicted to be $8 trillion in 2023 and going to rise to $10.5 trillion in 2025 (Brooks, C. (2023)). The need to reduce or mitigate these attacks is also increasing. The topic that will be discussed in this report is elasticsearch. Which is used to monitor traffic for any suspicious activity or attacks. Using pre-existing attack data [1] obtained from multiple honeypots, this project will aim to gain further insight into how threat actors conduct attacks, the services and ports they attack and the characteristics of said attacks. The project will aim to utilise machine learning to identify outlying data. A brief overview of what elasticsearch and the elk stack are will be written, then summarise some recent literature regarding this topic, along with how data analytics relates to cyber security and what anomalies are and how they are used. The methodology that will be implemented and how it related to this project will also be discussed. The setup of the ELK Stack is located in Appendix A. The progress documentation can be found in Appendix B.

1.1 Plan of Completion

The first step to competing this project is setting up the ELK stack on VMware, and then writing up the steps it took to set it up. The deadline for this is one day but two days if any problems occur.

The next step is to write a literature review on research that is related to this project. A detailed description of what the ELK stack and Elasticsearch are and how data analytics relates to cyber security will be discussed as Anomalies will be used within Elasticsearch. The deadline for this is two months with a hard deadline of 3rd of July.

Lastly network traffic CSV will be imported into Elasticsearch and analysed for patterns with regards to attack metrics and from here conclusions can be formed on the most common attacks. The deadline for this is three weeks.



[1] https://www.kaggle.com/datasets/casimian2000/aws-honeypot-attack-data

2.0 Literature Review

2.0.1 Overview of the ELK Stack

ELK is the abbreviation of the projects Elasticsearch, Logstash and Kibana. With elasticsearch being the most popular. With Elasticsearch the data can be visualised more easily.

Elasticsearch is search and analysis system that was created to respond a vast number of used cases. The types of data that can be analysed are structured, semi structured, or unstructured, and to launch different types searches.

Logstash is a server-side pipeline that is used to process data. Its main aim is to integrate data from a vast number of sources. Once this is completed it will transform and send them to the elasticsearch.

Kibana is used to visualise the stored data within elasticsearch. It has the capability to take the data that has been collected and then build analytical graphs, diagrams, and data tables, which are able to be shared.

There are some interesting features such as search performance, indexing and scalability. Safety was not considered as the main objective when the elastic stack was first developed. (Abdou, F, Lemine, M, Farouk, M. (2019)).

2.0.2 Advantages of the ELK Stack

There are various advantages to using the elk stack as discussed in the paper 'Automated Switching Crash Analysis using ELK for Log Analysis'.

Unlike other engines such as SPLUNK which can cost up to $400 a month, the elk stack is free. It provides large storage space and dynamic allocation resources, unlike Splunk and Mongo DB which are limited with the amount of storage, which is usually only 256 MB. This making it useless when it comes to using crash files. Security features required are provided and helps user privileges are restricted for their protection. It has a user-friendly GUI 'Kibana' that has a lot of filtering options. (Supreeth, S, Rajendran, S. (2022)).

2.0.3 Disadvantages of the ELK Stack

Complex Management – The ELK stack is downloaded in large amounts every month and downloading the software is the easy part. There is a multiple step procedure that is used to deploy the stack. This includes configuring log parsing and ingestion, creating a data pipeline, making sure the handled exceptions are secure and avoiding any data loss. These are just a few of the steps needed. The creation of a data pipeline is crucial for any organisation that deal with large amounts of data and rely on insights that are accurate and timely, which is used to drive decision-making. Deploying and managing the ELK stack can be

quite complex, and if corporations do not have the necessary help and mastery they will need to invest in training or employ an expert an ELK stack developer.

High Ownership Costs – Although the ELK Stack software is free, resources will be needed to build, grow, and maintain the infrastructure. The cost of data storage will be dependent on if it is deployed on-premises or in the cloud. To deploy the ELK stack, there will need to be at least one full-time worker who will configure, sustain, and customise it. The cost of the infrastructure will grow over time. This shows that open-source do not always mean low cost.

Tranquillity & Uptime Problems – Elasticsearch indices are the main reason of the fluctuation of the ELK stack. These indices contain records with log data, which Elasticsearch will query or interpret, which have been reported by users. If an index size exceeds the limitations of the nodes data storage, then the indexing starts to fall, which can lead to the loss of data or crashes.

Data Retention Trade-offs – There might be data usability issues and trade-offs for customers between data retention and expenses when the volume of data boosts. This is due to the two distinguishing features of Elasticsearch: Sharding and Replicas. More resources, disk space and nodes will be needed to gain the full benefits of Sharding and duplication.

Scaling Issues – The ELK stack has some scaling challenges and these are due to a number of factors: the large indices are unstable, replication and sharing the data is not economically sound, a rapid increase in TCO caused by organisations increasing the daily ingestion of log files. The ELK stack is scalable but with the challenges and the cost the benefits are outweighed, especially if it is compared to some alternatives. (Ali, H. (2023)).

2.1 Overview of Elasticsearch

Elasticsearch is full-text search engine that is written in Java and is open source. It has been designed in a way that makes it distributive, scalable and its capability is near real-time. It is easy to use and is set with default configurations which is more than enough for a standalone use without needed to tweak anything. Users usually fine tune several of the parameters. Node is the name for running an instance, and elasticsearch cluster is when two or more nodes form. All an elasticsearch cluster needs to get set up is a value within the configuration file which is the name of a cluster. Nodes are discovered buy the elasticsearch within the network and then bind them together int clusters. (Kononenko, O, Baysal, O, Holmes, R, Godfrey, M. (2014)).

2.1.1 History of Elasticsearch

During 1999 an interpretation of a search engine called Lucene was created by Dough Cutting. The was powerful but could only be run on a single machine. The more it evolved the stronger the open-source community became.

The creator of Elasticsearch, Shay Bannon built an Object Mapping framework called ‘Compass!’ in 2006. The second version of compass later on and he was planning on creating a third but realised he wanted something larger.

During 2009 he stopped maintaining Compass when he came to the conclusion that search needed to be in all applications. This is when he created Elasticsearch which would be an open-source, distributed RESTful engine for every kind of application and data.

Elasticsearch was realised to the public in 2011 the Elasticsearch Inc. was formed in 2012 along with Logstash and Kibana, and with all combined create the ELK stack. Elasticsearch was offered as a service in the cloud during 2015.

During 2021 there were disputes with AWS and other vendors who were taking advantage of Elasticsearch so they had to change their open-source license. (Souris, S. (2022)).

2.1.2 Background of Elasticsearch

Although traditional RDBMSs and elasticsearch are different in many ways, there are analogues in a lot of the core concepts in the RDBMS at the higher levels of Elasticsearch. All of the data within Elasticsearch is stored indices. Indexes in Elasticsearch are the same as databases in RDBMS. It has the ability to store various types of documents, update them and search for them. The documents are a JSON object, which is very similar to a row in a table within RDBMS. The number of fields that are within a document is zero or more. These fields are either a primitive type or a structure that is more complex. Even though the documents have a document type they are schema-free which means two documents that are the same can have different sets of fields. Document type can define the set of fields of a specific document.

Apache Lucene is what Elasticsearch is based on, and each index consists of Lucene indices which are called shards. The value of the amount of shards in an index is fixed before the index is created. The job of the Elasticsearch server when a document is added to the index is to define the shards that are responsible for indexing and storing the document. It does this by balancing the loads between any shards that are available, and improving the performance overall. This is because the shards can be used concurrently. Automating shards is only one key part of the distributed nature of Elasticsearch, the other is automatic distribution of shards between the nodes in a cluster. An example of this is if we have 6 shards, they will all be added to the same node, then if another shard is added after half of it will be added to a new node which will result in there being two nodes with three shards each. No matter how many shards within an index, or the amount of nodes that occupy it, the index will always be seen as a single entity to a client. (Kononenko, O, Baysal, O, Holmes, R, Godfrey, M. (2014)).

2.1.3 Strengths

There are several strengths to using elasticsearch some of which has have been discussed within this research paper. They mention 3 main ones, scalability, agility, and performance.

Scalability – A rational dataset is usually favoured for storing data on elasticsearch, though the data is usually stored within a single database. This means that if more data is stored it will need a more powerful server. The database will need to be sharded as the server can be pushed to its limit very quickly. The shards will then be separated into different servers. Shards of an index are automatically distributed across nodes of clusters by elasticsearch which will load them equally. If anymore data is needed elastic search is the best choice as it scales horizontally.

Agility – When thinking of agile data, we think of the number of updates or new records, it constantly changing structure of a logical piece of information or documents or both combined. A good database to use is a relational one as it is good at changing and adding data. As long as the amount of data is not too much. The larger the amount of data the more maintenance a database will need. Elasticsearch is able to handle agile data because the shards are independently being indexed/refreshed, and indices are being continuously refreshed with fixed time intervals. This indicates that the likelihood of a shard accumulating a lot of unrefreshed data is unlikely. A database in the RDBMS world is fixed and known before the first record is released. When there is an update, the schema has to change and must be extrapolated to the records already within the database. The process can be slow if the database stores a large amount of data. If a domain is used that has documents with too many additional fields, it can cause the database to have large sparse tables that end up wasting disk space for storing NULLs. Documents in indices are not subject to schema in elasticsearch. It automatically updates mapping when a new document is added to an index and if there are new fields within the document. It has the ability to automatically alter any data types of a field if a value needs a wider type.

Performance – To get the best practice for relational databases it is indicated that each one has to go through the normalisation process during the design stage. A dataset can be separated into a number of different tables by converting it into a more normal form. This

helps to mitigate the redundant information. Even though it being normalised can be beneficial to the create, update, and remove operation, it can cause problems for the read operations. SELECT statement most of the time hit more than one table and have to be joined before any filtering condition is applied. Even though RDMBS is able to handle these operations efficiently it can be time consuming if it requires complex schemas. With elasticsearch being document-orientated there is no need for it to spend time gathering the data. Furthermore, shards that are within an index are in search of documents satisfying several filter criteria concomitantly. The result of this is the data being combined and returned. Elasticsearch stands out from a lot of other systems because of the combination of scalability, agility, and performance, within the one system. (Kononenko, O, Baysal, O, Holmes, R, Godfrey, M. (2014)).

2.1.4 Weaknesses

Security – As good as some of the features are within elasticsearch, it lacks security features such as authentication and access control. The indices can be easily deleted if an adversary knows the URL of the server. This will allow them to also shut down the cluster. A proxy or firewall will be needed to protect the data in this case.

Learning Curve – Writing simple queries is easy with the JSON origin that elasticsearch uses, however it can become quite complicated when nested objects are involved. A nested query has to be used in elasticsearch when a filters condition is on a field from a nested object. This requires the knowledge of how a certain document is stored and analysed by elasticsearch. It has NoSQL system, lacks transactions, JOIN operation, possible inconsistencies within data, etc. (Kononenko, O, Baysal, O, Holmes, R, Godfrey, M. (2014)).

2.2 Overview of Logstash

Logstash is a data collection engine that is in real-time and has the ability to consume messages from multiple sources, such as HTTP, messaging queues and logging frameworks. Logstash normalises inputs that have their own inherent structure and will bring them in a consistent form. Once the message has arrived it will be transformed to a JSON-like event. They consist of key-value pairs. All of the incoming events can be enriched with additional data such as simple timestamps, and the content within these timestamps can be modified. Message will be dispatched to one or more destinations once it has been received and processed. Data storage such as 'MongoDB, Amazon S3, Hadoop or of course Elasticsearch', are supported for any possible targets.

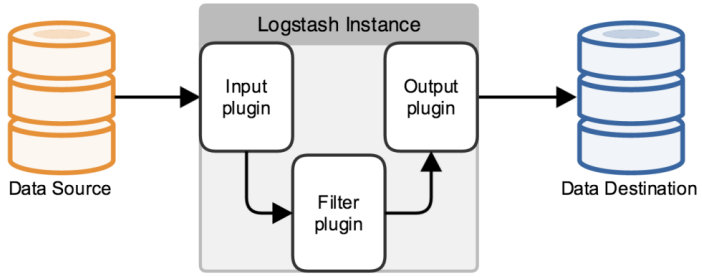


Figure 1. Logstash workflow (Kleindienst, P. (2016)).

Installing Logstash on Ubuntu is simple and only needs these three steps:

```
`$ wget -qO - https://packages.elastic.co/GPG- \`
`KEY-elasticsearch | sudo apt-key add -`
`$ echo 'deb http://packages.elastic.co/logstash/ \`
`2.2/debian stable main' | \`
`sudo tee /etc/apt/sources.list.d/logstash-list`
`$ sudo apt-get update &&`
`sudo apt-get install logstash`
```

The repositories public key will need to be downloaded, the repositories added to the sources and then for Logstash to be installed once the package list has been refreshed. The Logstash will automatically start once the installation is finished. (Kleindienst, P. (2016)).

2.2.1 Advantages of Logstash

It offers regex pattern sequences which identify and parse several fields in an input event.

There are a selection of web servers and data sources that are supported by the Logstash. These are used to extract logging data.

Several plugins are supported which are used to parse and transform all of the logging data into the users desired format.

With Logstash being centralised processing and collecting data from servers is much easier.

It supports network protocols, databases, and other services as a destination source, which is used for logging events.

It uses HTTP protocols which can be used to upgrade the Elasticsearch version without having to upgrade Logstash. (Tutorials Point. (2023)).

2.2.2 Disadvantages of Logstash

The process of logging the data can be affected in a negative way because of the use of HTTP.

It can be complex and will need a good understanding and analysis of the logging data that is inputted.

The correct sequences of patterns will need to be found by the user because the filter plugins are not generic. (Tutorials Point. (2023)).

2.3 Overview of Kibana

Kibana is a visualisation platform that was designed for Elasticsearch, which provides web-based interfaces that can search, view, and analyse any stored data within an Elasticsearch cluster. It is divided into four components - Discover, Visualize, Dashboards and Management.

All of the Kibana internals are configured within the *Management* section and is where index patterns need to be set. Timestamp fields will need to be set if the index patterns contain any time-based events. These fields will need to be based on which data will be sorted and filtered by Kibana. It is based on a set of indices will be used to fulfil selected index patterns, show the list of index fields, alongside the type and properties. The field formatter is able to be modified which will alter how the value of the field is shown within the Kibana GUI.

Pure data entities are allowed to be browsed and analysed interactively with the use of *Discovery*. This section categorises Elasticsearch documents that are based the index patterns. Apache Lucene query syntax is used to help search and filter by time or document properties easily. The amount of documents that match the search query can be viewed or the statistics of the field value.

Kibana has plugin-based architecture which makes it easily extended to suit any particular need. Data can be visualised with the visualisation plugin. Tables, charts, maps, histograms, and many other forms are way in which information can be shown. Large volumes of data can be shown easily with pie charts, bar charts, line, or scatter plots. A downfall is it cannot export raw data from Elasticsearch. Using ES2CSV tools can fix this problem. What this does is allow the export of Elasticsearch data in CSV form.

The dashboard section allows the combination of several saved Discoveries and Visualisations into one view. Any elements within the dashboard can be arranged and resized when needed. They can be saved, shared, or embedded into other webpages easily.

The main downfall of Kibana is the lack of user management. Using Nginx proxy is the most simple solution that is used to restrict selected user’s access. Commercial support can be purchased for elaborate access control, if it is required and will provide a license for X-Pack, which allows for the configuration of sophisticated security options as well as enable reporting and alerting features. There is a possibility that Searchguard can be used. This is a plugin that encrypts, authenticates, and authorises. All of the basic features are free as well as enterprise features for personal and non-commercial projects. For commercial use the enterprise features are licensed. ReadonlyREST is a plugin that has a similar set of features under the GPLv3 license. (Bajer, B. (2017)).

2.3.1 Advantages of Kibana

Kibana contains visualisation tools that are browser based. The main uses for this are to analyse a large number of logs in the form of line graph, bar graph, pie charts, heat maps etc.

It is easy to understand for beginners.

Reports can be easily converted from visualizations and dashboards.

It can analyse complex data easily with the help of canvas visualisation.

Data can be compared backwards to better understand the performance with the use of timeline visualisation. (Point, T. (2023)).

2.3.2 Disadvantages of Kibana

If the version is mismatched it can be tedious to add the plugins.

There are issues when trying to upgrade to a new version. (Point, T. (2023)).

2.4 Cyber Attack Detection Using Open-Source ELK Stack

The detection of cyber attacks is far from being one of the most active research areas in Information Security. There are several papers which discuss threat hunting implementations on security system. This paper discusses that in B. Raja, K. Ravindranath, and B. Jayanag paper they Packetbeat as being one of the solutions when it comes to monitoring and analysing anomaly activity within a network. With their paper they only show how to install Packetbeat and the ELK Stack.

The author of the next paper created a system that was used for the detection of malicious events using the ELK Stack and cyber threat intelligence. All of the data that is used will be sent from Elastic Beats to Logstash, and will then be forwarded to Elasticsearch. After this it will then be sent to threat intelligence for correlation with feeds. Then lastly it will be analysed to detect any malicious activity.

The author of paper 3 discusses the detection and analysis of malicious Windows events. The ELK Stack was used for monitoring logs, to identify malicious activity Sysmon was used and Winlogbeat was used to ship the Windows event logs to the ELK Stack. When a malicious hash is found it will be searched for on VirusTotal manually.

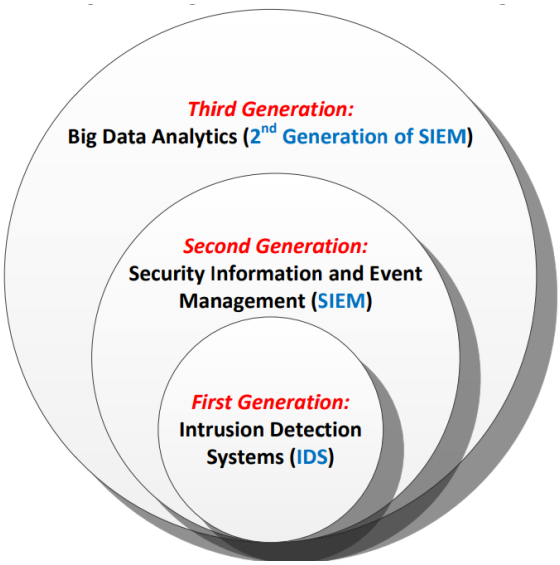
The fourth paper that is discussed talks about a solution for detecting and analysing Linux events that are malicious with the use of Filebeat and the ELK stack. Firewall logs will be collected through syslog-ng and then will be stored in /var/log/. The logs will then be forwarded to either Logstash or Elasticsearch with Filebeat. All of the events will be observed and analysed in Kibana. Moloch was installed on Elasticsearch and used to collect all of the packets on the network that were being transmitted on the server-side.

All of the authors including H. S. Kumar discuss how intrusion detection uses the ELK Stack. A Kali Linux machine was set up so they could perform some attacks (Port Scan -Nmap and Brute Force) against two stations. The logs that are generated will then be sent through rsyslog to a server with the ELK Stack running on it. In their paper they do not present any setup or config file. There is also no third-party solution integrated into the ELK Stack which would give the values to generated logs or provide additional information.

(Stoleriu, R, Puncioiu, A, Bica, I. (2021)) the authors of this paper compare their solution against the solutions from the papers discussed above. Their solution is focused on network and host monitoring (Linux and Windows) for cyber attack detection. Machine learning is used within the system and will find any anomalies within the network traffic (based on the Packetbeat logs). Tools such as Sysmon and auditd are used for fine-grained logging and then Elastic Beats will be used to help send the generated logs (Winlogbeat, Filebeat). The ELK Stack will be integrated with GeoIP processor and any different Threat Intelligence platform (VirusTotal, MISP). (Stoleriu, R, Puncioiu, A, Bica, I. (2021)).

2.4.1 How Data Analytics Relates to Cyber Security

Information security that is data related has been used for decades for bank fraud detection and anomaly-based detection systems. The vast amount of data generated within a fraud detection system that is generated means it can be considered as a Big Data analytics system. This is because they generate millions of data instances and events every day for medium or large organisations. The context of data analytics has evolved three generations according to intrusion detection as shown in the figure below.



(Rassam, M, Maarof, M, Zainal, A. (2017)).

The first generation of the intrusion detection system was used to identify any security breaches security mechanisms missed, such as risk assessments, malware detections, and other network perimeter security tools.

With the second generation which was the SIEM (Security Information and Event Management) it has an additional rule which is '*aggregating and filtering alarms from different sources'* (for example IDS sensors) of the first generation.

The second generation of SIEM is considered to be the third generation of Big Data security analytics. This generation will advance the effect of the second one and does this by the time consumed being reduced in security event information that is correlating and consolidating. It can be used in forensic analysis of cyber threats by using context and ling historical data.

There will need to be requirements for any Big Data analytics based cyber security solution that can cope with continuous and rapid growth of cyber threats that are sophisticated. The requirements that should be considered are the characteristics of Big Data as well as to the security demands. The requirements that should be considered are. (Rassam, M, Maarof, M, Zainal, A. (2017)).

1. *Dealing with Multi-Sources Data*

   The growth in data sources can be huge and can be considered by security systems based on Big Data analytics context. Some of the sources include '*firewall logs, active directory files, operating systems events logs, SIEM data, IDS data, SQL server logs, NetFlow data, threat intelligence data, and others'*. These data sources have existed for a long time, although they were not used in the context of Big Data analytics all together. It is required to get insights that are useful to detect and prevent threats so that multiple sources can be considered at the same time.

2. *Large Scale Data Management*

   The volume of data increases as the data sources grows. This in turn becomes an obstacle for cyber security systems that are Big Data analytics based. Therefore, this requirement should be considered when designing such a system, so that it can properly collect, process and retrieve data that is useful in a tamely and effective manner. The platforms that should use efficiently when designing any Big Data analytic system are cloud computing-based technologies such as cluster and grid platforms. It should be able to facilitate the storage, processing, retrieving data and drawing any useful insights about system events that may occur.

3. The number of data types that are able to be encountered can be increased by the rate of data generation, this can range from highly structured data to highly unstructured datasets. Most of the data that is used in security analytics is numerical and from other data types. Although these days very unstructured data can be collected from sources such as '*e-mails, blogs, social networks activities, threat feeds, and combinations of these and others sources'*. This means that if these data types are to be considered together, the design of the Big Data analytics system and tools needs to be designed properly.
4. Within cyber threat intelligence Visualisation is a key factor, and it provides an analysis of security data that is graphically descriptive. The way it reveals data anomalies and instructions is with the visualised connection between devices, events, locations, signatures, and IPs. This makes visualisation critical when trying to understand the connection and to extract insights about how the network system is behaving. A network visualisation tool that was created to visualise cyber threats in order for the user to perform better and more effective data analysis is called 'Keylines'. It is used to extract hints that are important from complex connected data. There are four main capabilities within this tool which are: (i) the analysis of software threats and vulnerabilities; (ii) the detection of anomalous logins; (iii) the identification of patterns in data breaches; and (iv) the detection of malware propagation patters over a period of time. A good thing about these applications is that users can be involved when discovering patterns and anomalies. It does this by turning the raw data into interactive charts. It was pointed out that visualisation tools are used to help security teams understand the relation and to track the historical patterns that are amongst security data elements. Another visual analytics system that has been designed is called Visual Analytics Suite. The way it was designed was

by combining multi-criteria clustering techniques. There are three types of interactive visualisation that is used, which are treemap, node-link and chord diagrams. Just like Keylines VACS also aims to gain insight from a variety of threat landscapes. VACS's design is a dashboard interface for querying and receiving information to investigate suspicious hosts.

5. Cloud computing, distributed computing, grid computing, stream processing, Big Data modelling, Big Data structure, and software systems are some of the main infrastructure technologies that support Big Data and these needs to be studied thoroughly. With any Big Data application, including cyber security it has been made clear that the data evidence of attacks and security breaches that has been utilised will be growing across the three V dimensions, which are volume, velocity and variety. The result of this growth is it hardens the detection of attacks using traditional technology. If the traditional information retrieval system is only used, it will make it more difficult to detect sophisticated APT attacks when being used to design traditional IDS. Advanced technologies such as MapReduce structure should be used instead. There is a higher chance of the detection system of APT's being able to handle sophisticated APT unstructured data with different formats and that has been collected from different sources. when using MapReduce implementation. Examples of these sources are system logs, IDS, NetFlow, Firewalls, and DNS systems during a long period of time. MapReduce out performs SQL based detection systems due to its massive parallel processing power. The design of MapReduce as map and reduce function means it is easier for users to incorporate more detection algorithms. The distribution is transparent to users (who directly work with specific data) because of the design. When large-scale distribution systems are used the analysis of large amounts of data will be simplified and will provide a mechanism that will be used to reveal more paths to attacks and any targets that are used to detect difficult threats that are unknown. (Rassam, M, Maarof, M, Zainal, A. (2017)).

2.5 Anomaly Detection

Anomaly detection system is used to develop a model of behaviour that is normal and then it will define any activity that strays from predilections that have been generated as anomalous by the model. This paper discusses the authors of another paper explain three types of anomalies they worked with. Foreign-symbol anomalies, foreign n-gram anomalies and rare n-gram anomalies. When a character or item are encountered by an IDS this means that it is the first time a foreign symbol anomaly occurs. When a previous unseen sequence of characters appears, this is known as foreign n-gram anomaly. When character sequences appears more than once but below a specific user threshold, for example 5% at a time, this is when rare n-gram anomaly occurs.

Changes of behaviour within a network is what network-based anomaly detector looks for. Even if a specific sequence of n-gram or packet may not be rare, but if there is a sudden or unexpected increate to the number or rate of occurrences, this could be considered to be anomalous. Nine anomaly behaviours have been identified by Lakhina et al in-connection information from network traffic. This was detected by using an entropy-based approach. If there is a behaviour change that strays to far from the baseline activity will be classed as anomalous. (Gates, C, Taylor, C. (2006)).

2.5.1 Anomaly Predicting Model

A lot of different machine learning algorithms are used in anomaly prediction, logistic regression, PCA-Q statistics, support vectors machine and ensemble learning methods are examples of some of the algorithm.

The basis of logistic regression is the assumption that the sample will obey Bernoulli distribution and will be solved with the maximum likelihood estimation and gradient descent. The fields it is the most used in are classification, prediction and evaluation. Defining the hypothesis function, '*construct the loss function, minimize the parameters of the objective function through the loss function, predict the test data through the solved objective function and evaluate the predicted data*' is the basic process. The methods and formula involved are detailed in the following way:

Hypothetical functions can be regarded as the result of the linear regression equations. The following formula represents linear regression:

$$\theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n = \sum_{i=1}^{n} \theta_i x_i = \theta^T x$$

The sigmoid function is:

$$g(z) = \frac{1}{1 + e^{-z}}$$

$h_\theta(x)$ When formula one is substituted to formula two the following formula is obtained:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

signifies that the probability of a sample is predicted as positive with the probability that the same will be predicted as positive and a negative is easily obtained as follows:

$$P(y = 1|x; \theta) = h_\theta(x),$$
$$P(y = 0|x; \theta) = 1 - h_\theta(x)$$

This formula can be combined into the following formula:

$$P(y|x; \theta) = (h_\theta(x))^y (1 - h_\theta(\xi))^{1-y}$$

When combined the formula is as follows:

$$P(y|x; \theta) = (h_\theta(x))^y (1 - h_\theta(\xi))^{1-y}$$

(Wang, B, Hua, Q, Zhang, H, Tan, X, Nan, Y, Chen, R, Shu, X. (2022)).

2.5.1.1 Constructing Loss Function

If the likelihood function is taken from the probability representation of the result that is predicted, will obtain the following function:

$$L(\theta) = \prod_{i=1}^{m} P\left(y^{(i)} | x^{(i)}; \theta\right)$$

$$= \prod_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right)\right)^{y^{(i)}} \left(1 - h_\theta\left(x^{(i)}\right)\right)^{1-y^{(i)}}$$

Logarithmic operation is used to obtain the log likeliness function:

$$l(\theta) = \log L(\theta)$$

$$= \sum_{i=1}^{m} \left(y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right)$$

Assume that:

$$J(\theta) = -\frac{1}{m} l(\theta)$$

Final form of loss function:

$$J(\theta) = -\frac{1}{m} \times \sum_{i=1}^{n} \left(y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right) \quad (9)$$

To obtain the parameters of the objective function is to minimise the loss.

The gradient of the loss function is:

$$\frac{\delta}{\delta_{\theta_j}} J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\frac{1}{h_\theta\left(x^{(i)}\right)}\frac{\delta}{\delta_{\theta_j}}h_\theta\left(x^{(i)}\right) - \left(1-y^{(i)}\right)\frac{1}{1-h_\theta\left(x^{(i)}\right)}\frac{\delta}{\delta_{\theta_j}}h_\theta\left(x^{(i)}\right)\right)$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\frac{1}{g\left(\theta^T x^{(i)}\right)} - \left(1-y^{(i)}\right)\frac{1}{1-g\left(\theta^T x^{(i)}\right)}\right)\frac{\delta}{\delta_{\theta_j}}g\left(\theta^T x^{(i)}\right)$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - g\left(\theta^T x^{(i)}\right)\right)x_j^{(i)}$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j^{(i)}$$

θ Is the final update, α is the learning rate.

$$\theta_j := \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j^{(i)}.$$

The logistic regression model once it has been determined will use the maximum likelihood estimation method to realise the minimum divergence. This is done to calculate the model's parameters.

PCA which stands for Principal Component Analysis is an analysis technology with its main aim to transform several indications into several comprehensive indicators. It does this by using the idea od dimension reduction. The dimensions of a dataset will be reduced whilst the feature that is contributed to the dataset is being maintained. Wang, B, Hua, Q, Zhang, H, Tan, X, Nan, Y, Chen, R, Shu, X. (2022)

The following is the algorithm flow:

Input: n-dimensional sample set $D = \left(x^{(1)}, x^{(2)}, \cdots, x^{(m)}\right)$, reduced to dimension n' (n' $<$ n)
Output: sample set after dimension reduction D'

Step 1. Centralize all samples: $x^{(i)} = x^{(i)} - \frac{1}{m}\sum_{j=1}^{m}x^{(j)}$;
Step 2. Calculate the covariance matrix of the sample: X^T;
Step 3. Eigenvalue decomposition of the matrix XX^T;
Step 4. Extract the eigenvector $(w_1, w_2, \cdots, w_{n'})$ corresponding to the maximum n' eigenvalues;
Step 5. Convert each sample $x^{(i)}$ in the sample set into a new sample $z^{(i)} = W^T x^{(i)}$;
Step 6. Get the output sample set: $D' = \left(z^{(1)}, z^{(2)}, \cdots, z^{(m)}\right)$.

Wang, B, Hua, Q, Zhang, H, Tan, X, Nan, Y, Chen, R, Shu, X. (2022)

After the value of n' has been dimensionally reduced it is not specified, instead the dimensions will be reduced to the principle component threshold t, which is situated

between (0,1). If the eigenvalues are $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_{n-1} \geq \lambda_n$, then n' is obtained with the following formula:

$$\frac{\sum_{i=1}^{n'} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq t$$

The reduction of the dimensions of log data is the only thing that the PCA method can do. When it comes to data that uses feature extraction and dimensionality reduction, the established statistical hypothesis test method will be used to judge if log data strays from the principal component model. The deviation procedure of the predicted value is described by the Q static (i.e., Square Prediction Error statistic). This is from the principal component model at specific times.

The Square Prediction Error (SPE) statistics within $\widetilde{S}$ space is defined with:

$$SPE = \| \widetilde{C}_x \|^2 \leqslant Q_\alpha^2.$$

The changes within the data that cannot be explained by the principal component model is represented by SPE or Q statistics. Any abnormal situation within the process will be caused by the SPE being too large. The data that is established by the model will not be applicable when the process runs normally. Assumptions is what the calculations of the control limit are based on. When is the test level is α, the SPE control limit can then be calculated as shown in the following formula:

$$Q_\alpha = \theta_1 \left[\frac{C_\alpha \cdot \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{\frac{1}{h_0}}$$

where:

$$\theta_i = \sum_{j=k+1}^{n} \lambda_j^i (i = 1, 2, 3) \qquad h_0 = i - \frac{2\theta_1 \theta_3}{3\theta_2^2}$$

In the above equations λ_i is the eigenvalue of the covariance matrix of X, is the vital value for the normal distribution at the test level α, k the amount of principal components that are retained within the principle component model is represented by with n being the amount of principal components.

The linear classifier that has the largest interval that is defined within the feature space is a Support Vector Machine (SVM). Kernel techniques are also included, this makes it a classifier that is nonlinear.

The strategy which SMV uses to learn from is interval maximization. This can be formalised as a problem of solving convex quadratic programming. Optimization algorithm that is used to solve convex quadratic programming is its learning algorithm. The basic idea of this is to solve the separation hyperplane. This can be used to split the training dataset correctly and it has the largest geometric interval.

Suppose a training dataset on the feature space is given:

T = {(x_1, y_1), (x_2, y_2), … , (x_N, y_N)}, where $x_i \in \mathbb{R}^n$, $y_i \in \{+1, -1\}, i = 1, 2, \cdots$ N and x_i is the i-th feature vector, y_i is the class mark. It is positive when it is equal to +1, and negative when it is equal to -1. The assumption can be made that the training set is linearly separable.

Geometric interval: the dataset T and hyperplane w -x +b = 0, the geometric spacing interval of hyperplane about the sample point (x_i, y_i) is defined as the following:

$$\gamma_i = y_i \left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \cdot \boldsymbol{x_i} + \frac{b}{\|\boldsymbol{w}\|} \right)$$

The following shows the geometric intervals minimum value within the hyperplane about all of the sample points:

$$\gamma = \min_{i=1,2\cdots,N} \gamma_i$$

The above equation shows the maximum separation hyperplane that is solved by the SVM model can be expressed with the constrained optimisation problem shown below:

$$\max_{w,b} \gamma \quad s.t. y_i \left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \cdot \boldsymbol{x}_i + \frac{b}{\|\boldsymbol{w}\|} \right) \geq \gamma, i = 1, 2, \cdots, N$$

γ is used to divide both sides of the constraint and to obtain the following formula:

$$y_i \left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|\gamma} \cdot \boldsymbol{x}_i + \frac{b}{\|\boldsymbol{w}\|\gamma} \right) \geq 1$$

With ||w|| and being scalar, for the sake of simplicity we will let

$$\boldsymbol{w} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|\gamma}$$
$$b = \frac{b}{\|\boldsymbol{w}\|\gamma}$$

Then get the formula:

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, i = 1, 2, \cdots, N$$

Maximization γ is equivalent to maximization $\frac{1}{\|w\|}$, and the minimization it is equivalent to is $\frac{1}{2}\|w\|^2$, which means the following will be used to represent the maximin separation hyperplane of the SVM model:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad s.t. y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, i = 1, 2, \cdots, N$$

It is a convex quadratic programming problem that consist of inequality constraints. The Lagrange multiplier method is used to solve its dual problems.

The goal of the supervised learning algorithm within machine learning is to learn a model that is stable with good performance in every aspect. However, the results it actually generates are mostly not ideal ones and at times only multiple preferred models can be obtained. Ensemble learning's fundamental idea is that even a weak classification that gets a wrong prediction can be corrected by other weak classifications, as to reduce the variance, deviation or the improvement of predictions.

Below there will be an explanation of all three of the ensemble learning.

Firstly, sequences ensemble method is boosting and the very basic learning participating are generated in series. The classifiers that are at the base are stacked layer by layer. During training, higher weight are given to the sample that are wrongly divided by the previous base classifier, given by each layer. The results will be weighed depending on the results of each layer classifier. The dependencies are used to assign higher weights to any samples that are incorrectly mark in the pervious training. It does this between basic learners and the overall effect of the prediction can be improved.

Next is the parallel ensemble which is bagging. There are no strong dependencies between the basic learners and the learners who train on parallel within the training process. The training set can be dividend into numerous subnets which can be used to make the basic classifiers independent of each other. Separate judgements are made by the each individual

and the voting is what makes the final collective decision. The main principle of this method is to make use of the independence that is between the learners, which is good for reducing errors by averaging.

The last method is the model fusion ensemble. What this is, is stacking which means it uses the training model to combine other models. First of all, numerous models are trained, then the output of these models will be used as input to train the model to get the final output. In theory stacking can represent two ensemble learning methods as long as the combination of models is adapted. Fig. 1 shows that a series of classifications models can be obtained by obtaining training sets by using bootstrap sampling. The output of this will be used to train a second layer classifier.

The characteristics of ensemble learning methods are: several classification methods are gathered to improve the accuracy of classification (these can either be the same or different algorithms); Base classifiers are constructed by ensemble learning methods and does this from training data. How the classification is performed is by voting on the predictions of the basic classifiers. Integrated classifier performs better than a single classifier.

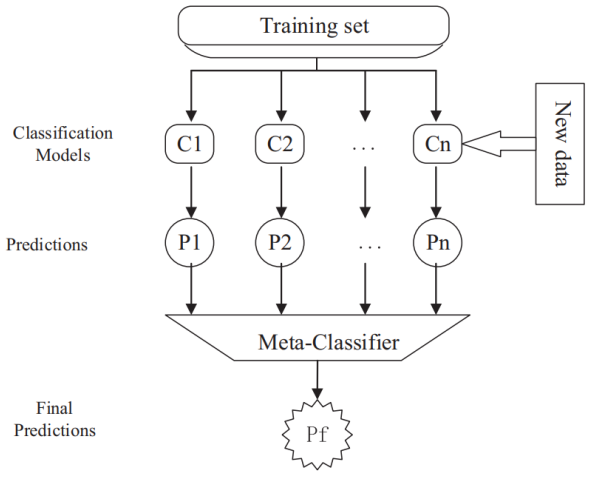


Fig. 1 Stacking - hybrid model of ensemble learning method. (Wang, B, Hua, Q, Zhang, H, Tan, X, Nan, Y, Chen, R, Shu, X. (2022)).

2.6 Data Classifications

A problem that is in many decision-making tasks is data classification. A lot of these tasks are instances of classification problems or they can be articulated into a classification problem. Examples of these are predictions or forecasting tasks, diagnosing tasks and pattern recognition. A type of data classification technique is discriminant analysis. What this does is it determines groups of units based on the observed score. What this technique does is it classifies observations or units of groups or class membership that are unknown, with the use of values of a set of variables that are associated with each observation or unit.

The main methodologies associated with classification are statistical methods, mathematical programming approaches, neural networks and other approaches that are machine learning based. The statistical classification model can be traced back all the way to Fisher's linear discriminant analysis. This analysis technique maximises the ratio between groups and within group variances. It is an optimal method for any situation where the fundamental populations are multivariate normal and all of the groups different groups have equal covariance structures. Quadratic discriminant analysis when multivariate populations have unequal covariance structures is recommended by (Fisher, 1936; Smith, 1947). When statistical classification models are not able to provide good or satisfactory classification results, this happens when normality assumptions are not provided. Techniques of mathematical programming are proposed to a variety of data classification problems. A paper that was written by Fred and Glover (1981a, 1981b) had activated an excess of research contributions on a mathematical programming approach. Mathematical programming discriminant analysis at its most simple form generates a discriminant function that is used to separate training samples into specific groups according to known memberships. There are several mathematical programming techniques that are used for optimisation, which includes minimisation of the sum of deviations, maximisation of minimum deviations, goal programming, mixed-integer programming and hyper-box representation. Artificial neural network is popular for solving business and technical problems that uses predictions and has a wide range usage are in the classification problems. An important issue when it comes to neural networks is training of the networks.

What is most used search technique when training neural networks is backpropagation. algorithm. This algorithm has negative features which are being captured in local solutions and in some cases having low classification performance. Alternatives are provided to help prevent these disadvantages. An alternative algorithm for training neural networks is genetic algorithms, binary-coded genetic algorithms especially have been compared with backpropagation within data classification, regression and forecasting. Studies show that genetic algorithms outperform with training than backpropagation.. (Örkcü, H, Bal, H. (2011)).

2.6.1 Structured Data

Data that is structured includes mostly text, which can be easily processed. The data can be easily entered, stored and analysed. The form they are stored in are rows and columns, which can be managed easily with structured query language. A model that supports structured data is relational model. What it does is support structured data and manages the for of rows and tables then processes the content within the table. Another thing that supports structured data is XML. XML is the form that most content within web pages is in. All of the content of is included within structured data, and companies like Google use the structured data to understand the content of the page. What this means is that most of the searches made on Google are done with the help of structured data. Since the databases have started network, hierarchical, relational, object relational data model deal with structured data.

2.6.1.1 Characteristics of Structured Data

1. There are several data types within structured data such as date, name, number, character and address.
2. They are arranged in a defined way.
3. SQL is used to handle structured data.
4. It relies on schema and is schema based.

The data can interact with computers easily. (Praveen, S, Chandra, U. (2017)).

2.6.1.2 How to Manage Structured Data

The wat that structured data is managed it by using relational databases for example Excel sheets or structures query language (SQL) database. What this kind of database is based on is the relational model, which represents data that is in a tabular form. This helps

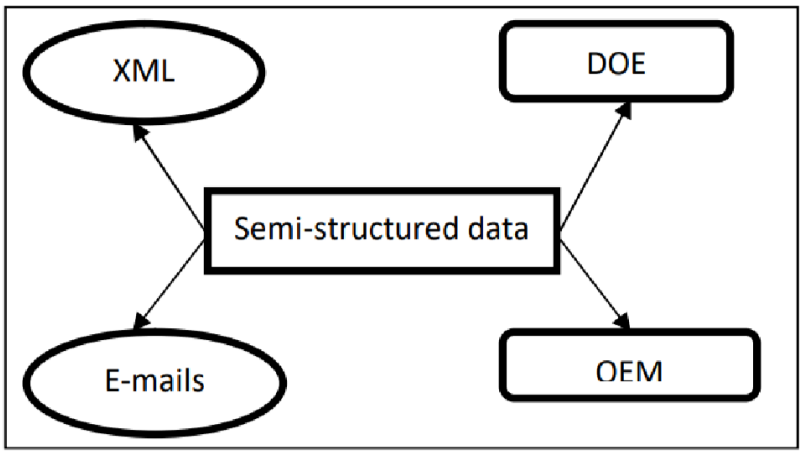
organisations to create relationships between a variety of data points and for structured data to be inputted, searched and manipulated.

Structured data needs to be structured into a data model before it can be placed into a database. This is because it is schema-on-write. Defining a schema based on the data is how the model is established and will then create tables or entities. The relationship between entities will then need to be established. Lastly the SQL script will need to be written in order to produce the relational database that will be used to store the structured data. Once this is done it can then be accessed and manipulated to suit the users needs.

Once the relationship between the data points has been established the SQL script needs to be written up. Sources that structured data can come from are online forms, network logs, sensor data, and points-of-sale. The data can then be used to drive machine learning when it is used in algorithms. What this does is search and analyse data and then generate reports. (Elastic. (2023)).

2.6.2 Semi-Structured Data

What is included in semi-structured data are emails, XML and JSON. When semi-structured data is expressed with edges, labels and tree structures, it will not be fit for a rational database. Trees and graphs help to represent these structures and have attribute labels and are also schema-less data. Graph based data models can be used to store semi-structured data. A model that supports JSON is MongoDB which is NOSQL. Semi-structured data contain self-describing tags. Some of the most known data models that use semi-structured data are Objects Object Model, Data Exchange Model and Data Guides. Some of the concepts of semi-structured data models are document instance, document schema, elements attributes, elements relationship sets. (Praveen, S, Chandra, U. (2017)).



Attributes of Semi-Structured Data (Praveen, S, Chandra, U. (2017)).

Examples of Semi- Structured Data

```
{
Row:{Emp_id:” 12345”,Emp_name:”Ram”},
Row:{Emp_id:” 56786”,Emp_name:”Hari”},
Row:{Emp_id:” 67858”,Emp_name:”Shyam”},
Row:{Emp_id:” 90890”,Emp_name:”John”},
}
```

2.6.2.1 Characteristics of Semi-Structured Data

1. This data is based on Schema.
2. Label and edges represent the data.
3. A variety of webpages are used to generate the data.
4. There are multiple attributes within the data. (Praveen, S, Chandra, U. (2017)).

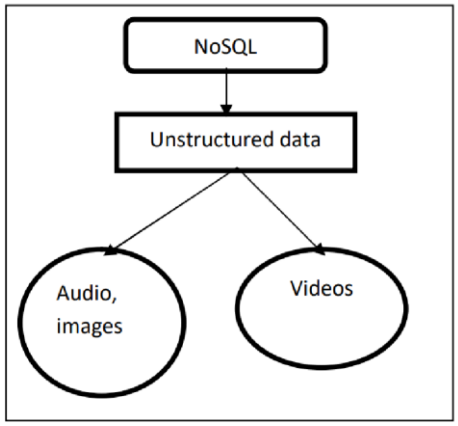
2.6.3 Unstructured Data

Videos, images and audio are included in unstructured data. Any data that increases is known as unstructured data. It is not suitable for rational databases and in order for them to store any data NoSQL databases need to be used. There are four families of NoSQL database: key-value, column oriented, graph-oriented and document-oriented. A lot of popular organisations such as Amazon, LinkedIn, Facebook, Google and YouTube deal with NoSQL data so replaced their convention database to one that is NoSQL.

2.6.3.1 Characteristics of Unstructured Data

1. This type of data is not schema based.
2. Not suitable for relational databases.
3. There is about 90% of data growing today.
4. Digital media files, Word doc, pdf files are included.

NoSQL databases store the data. (Praveen, S, Chandra, U. (2017)).



Attributes of Unstructured Data (Praveen, S, Chandra, U. (2017)).

2.6.3.2 How to Manage and Analyse Unstructured Data

When it comes to unstructured data there is no structure that can enable easy management and analysis. A structure needs to be defined which will help to manage the data and then it can be analysed more easily. Storing, organising and securing the data will then be allowed.

Once this is done the unstructured data will then be ready to process and analyse. Which provide organisations with insights that are actionable.

There are several tools and techniques that can help to manage and analyse unstructured data.

Natural Language Processing (NLP): What the NLP technology foes is it focuses on the interactions between computers and humans by using natural language. Its main goals it to read, decipher, understand and make sense of human linage in a way that is valuable.

Machine Learning (ML): Machine learning is a part of artificial intelligence and is used to enable computers to learn and make decisions that are data based, it improves the performance over a period of time without the need to explicitly programme it. The way that it identifies patterns within structured and unstructured data is by using statistical techniques, which then will make predictions or decisions.

Data Lakes: The variety and volume of unstructured data means it can be stored in data lakes or where the data was originally created (at the edge). Data lakes are usually used for large volumes and a variety of types of data. They accommodate data in a native format. This means video, audio, text and documents can be stored in the same place. (Elastic. (2023)).

2.7 ELK Stack vs Splunk

Set up and Maintenance

Splunk is easier to configure and set up because it is a proprietary software, unlike the ELK Stack. Both ELK and Splunk can sit on a user's physical data centre because they both support on-premises and SaaS deployment. The users will then be able to deploy them to the cloud.

Storage

Splunk uses indexes that are made up of file buckets to store its data. The structure of these buckets enables Splunk to see if the data incudes terms or words. They contain data that is compressed and raw. Once the data has been compressed it will be reduced to 15% of its original size. This helps Splunk to store more efficiently.

Data that is stored within Elasticsearch is stored as an unstructured JSON document. The documents have a set of keys (names of fields or properties) with corresponding values (strings, numbers, Booleans, dates, arrays of values, geolocations, or other types of data). All of the content of the stored documents are indexed. Documents are then fully searchable at the same time as needing more storage space.

Query Language

The syntax that is on Kibana is query and it is based on Lucene Query syntax, unlike Splunk which uses its own Search Processing Language. It is easy to learn Lucene query syntax because it is very similar to scripted languages. SPL is a proprietary language that will support the search pipelines.

The main difference between the Lucene queries and SPL syntax is SPL supports search pipelines which will chain together commands by using a pipe character. What this does is it

allows the output of a command to be used as the input for the next one. Whereas Lucene query syntax is a lot more simple and the outputs form the query will be directly generated.

Indexing

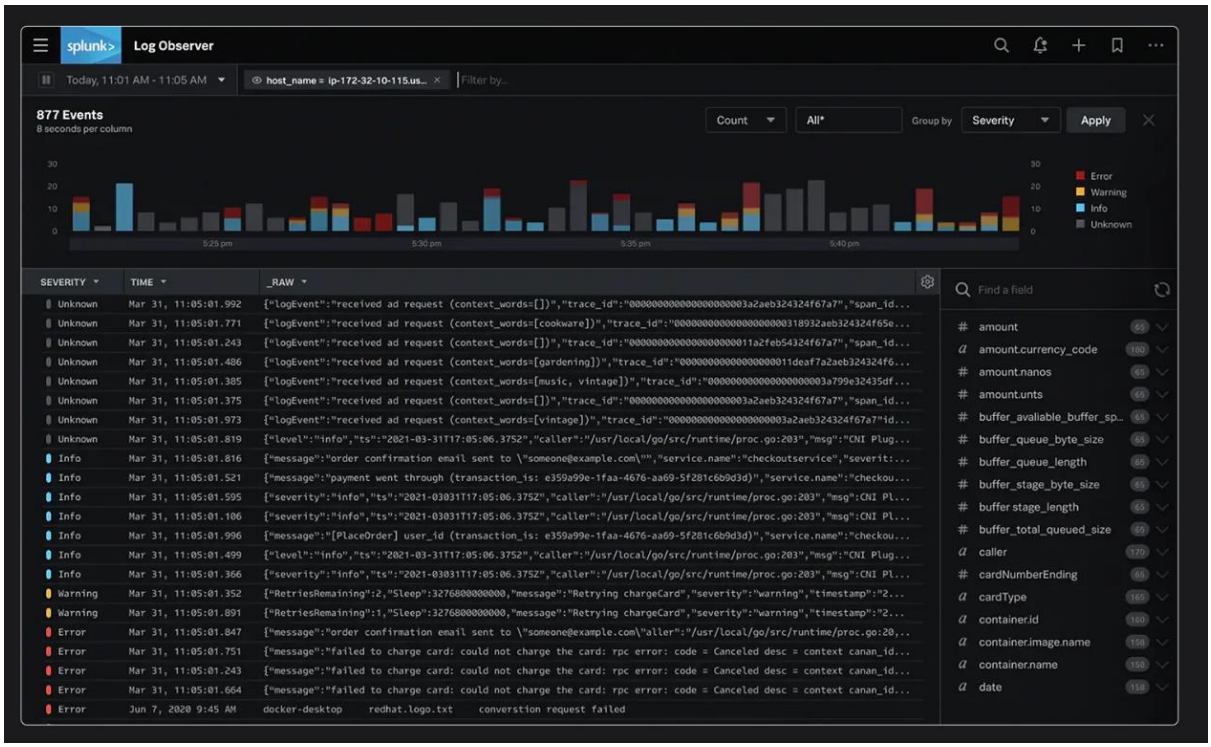
Elasticsearch indexing is a collection of documents that are correlated. The data structure that is uses is known as inverted index and has been created to provided full-text searches. Lists of unique words that appear in documents will be generated from the data structures and the documents the words occur in will also be identified. An Index API is used for the indexing, which means a user can add or update a JSON document within a specific index.

Indexers are used to index data within Splunk that is coming from the Splunk forwarder. The logs of data that is generated will be broken into lines by the indexer and then timestamps will be identified which will then be used to create individual events. Metadata will be annotated into it. Data can only be parsed if it from a universal forwarder or else it directly indexes the data. Transformation rules which are defined by an operator will be used to transform event data. The events will then be written onto a disk by Splunk, then point to them from an index file which can enable fast search across a large volume of data.

Data replication is another benefit of Splunk indexer. It keeps several copies of the indexed data, which means there is no need to worry about losing data. (Paliwal, M. (2022)).

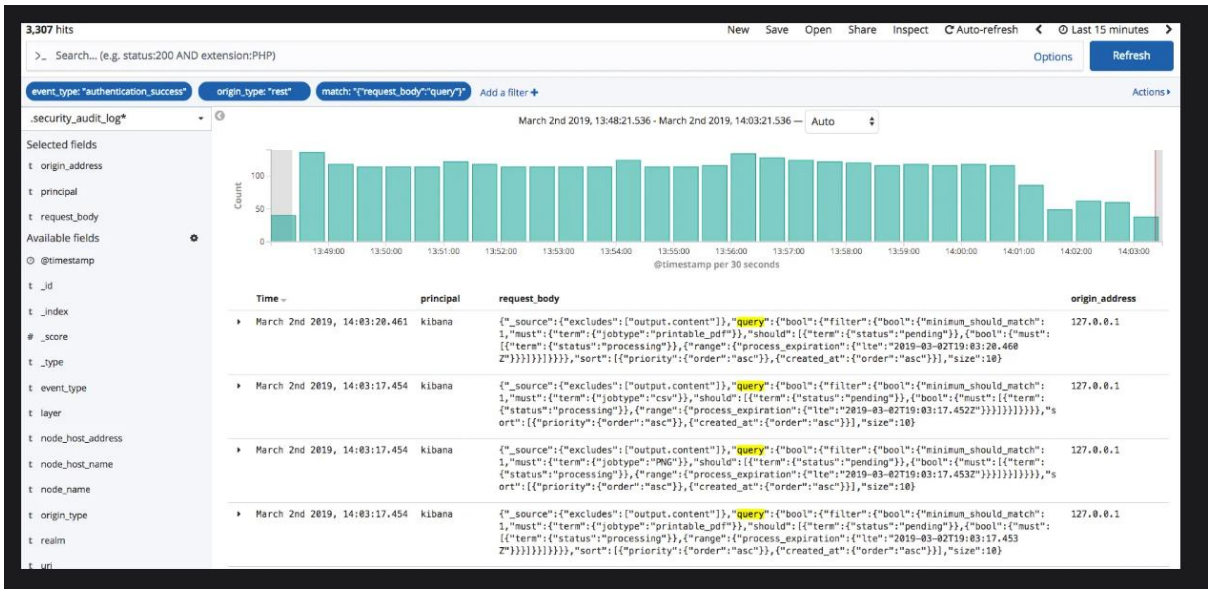
User Interface

When it comes to Splunk's user interface user management and control is provided by Splunk's web UI or the search head. It is also provided within xpack Kibana.



Splunk Log Observer (Source: Splunk website (Paliwal, M. (2022)).

Kibana on the other hand has a feature that allows a user to build a dashboard quickly. The users will have to make sure of the data types. For the aggregator function to work the data types have to be correct. The filtering of data within the ELK tack is much easier and more advanced.



Kibana Dashboard (Paliwal, M. (2022)).

Data Collection

Within Splunk's official document it states that the collection of data depends on the type of data source that a client is using. Below are the methods that Splunk uses for data collection.

- JSON objects that are gathered from events and metrics endpoints of the ingest REST API are collected with ingest service.
- Data that is collected from Splunk forwarder uses forwarder services.
- DSP HTTP event collector is used collect HTTP clients data and data sources from Syslog's.
- DSP Collect connectors are used to collect data from a number of types of data sources for example, Amazon S3, Amazon CloudWatch, Azure, etc. The way these collectors collect data is by running jobs on a schedule.
- When using steaming connectors, data will be collected from a number of sources such as Apache Kafka, Apache Pulsar, Google Cloud Pub/Sub, etc. The data will be received from these kinds of connectors and will be continuously emitted by the sources.

One of the ways to collect data an ingest within elasticsearch is by using REST calls. Elastic Beats is also used. What it does locally sit on a clients device and collect all of the logs which will then be sent to the aggregator (Logstash).

Pricing

The free version of the ELK stack means a license will not need to be paid for, although a lot of support and maintenance will be needed which can be costly. Splunk is a very pricy tool that is used by a lot of enterprises. (Paliwal, M. (2022)).

2.8 Justification of Research Papers Used

The research papers that were used in this thesis were chosen over other research papers because they relate more to this project. The first few papers used are explaining what the ELK Stack and Elasticsearch are and how it is used. A review of a research paper on malware attack detection using open-source ELK Stack has been discussed as it is related to this project the most. Then the next ones explain what anomalies are as machine learning will be used to monitor the behaviours and patterns of the data. What data classification is, how it relates to cyber security and how the elasticsearch is used to analyse and predict the patterns of structured, unstructured and semi-structured data. I have used these papers because they are written by reputable authors who are respected in their field.

3.0 Research Design

The research methodologies that will be conducting within this project are Data Analysis, both quantitative and qualitative (as both numerical and non-numerical data will be analysed), and Network Development Model. Network development methodology will be used when creating the network and the data analysis will be used when analysing patterns and data within elasticsearch.

3.1 Quantitative Data Analysis

This type of analysis is used to analyse data that is either already numerical or data that can be converted into numbers. Objects are described or interpreted statistically, by using numbers to construe data collection. To gain insight into this data algorithms, mathematical analysis tools and software will be used. 'How many, how often, and how much?' are questions that will be answered during the analysis of the data. The data is acquired surveys, questionnaires, polls, etc.

It uses both computational and statistical methods, with their main focus being statistical, mathematical, and numerical datasets. Descriptive statistical is the first phase, which is then, if needed followed by a more thorough analysis, in order to gain more insight, for example 'correlation, and the production of classifications based on the descriptive statistical analysis.

The two main data analysis are descriptive statistics which is used to explain phenomena and inferential statistic which is used to predict. (Eteng, O. (2022)).

3.1.1 Data Preparation Steps for Quantitative Data Analysis

The data used will need to be collected and cleaned before any analysis is performed. It is important to do so because this could cause problems such as wrong findings, wrong judgements, and misinterpretation.

The way to prepare the data is to merely convert it to a meaningful and readable format. The steps for this will be shown below.

Data Validation: this first step is to examine the data through the obligatory channels in order to see if it was collected correctly and if it meets the standards that were set out. The way this can be determined is by checking if all of the set procedures were followed, ensuring that the respondents were chosen solely on research criteria, then to check if the data is complete.

Data Editing: with large datasets there may be errors that occur. This can happen if a field has been filled out incorrectly or if someone accidentally forgot to fill it in. Data checks need to be performed to ensure that there is no unwanted data that may lead to inaccurate results.

Data Coding: this final step groups and assigns values to the data. This can be done with formatted tables and structure, which can help to represent the data more accurately. (Eteng, O. (2022)).

3.1.2 Quantitative Data Analysis Method

The method that will be implemented into this project is descriptive statistical. This method is used to describe a dataset and help to understand and summarise it by finding patterns from each sample of data. Absolute numbers that have been obtained from a sample will be provided, although this will not explain the foundation behind the numbers and are only used for the analysis of a single variable.

The methods used include:

Mean: The numerical average of sets of values will be calculated.

Median: The midpoint of a set of values is obtained using the median and the numbers will be arranged in a numerical order.

Mode: The values that appears the most in a dataset will be found.

Percentage: A group of respondents will be compared in relation to a larger group of respondents.

Frequency: The number of times a value is found will be shown.

Range: The highest and lowest values within the set will be shown.

Standard Deviation: This is used to show how close to the mean the numbers are.

Skewness: How symmetrical a set of numbers are will be shown, and will show if tis clustered into a smooth bell curve shape that will be in the middle of the graph or skewed towards the left or the right. (Eteng, O. (2022)).

3.2 Qualitative Data Analysis

This type of analysis is used to collect and analyse non-numerical data, such as (videos, audio, and text). The data will then be used to gain an understanding of concepts, opinion, or experiences. Any problems that have occurred can be examined in more detail. There are 4 main steps involved in this analysis type:

The first step is to familiarise yourself with the data. The data needs to be read and re-read several times to determine which piece is more valuable.

The next step would be to categorise and sub categorise all of the data. The categories are examples of thematic ideas. The data and the thematic idea will then be grouped together and be examined together. The first set of code will consist of a list of build themes.

The third step is to search for patterns and connections so that it is easier to see which data is important and to be able to identify any relationships between the dataset and the themes.

The last stage is concluding. (Valcheva, S. (2023)).

3.2.1 Qualitative Data Analysis Methods and Techniques

The type of qualitative analysis method that will be conducted in this project will be content analysis.

This kind of analysis is the most widely used and its main function is to understand the meaning of text data and how important it is. Processes and procedures that include categorisation of text data are also included. This is mainly for the classification and summarisation. The format of this text can be in documents, pictures, video, audio, etc.

The need for content analysis software programs in this era are vital. These programs are good for examining unstructured text data such as ‘documents, emails, social media, chats, comments, news, blogs, competitor websites, marketing surveys questions, customer feedbacks, product reviews, call centre transcripts and even scientific documents. (Valcheva, S. (2023)).

3.3 Network Development Method

The Network Development Life Cycle is model behind the network design process. The significant word that describes the term of the network development life cycle is the word 'cycle'. It demonstrates the incessant nature of the network development. When the network is designed from scratch it will need to start at the analysis stage. Whereas networks that already exist are continuously progressing from one stage to the next. When monitoring these networks, the statistics that would be produced would be management and performance, using the protocol Simple Network Management Protocol (SNMP).

There are two natures that the network design can, either physical or logical. Physical network designs are used to arrange and interconnect physical network circuits and devices, whereas logical configures and defines any service that runs on the physical network, for example, addressing schemes, routing schemes, traffic prioritization, security and management. Network design changes will be simulated with the use of a sophisticated network simulation software package or by using a prototype within a test environment. (Goldman, J & Rawles, P. (2001)).

3.3.1 Network Development Life Cycle Steps

Prepare – During this stage the requirements that are needed will be established, a network strategy will be developed, and suggesting a high-level conceptual architecture and then to search for technologies that will support the architecture the best.

Plan Phase – The network requirements will be identified during this stage, on the basis of the goals for the network, in the location of the network installation and who decided on the network services. Assigning a place where the network will be installed is also part of this phase. A gap analysis will be performed to see if the system infrastructure is able to be support by the existing infrastructure, site and operational environment. A project plan can be created to help manage all of the tasks, responsibilities, critical milestones and resources that are needed to implement those changes within the network.

Design Phase – This stage drives the network design specialist activities that were originally planned our in the previous stage. The network will be designed exactly with the initial requirements and will add additional data if required during the analysis and network audits. A network design specification will be produced and it is a comprehensive design which will meet every requirement and integrate specifications that support availability, reliability, security, scalability, and performance. The basis for the implementation stage uses the design specification.

Implementation Phase – Once the design has been approved that is when the implementation starts. The network will be built using the designs specifications with the goal of combining the devices within the network without disturbing the existing network or creating vulnerabilities.

Operate Phase – This stage is to test the design to see if it appropriate. It is used to sustain the health of the network through day-to-day operation. Any initial data from the network can be detected with fraud detection and correction and performance monitoring.

Optimise Phase - This phase is based on network management with the main goal to find and resolve any issues that have occurred before any real problems arise. If proactive management is not able to predict and reduce the failures then reactive fault detection and correction will be used. During this stage the network may need to be redesigned if too many problems or errors arise, if it does not perform up to the expectations, or when any mew applications have been identified that can support the technical requirements. (Ques10. (2020)).

3.4 Evaluation and Justification of using Methodology

This project will use qualitative methodology because the non-numerical data within the network traffic dataset that will be analysed shows the countries where the attack has come from and the protocol used by each individual attack. Quantitative methodology is being used because the numerical data that will be analysed will be the date and time that an attack happened, the source IP address from the computer that performed the attack, along with the source port and destination port. Graphs will be created to make it easier and quicker to analyse and to draw conclusions of all of the data. Machine learning will be used to detect anomalies and patterns within the data. Lastly network development methodology is being used with setting up the ELK Stack on a network. This was set up and configured on an Ubuntu VM through the command prompt.

The main problem with the data that has been collected is that it is from a third-party source on Kaggle. Kaggle is a website with a quarter of a million dataset and machine learning scripts.

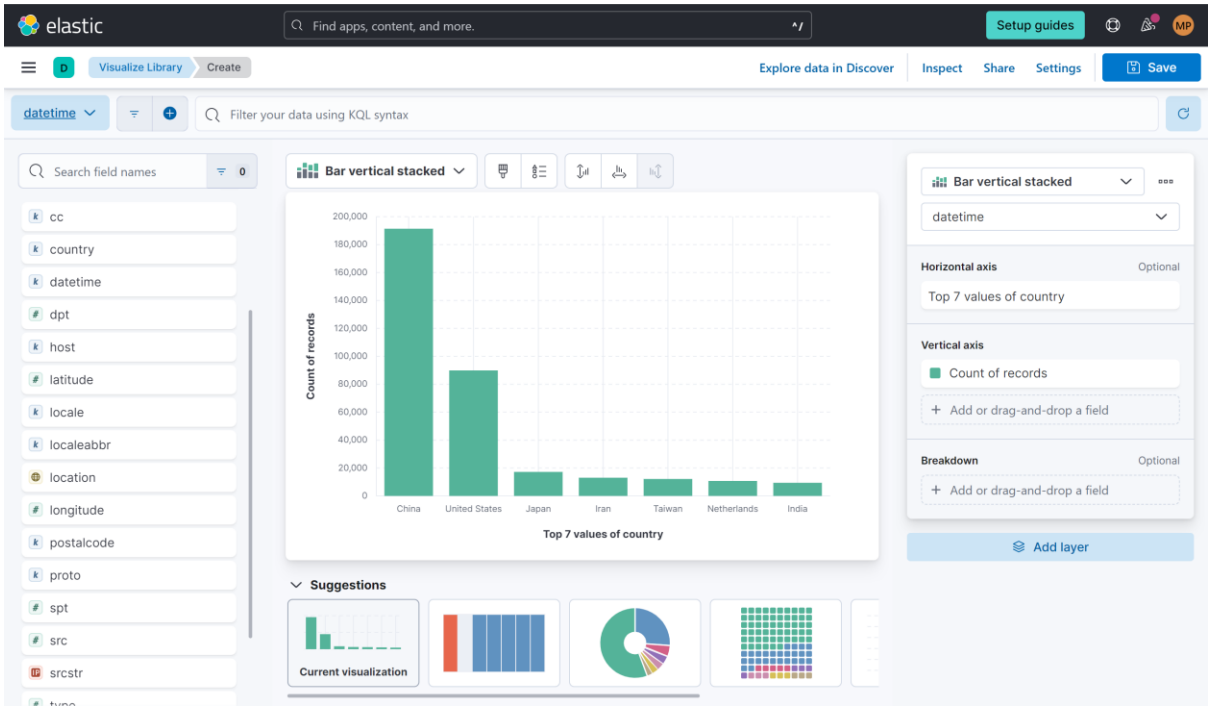
4.0 Practical Work

The data I will be analysing is taken from Kaggle. This data contains honeypot attack data whereby the attacks have taken place on a number of AWS instances. The purpose of analysing this data is to recognise trends, patterns and common attack types used by threat actors. The data has over 450,000 data points and spans over six months of attack data. The ELK Stack will be leveraged in order to gain insight into this data. Conclusions can be drawn and protection from future attacks can be stopped.

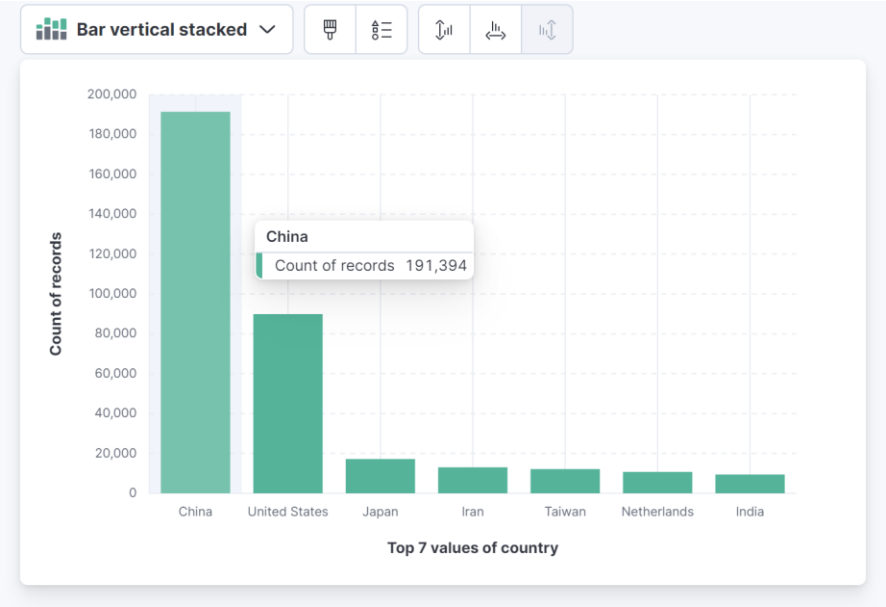
This dataset has the following datapoints, date/time, host, src, proto, type, spt, dpt, scrstr, cc, country, locale, localeabbr, postalcode, latitude and longitude.

4.1 Analysis

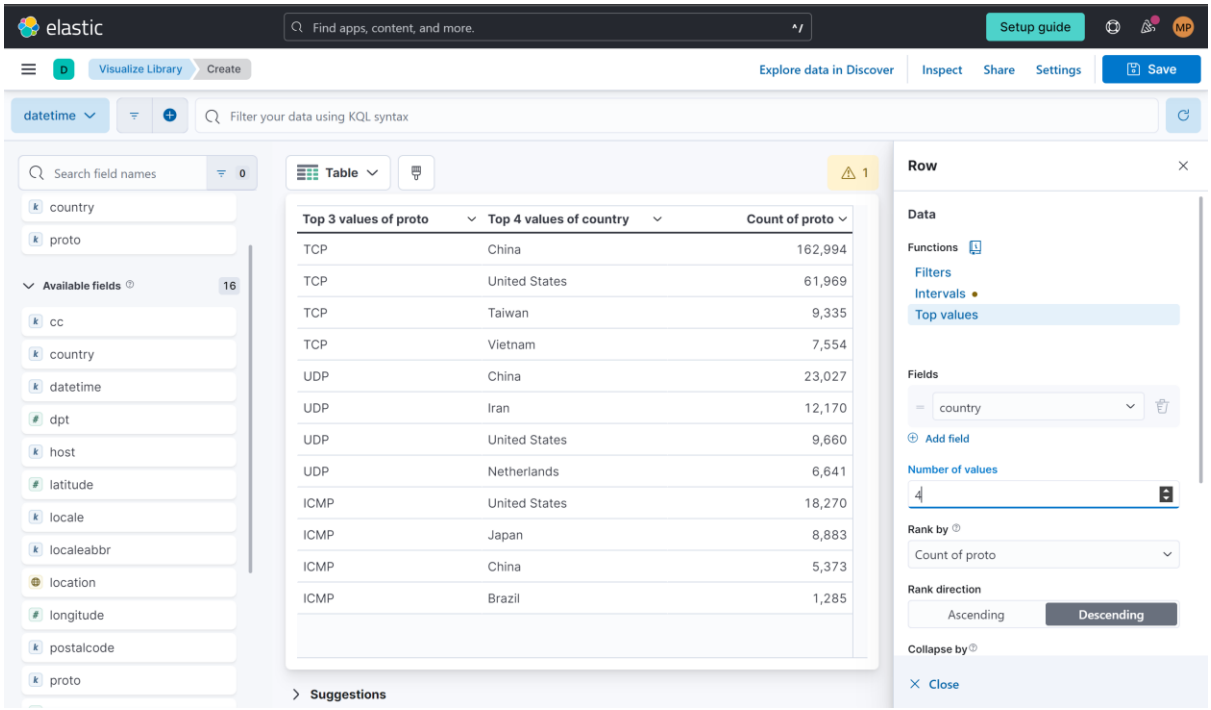
The first piece of analysis conducted was discovering the patterns associated with the country where the attack originated. As you can see from the analysis below, the overwhelming majority of attacks originated from either China or America.



Of the 450,000+ attacks 191,394 originate from China. This shows that Chinese capability and intention is the most prevalent over the six-month period.

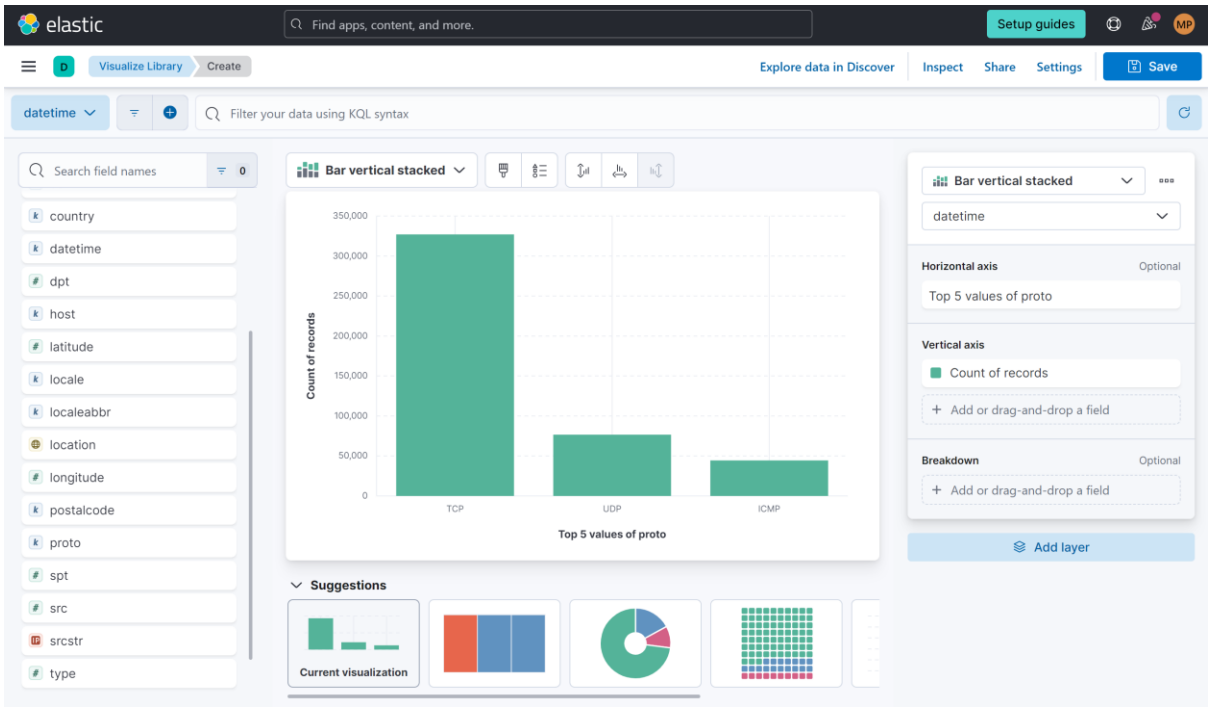


Here are the top four countries grouped by protocol. It is clear China is leveraging all three protocols that are in the dataset, with TCP being the most with 162,994 counts. China and United States both feature in the top four countries for all three protocols. This could show varied attacks and could hint towards a more advanced threat. Brazil feature in the top four ICMP protocol origin which could hint towards less sophisticated threat actors, given the simplicity of ICMP based attacks (such as ping requests).

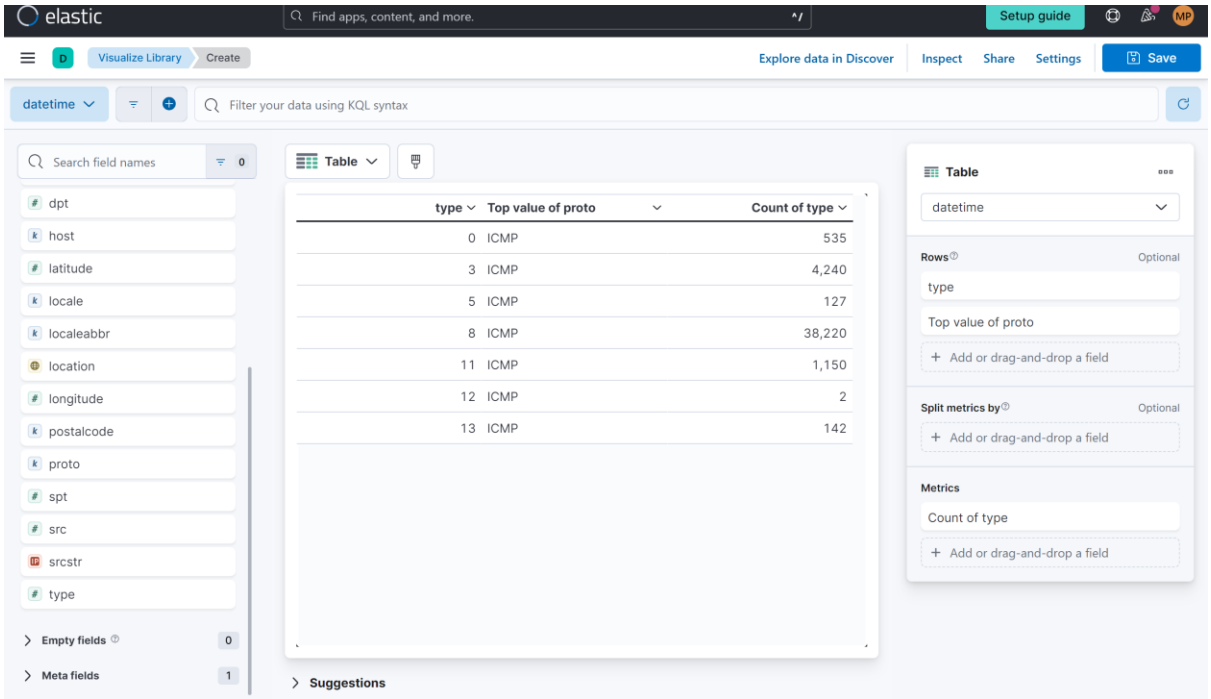


This shows the analysis of the protocols leveraged by the threat actors. It is clear from the analysis that the overwhelming majority of attacks leverage the TCP protocol. This makes

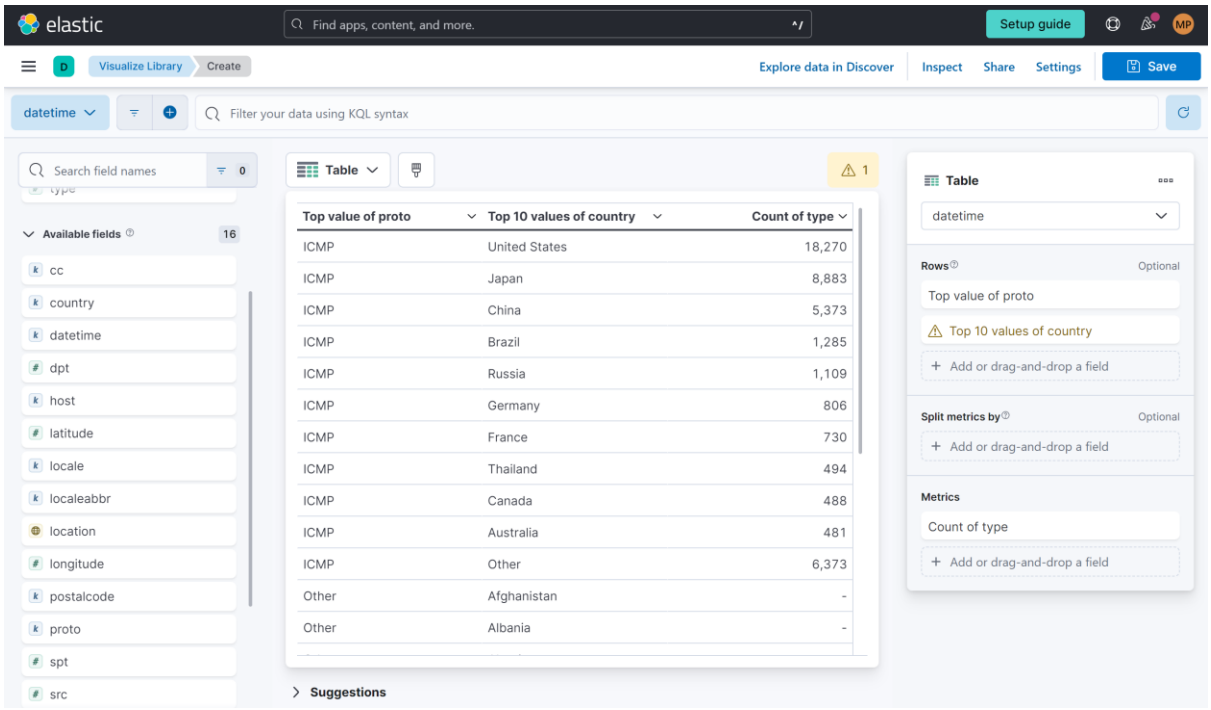
sense as most attacks will be web based and the https protocol rely on the TCP stack. Further to this ICMP has also been leveraged. Threat actors could be using this protocol to perform discoverability exercises.



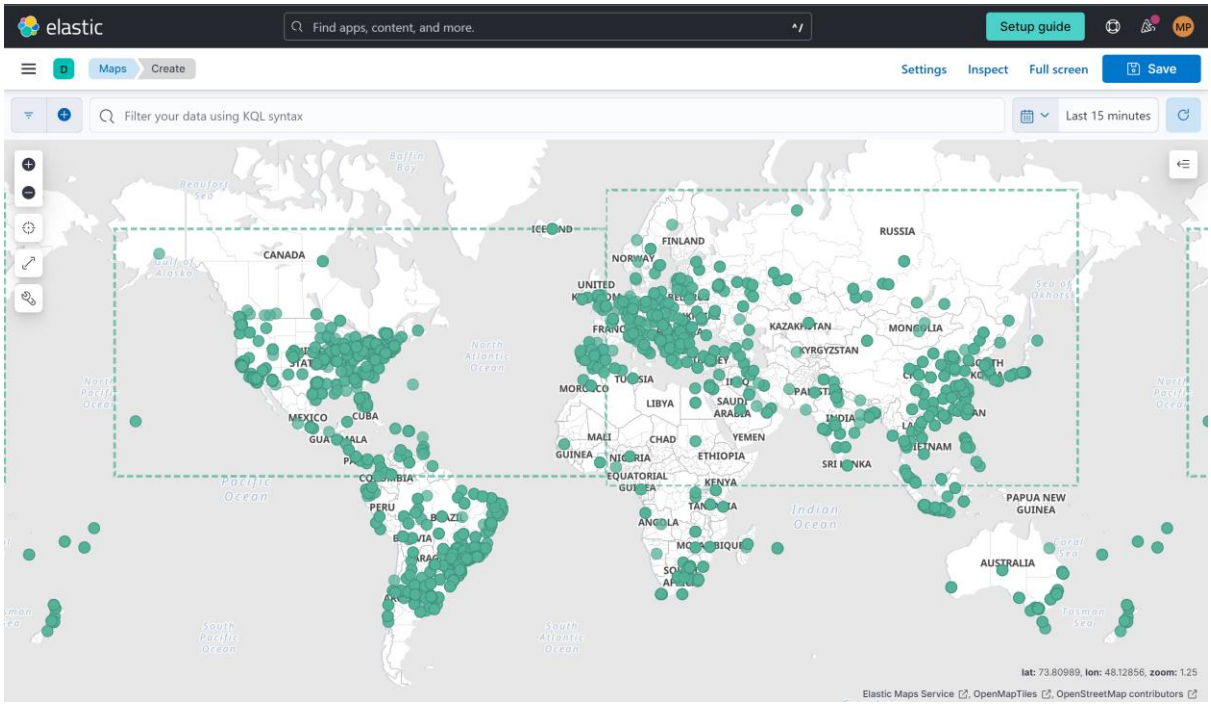
Further analysing the ICMP attack traffic we can view the ICMP type as per the ICMP protocol. As shown in the analysis below the most common ICMP type is type 8, which is an Echo request. This backs up earlier analysis and shows that threat actors are using ping requests to confirm an endpoint is in fact live.



Further analysing the ICMP traffic by country, it is clear that the threat actors based in the United States are leveraging ICMP more than other countries with Japan and China in second and third place.

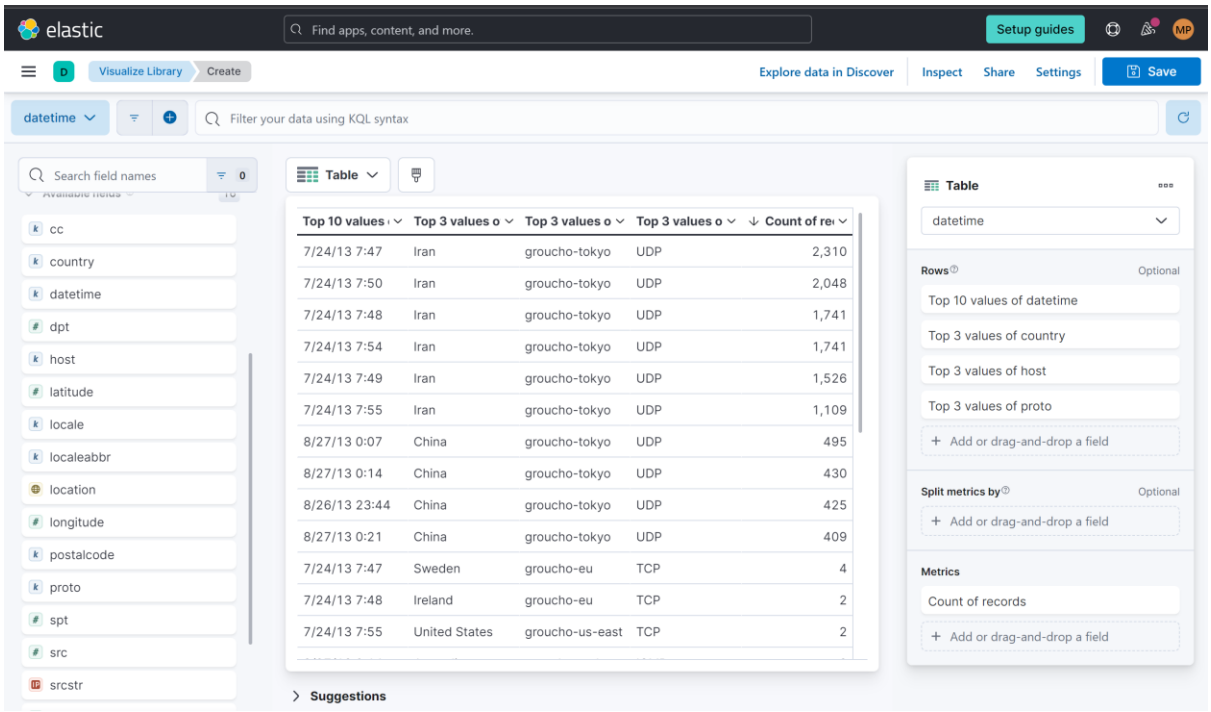


Here are the locations of each attack carried out over the six-month period. As we can see the majority of attacks do in-fact originate from China and the United States.

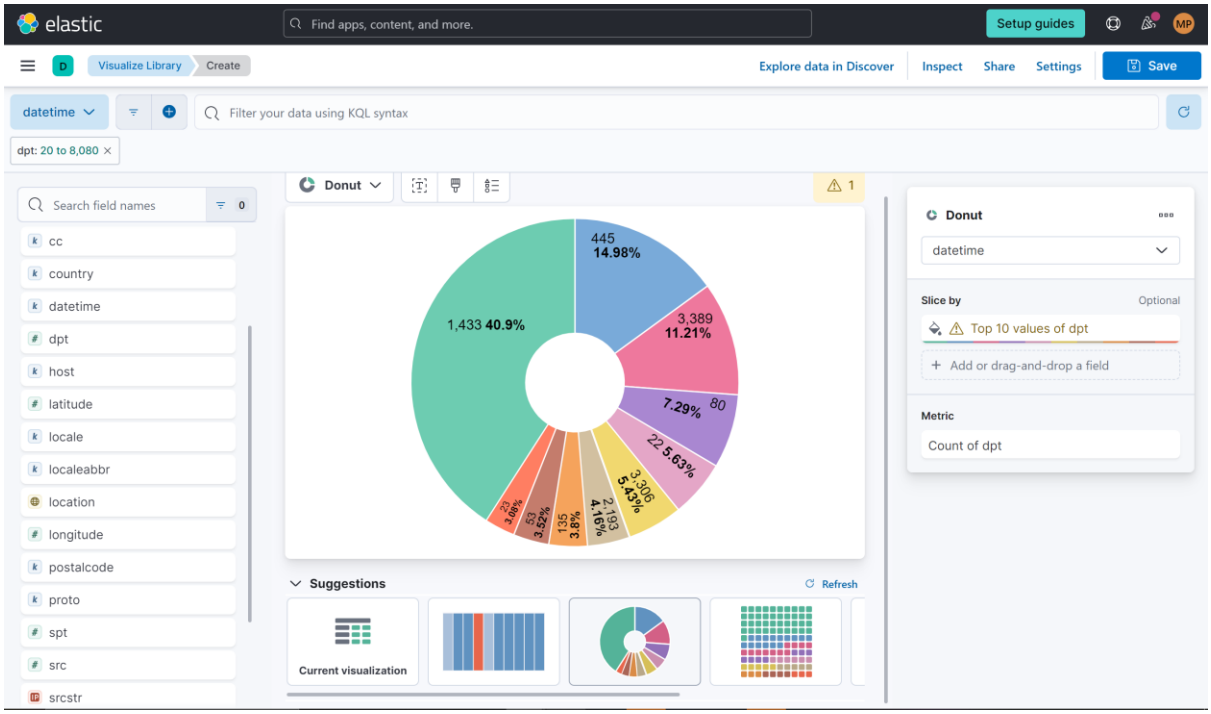


Here is the analysis of most of the dates of which the number of attacks were at its greatest during the six-month period. We can see the host which is the name of the honeypot under attack and its region. We can also see the number of attacks carried out, the origin of the

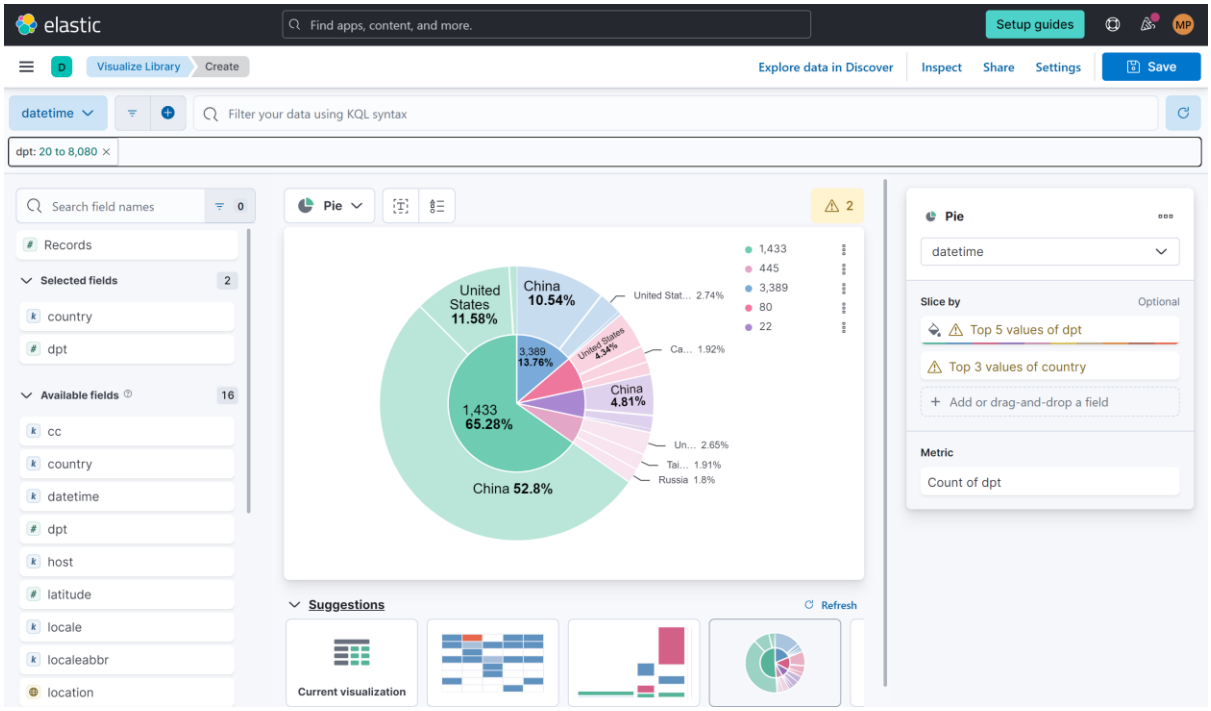
attack and the date and time of the attacks. Analysis shows that Iran on the 07/24/13 started an attack on Tokyo at 07:47 and ended at 07:55. In one minute they carried out 2,310 attacks and with the total attack being 8,949 over eight minutes. All of these attacks leveraged the UDP protocol.



So far we have established the main threat actors originate from China and the United States, who leverage all three protocols in the dataset. Analysis below shows the ten most common attacked ports represented by the ‘dpt’ column in the dataset. The most common attacked port is 1433. Usually on this port MSSQL operates (Microsoft SQL). From the analysis so far we can determine the American and Chinese threat actors attack MSSQL, SMB (port 445) and remote desktop which is port 3389. It is surprising that port 80 is not more commonly attacked. Representing only 7.29% of attacks in the six-month period.

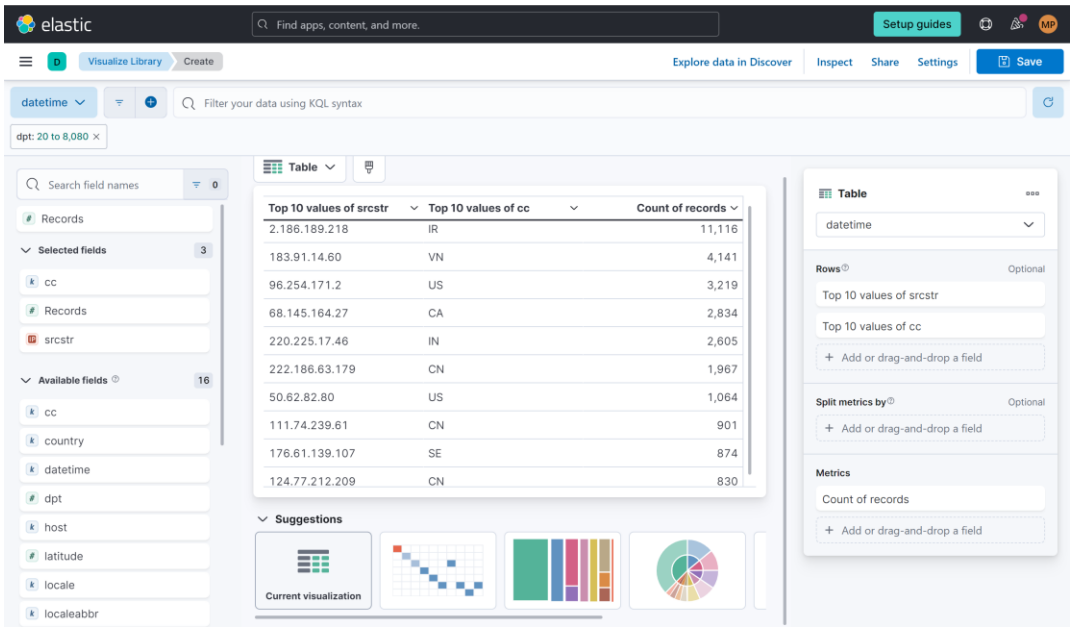


Further analysing the link between country of origin and the attacked service we can see the top three countries and the five most common attacked services. Threat actors from China attack MSSQL far more often than any other country, as is the case with the other top two attacked services, RDP and HTTP. American based attackers seem to prefer attacking port 80, HTTP, more than their counterparts. Analysis shows that Russian and Taiwanese based attackers also appear to prefer attacking port 445, SMB.

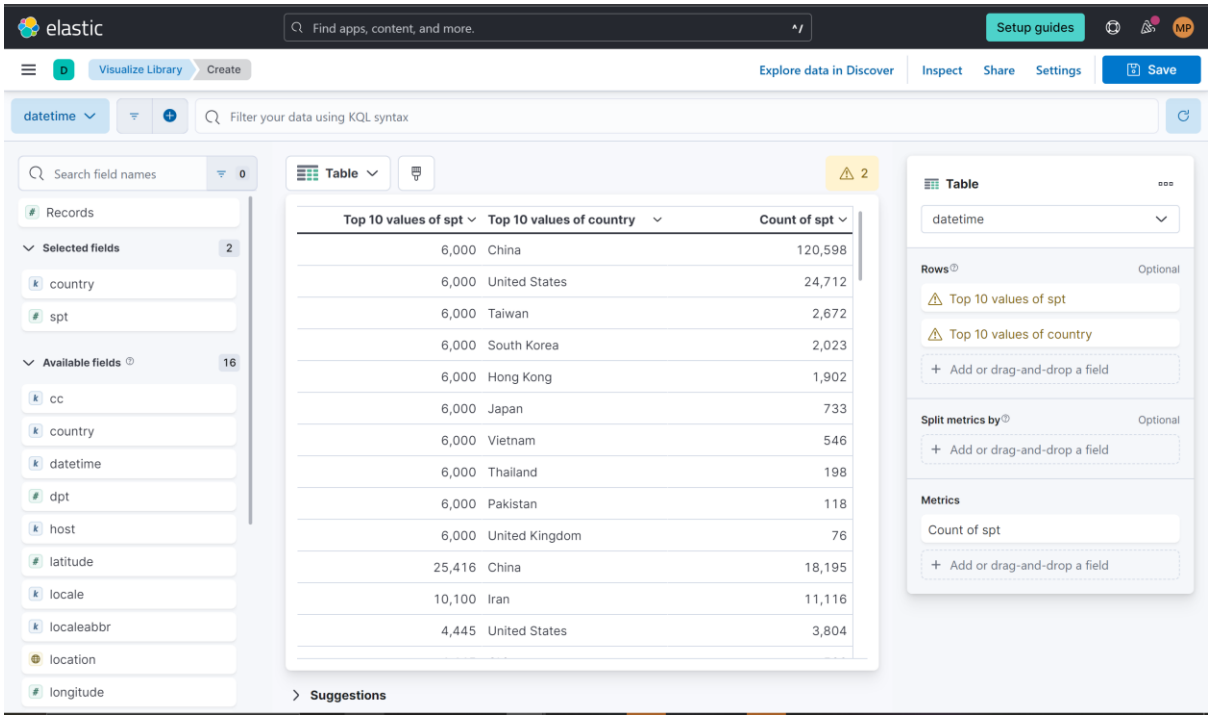


Analysis below shows the top ten IP addresses ordered by the number of times an attack originated from that IP address. We were also able to associate a country code with the IP address. Represented by the cc column. We can clearly see the most persistent threat actor

originated from Iran with a total of 11,116 attacks over the six-month period. This could show an advanced campaign originating from this country. It is interesting to see it is the only Iranian IP address in the top ten IP addresses. This could indicate alone attacker or a sophisticate campaign whereby the threat actors hide their real location.

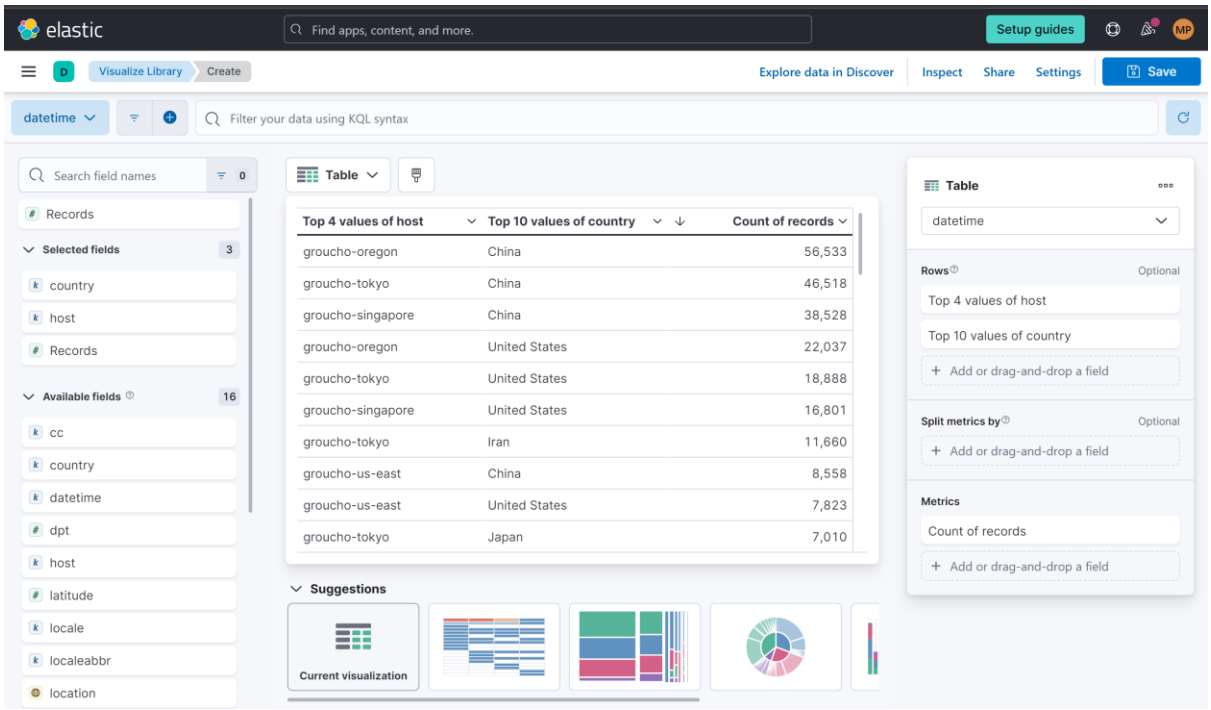


Here is the analysis of the top ten country of origin along side the source port leveraged by the threat actor of the country in question. Most threat actors from where most of the attacks originated leverage port 6000. This is shown below. Again, Chinese threat actors make use of this port the most during the six month time period.

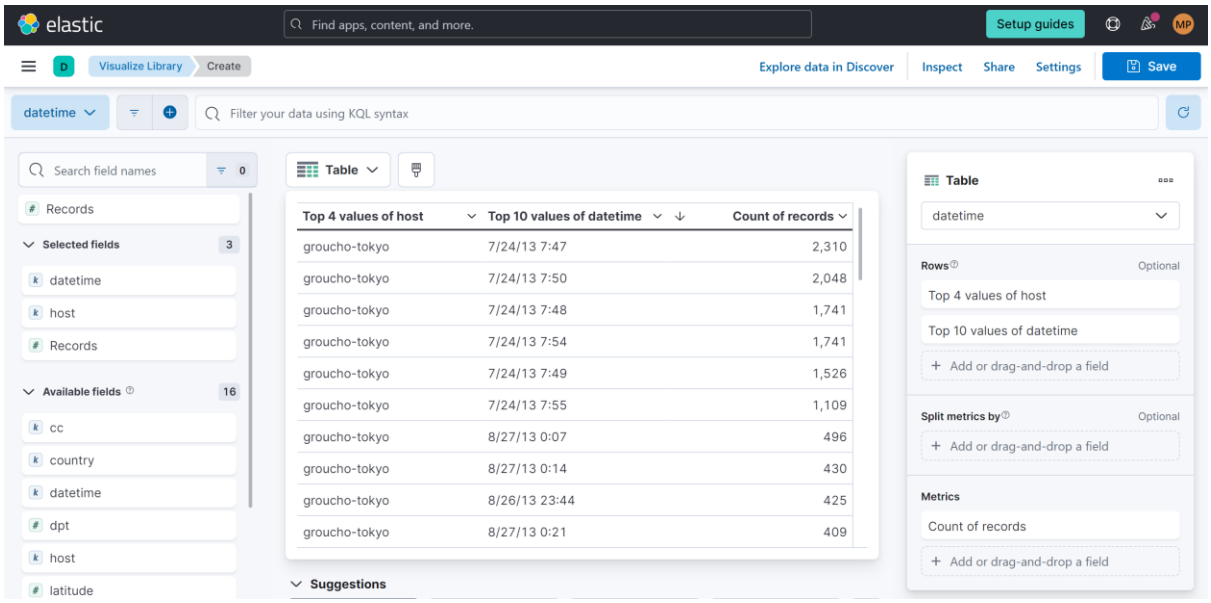


Performing analysis on the attacked hosts versus the country of attack origin presents some interesting data. As shown below, the most attacked host from the four within the dataset is

the node residing in Oregon, USA, followed by Tokyo, and Singapore. The origin of attack with the highest frequency of attacks is, in each case, China. Interestingly, The threat actors based in United States appear to attack the honeypot residing in their own country, shown clearly below.



In conjunction with this analysis, presented below is the host detection alongside the highest frequency of attacks. The highest number of attacks as shown previously was on 7/24/13, all attacks which were aimed at the honeypot located in Tokyo. The highest frequency of attacks, for the top ten, are all aimed at the Tokyo honeypot.

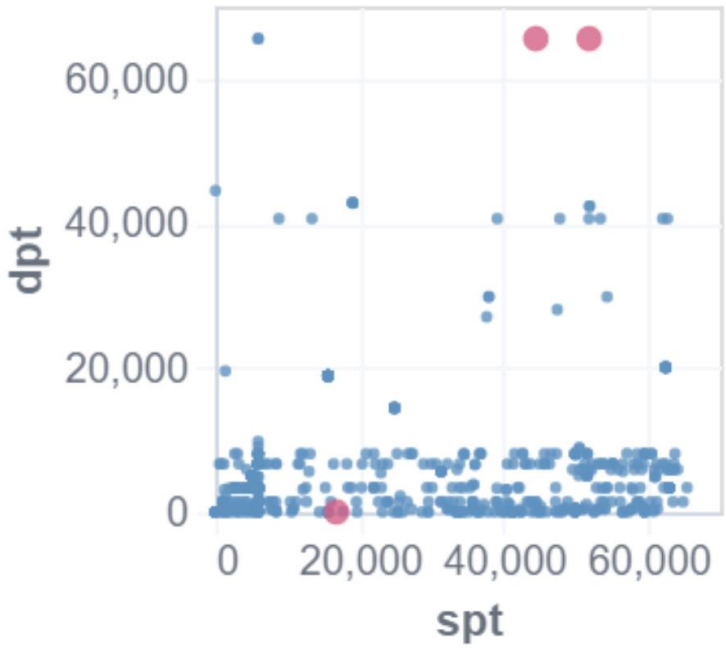


4.2 Machine Learning

Within the ELK stack, the ability to detect anomalies within data sets exists as part of the default configuration. This feature can be useful in this instance to show what attacks are not part of the norm, and what this can tell us. The ELK stack uses an outlier measurement to determine the anomalies, represented as an integer between 0 and 1; the higher the number, the higher the chance the representation is an outlier, or anomaly.

Based on the data set used for the analysis, investigating the link between the attacking source port and victim destination port presented itself as the most obvious linked data to determine outliers; both of these values are numerical and comparable in terms of use.

A threshold outlier score of 0.7 was set against a sample size of 1000 from the dataset. Viewing the results below, we can ascertain that most of the attacks were similar in nature.



Considering the source and destination port data, it is clear the source port does not seem to follow a pattern used by most attackers. The destination port clearly, for the most part, is in the sub-10000~ range, shown above. However, there are three outliers, represented by the larger pink circles.

Two of these outliers present an unusual desination port, 65500, which is not commonly leveraged by any common services. The third outlier presented attacking a common port, 22, which is commonly used for SSH. However, the high attacking port resulted in an outlier being presented.

5.0 Conclusions & Recommendations

It is clear from the data that threat actors around the world are actively scanning for and, most likely, exploiting vulnerabilities that exist on public facing assets. The dataset leveraged by this project investigated 451,581 attacks over a 189-day period. This equates to roughly 1.65 attacks every second. With this in mind, there are a number of patterns that have emerged from the analysis of the data:

1. China and USA are the most common origin of threat. Threat actors over the observed six-month period were predominantly from these two countries. Any individual or organisation wishing to place assets on the public-facing Internet should be aware of the capabilities of threat actors within these two counties.
2. MSSQL, SMB and RDP are the most attacked services. Throughout the time period observed within the dataset, these three services were attacked the most often. This could be due to the level of access granted by these services (in the case of RDP) or because of the data and information stored by these services. It is surprising to see that HTTP is not the most common attacked service. However, given the advancements in WAF technology and the protection given to web services in the modern era, threat actors could be moving onto lesser protected services. Individuals and organisations should protect these services, and if possible, do not expose them to the public facing internet unless absolutely necessary.
3. Most attacks originate from the same port. Analysing the data, most threat actors leverage port 6000 as an attacking port. This is not a commonly used port by any well-known service. Connections to an asset which has a source port of 6000 could prove to be a good indicator of compromise (IoC) based on the dataset.
4. ICMP is leveraged by threat actors based in USA extensively. Allowing any threat actor to gain confirmation that the asset is live on the internet allows them to further enumerate and possibly exploit the system in question. ICMP echoes should be disabled for any endpoint with this in mind, unless absolutely required.
5. The majority of attacks leverage the TCP protocol. This makes sense in relation to the services which are being attacked. It is difficult to block all TCP traffic, but users and organisations should be aware that these services are attacked the most.
6. There appears to be no patterns presented in terms of the honeypots attacked and the origin country of the threat actor. This is a humble reminder that, most of the time, attacks can be domestic or international in nature. Threat actors can vary in terms of levels of sophistication, tools leveraged, and intent.

Whilst it is clear that the ELK stack does have its advantages to data processing, data visualisation and analysis, alongside machine learning, there are a number of features which are lacking which could have increased the value as a result of the analysis of the data. This includes the lack of ability to perform machine learning analysis, and outlier analysis, on data that is not integer based. It would have been useful, in relation to the aims of the project, to perform further outlier analysis based on locations of attacks and other geolocation data.

A further recommendation would be to monitor the attacks on honeypots which over a longer period of time. This could be useful to gain further insight into the intent of threat actors, and their targets. Allowing a service to run with zero authentication could also be of use; monitoring the intent of threat actors, once they have gained access to an asset, could provide a clear picture of the tools, techniques and procedures (TTPs) leveraged by threat actors, not only from an externally facing perspective, but also from an 'assume breached' angle.

6.0 Critical Self-Evaluation

My performance throughout my project, from my perspective, has been good. I have had a clear plan from the inception of the project and stuck to it to the best of my abilities. Following a clear schedule and sticking to timescales has been an aim of mine. Whilst it has not always been possible to follow the plan, this has not always been in my complete control. If I was to do this project again, I would aim to start the project sooner in order to avoid any issues surrounding timescales.

I feel as though my planning for this project has been to a standard I am happy with. Much like following and sticking to my deadlines, I have planned this project at each stage. Using Gannt charts, written plans and following the advice from my university mentors, the planning for this project has been successful, and I have finished on time.

Communication between my university mentors has been good, and I feel as though my side of my communication has always been timely and professional. I have always sought out help if I have needed it and always responded to any and all communication in a professional manner, on time.

The data set used within my project could have been made better in several ways. Whilst the analysis of the data provides some actionable recommendations, I feel as though the dataset itself could have contained more information, such as attack payloads used, User Agent strings (where applicable), and any credential stuffing data. This could have provided further insight into the intent of the threat actors, allowing us to draw more sound conclusions from the data. Further to this, accessing the lateral movement data which could have been leverage (if applicable), could again provide further insights into how threat actors from across the globe gain access, pivot and escalate privileges to affected assets. The data set could have also included data from a longer time period, allowing for a deeper understanding over a longer period of time.

Bibliography

Abdou, F, Lemine, M, Farouk, M. (2019). Toward a Secure ELK Stack. International Journal of Computer Science and Information Security. 17(7), pp.139-140. [Online]. Available at: https://www.academia.edu/40206223/Toward_a_Secure_ELK_Stack [Accessed 2 June 2023].

Ali, H. (2023). ELK Stack Guide: What Is Elasticsearch, Logstash, and Kibana?. [Online]. Software Suggest. Last Updated: 20th February 2023. Available at: https://www.softwaresuggest.com/blog/elk-stack-tutorial/# [Accessed 7 June 2023].

Bajer, B. (2017). Building an IoT Data Hub with Elasticsearch, Logstash and Kibana. International Conference on Future Internet of Things and Cloud Workshops. 5(1), pp.65-66. [Online]. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8113772 [Accessed 7 June 2023].

Brooks, C. (2023). Cybersecurity Trends & Statistics For 2023; What You Need To Know. [Online]. Forbes. Last Updated: 5 March 2023. Available at: https://www.forbes.com/sites/chuckbrooks/2023/03/05/cybersecurity-trends--statistics-for-2023-more-t [Accessed 10 April 2023].

Elastic. (2023). What is structured data?. [Online]. Elastic. Available at: https://www.elastic.co/what-is/unstructured-data [Accessed 1 July 2023].

Elastic. (2023). What is unstructured data?. [Online]. Elastic. Available at: https://www.elastic.co/what-is/unstructured-data [Accessed 1 July 2023].

Eteng, O. (2022). Quantitative Data Analysis: Methods & Techniques Simplified 101. [Online]. Havo. Last Updated: May 18th, 2022. Available at: https://hevodata.com/learn/quantitative-data-analysis/ [Accessed 25 March 2023].

Gates, C, Taylor, C. (2006). Challenging the anomaly detection paradigm: a provocative discussion. Proceedings of the 2006 workshop on New security paradigms. 6(6), p.22. [Online]. Available at: https://dl.acm.org/doi/pdf/10.1145/1278940.1278945 [Accessed 3 July 2023].

Kleindienst, P. (2016). Building areal-world logging infrastructure with Logstash, Elasticsearch and Kibana. BertschInnovationGmbH. 4(5), p.2. [Online]. Available at: https://hdms.bsz-bw.de/files/5021/elk_paper_patrick_kleindienst.pdf [Accessed 2 June 2023].

Kononenko, O, Baysal, O, Holmes, R, Godfrey, M. (2014). Mining Modern Repositories with Elasticsearch. Working Conference on Mining Software Repositories. 11(978-1-4503-2863-0), pp.328-331. [Online]. Available at: https://dl.acm.org/doi/pdf/10.1145/2597073.2597091 [Accessed 4 April 2023].

Mamidwar, S. (2022). How to Install Elastic Stack 8 on Ubuntu 20.04 LTS. [Online]. Foss RechNix. Available at: https://www.fosstechnix.com/how-to-install-elastic-stack-8-on-ubuntu-20-04/ [Accessed 8 June 2023].

Örkcü, H, Bal, H. (2011). Comparing performances of backpropagation and genetic algorithms in the data classification. Expert Systems with Applications. 38(4), pp.3703-3704. [Online]. Available at: https://pdf.sciencedirectassets.com/271506/1-s2.0-S0957417410X00124/1-s2.0-S0957417410009851/main.pd [Accessed 29 June 2023].

Paliwal, M. (2022). Elasticsearch vs Splunk - Which tool to choose for Log Management?. [Online]. Sig Noz. Last Updated: 03 November 2022. Available at: https://signoz.io/blog/elasticsearch-vs-splunk/ [Accessed 2 July 2023].

Point, T. (2023). What is ELK Stack?. [Online]. Tutorials Point. Available at: https://www.tutorialspoint.com/kibana/kibana_overview.htm#:~:text=Advantages%20of%20Kibana&text=Cont [Accessed 7 June 2023].

Praveen, S, Chandra, U. (2017). Influence of Structured, SemiStructured, Unstructured data on various data models. International Journal of Scientific & Engineering Research. 8(12),. pp.68-69. [Online]. Available at: https://www.ijser.org/researchpaper/Influence-of-Structured--Semi-Structured--Unstructured-data-on-v [Accessed 29 June 2023].

Ques10. (2020). Network Design Methodology. [Online]. Ques10. Available at: https://www.ques10.com/p/50368/network-design-methodology/ [Accessed 3 April 2023].

Rassam, M, Maarof, M, Zainal, A. (2017). Big Data Analytics Adoption for Cybersecurity: A Review of Current Solutions, Requirements, Challenges and Trends. Journal of Information Assurance and Security. 11(1), pp.134-136. [Online]. Available at: http://mirlabs.org/jias/secured/Volume12-Issue4/Paper14.pdf [Accessed 25 June 2023].

Souris, S. (2022). A Brief History of Elasticsearch: From Lucene to the Future. [Online]. Stathis Souris. Last Updated: 01 December 2022. Available at: https://www.stathissouris.com/blog/a-brief-history-of-elasticsearch-from-lucene-to-the-future [Accessed 5 June 2023].

Stoleriu, R, Puncioiu, A, Bica, I. (2021). Cyber Attacks Detection Using Open Source ELK Stack. International Conference on Electronics, Computers and Artificial Intelligence (ECAI). 13(13), pp.1-2. [Online]. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9515120 [Accessed 2 July 2023].

Supreeth, S, Rajendran, S. (2022). Automated Switching Crash Analysis using ELK for Log Analysis. International Research Journal of Engineering and Technology (IRJET). 2395-0072(7), p.6062. [Online]. Available at: https://d1wqtxts1xzle7.cloudfront.net/64609996/IRJET-V7I51153-libre.pdf?1601991119=&response-content [Accessed 10 April 2023].

Tutorials Point. (2023). Logstash - Introduction. [Online]. Tutorials Point. Available at: https://www.tutorialspoint.com/logstash/logstash_introduction.htm# [Accessed 4 June 2023].

Wang, B, Hua, Q, Zhang, H, Tan, X, Nan, Y, Chen, R, Shu, X. (2022). Research on anomaly detection and real-time reliability evaluation with the log of cloud platform. Alexandria Engineering Journal. 61(9), pp.7185-7186. [Online]. Available at: https://pdf.sciencedirectassets.com/270704/1-s2.0-S1110016822X00045/1-s2.0-S1110016821008711/main.pd [Accessed 20 June 2023].

-

Appendix A

Setup of the ELK Stack

Firstly, all of the system packages were updated using *'sudo apt update'*.

```
michelle6391@michelle6391-virtual-machine:~$ sudo apt update
[sudo] password for michelle6391:
Hit:1 http://gb.archive.ubuntu.com/ubuntu jammy InRelease
Get:2 http://gb.archive.ubuntu.com/ubuntu jammy-updates InRelease [119 kB]
Get:3 http://security.ubuntu.com/ubuntu jammy-security InRelease [110 kB]
Get:4 http://gb.archive.ubuntu.com/ubuntu jammy-backports InRelease [108 kB]
Get:5 http://gb.archive.ubuntu.com/ubuntu jammy-updates/main amd64 DEP-11 Metadata [102 kB]
Get:6 http://gb.archive.ubuntu.com/ubuntu jammy-updates/universe amd64 DEP-11 Metadata [269 kB]
Hit:7 https://artifacts.elastic.co/packages/8.x/apt stable InRelease
Get:8 http://gb.archive.ubuntu.com/ubuntu jammy-updates/multiverse amd64 DEP-11 Metadata [940 B]
Get:9 http://gb.archive.ubuntu.com/ubuntu jammy-backports/main amd64 DEP-11 Metadata [7,972 B]
Get:10 http://gb.archive.ubuntu.com/ubuntu jammy-backports/universe amd64 DEP-11 Metadata [12.5 kB]
Get:11 http://security.ubuntu.com/ubuntu jammy-security/main amd64 DEP-11 Metadata [41.6 kB]
Get:12 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 DEP-11 Metadata [18.5 kB]
Fetched 789 kB in 1s (1,325 kB/s)
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
124 packages can be upgraded. Run 'apt list --upgradable' to see them.
```

Next the apt-transport-https packages were installed. What this does is allow access to all of the https repositories. The command used for this was *'sudo apt install apt-transport-https'*.

```
michelle6391@michelle6391-virtual-machine:~$  sudo apt install apt-transport-https
[sudo] password for michelle6391:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed
  apt-transport-https
0 to upgrade, 1 to newly install, 0 to remove and 125 not to upgrade.
Need to get 1,506 B of archives.
After this operation, 169 kB of additional disk space will be used.
Get:1 http://gb.archive.ubuntu.com/ubuntu jammy-updates/universe amd64 apt-transport-https all 2.4.8 [1,506 B]
Fetched 1,506 B in 0s (16.9 kB/s)
Selecting previously unselected package apt-transport-https.
(Reading database ... 200635 files and directories currently installed.)
Preparing to unpack .../apt-transport-https_2.4.8_all.deb ...
Unpacking apt-transport-https (2.4.8) ...
```

OpenJDK 11 was installed using *'sudo apt install openjdk-11-jdk'*.

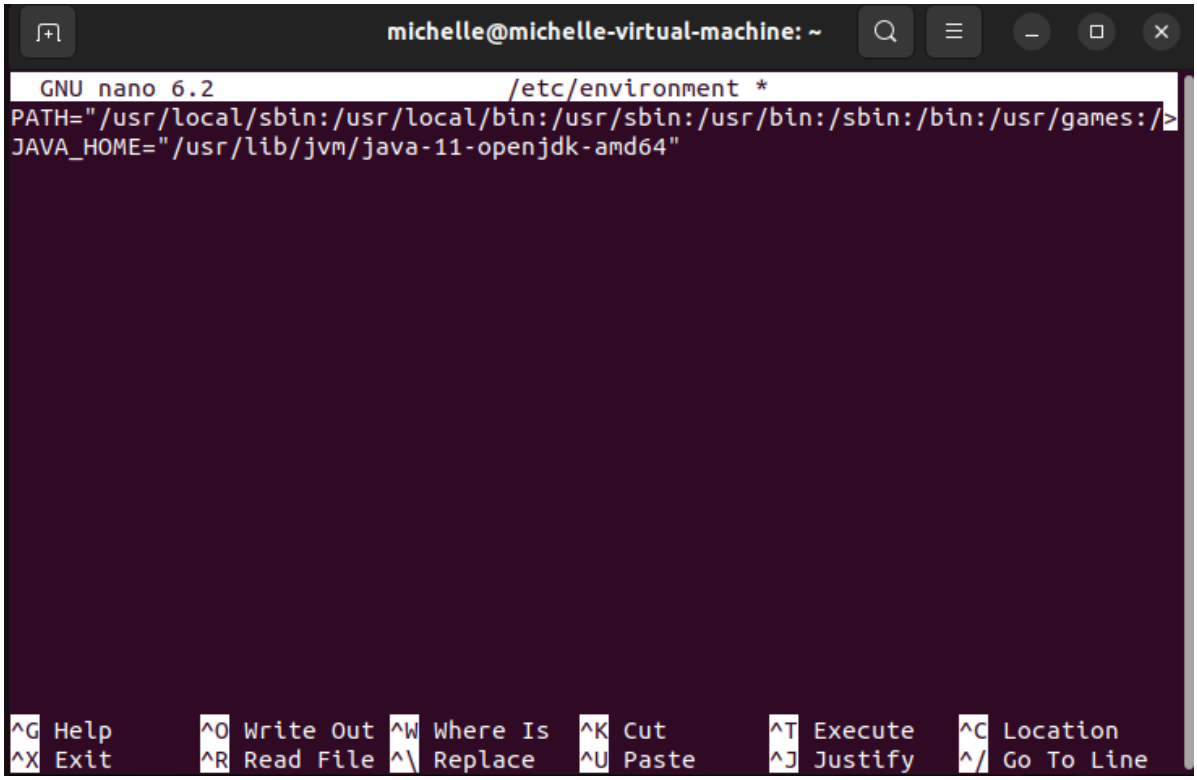
```
michelle6391@michelle6391-virtual-machine:~$ sudo apt install openjdk-11-jdk
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java
  libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev
  libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev
  openjdk-11-jdk-headless openjdk-11-jre openjdk-11-jre-headless x11proto-dev
  xorg-sgml-doctools xtrans-dev
Suggested packages:
  default-jre libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc
  openjdk-11-demo openjdk-11-source visualvm fonts-ipafont-gothic
  fonts-ipafont-mincho fonts-wqy-microhei | fonts-wqy-zenhei
The following NEW packages will be installed
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java
  libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev
  libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev openjdk-11-jdk
  openjdk-11-jdk-headless openjdk-11-jre openjdk-11-jre-headless x11proto-dev
  xorg-sgml-doctools xtrans-dev
0 to upgrade, 20 to newly install, 0 to remove and 125 not to upgrade.
Need to get 262 MB of archives.
After this operation, 413 MB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://gb.archive.ubuntu.com/ubuntu jammy/main amd64 java-common all 0.72build2 [6,782 B]
Get:2 http://gb.archive.ubuntu.com/ubuntu jammy-updates/main amd64 openjdk-11-jre-headless amd64 11.0.18+10-0ubuntu1~22.04 [41.6 MB]
Get:3 http://gb.archive.ubuntu.com/ubuntu jammy-updates/main amd64 ca-certificates-java all 20190909ubuntu1.1 [12.0 kB]
Get:4 http://gb.archive.ubuntu.com/ubuntu jammy/main amd64 fonts-dejavu-extra all 2.37-2build1 [2,041 kB]
Get:5 http://gb.archive.ubuntu.com/ubuntu jammy/main amd64 libatk-wrapper-java all 0.38.0-5build1 [53.1 kB]
Get:6 http://gb.archive.ubuntu.com/ubuntu jammy/main amd64 libatk-wrapper-java-jni amd64 0.38.0-5build1 [49.0 kB]
Get:7 http://gb.archive.ubuntu.com/ubuntu jammy/main amd64 xorg-sgml-doctools all 1:1.11-1.1 [10.9 kB]
Get:8 http://gb.archive.ubuntu.com/ubuntu jammy/main amd64 x11proto-dev all 2021.5-1 [604 kB]
Get:9 http://gb.archive.ubuntu.com/ubuntu jammy/main amd64 libice-dev amd64 2:1.0.10-1build2 [51.4 kB]
Get:10 http://gb.archive.ubuntu.com/ubuntu jammy/main amd64 libpthread-stubs0-dev amd64 0.4-1build2 [5,516 B]
Get:11 http://gb.archive.ubuntu.com/ubuntu jammy/main amd64 libsm-dev amd64 2:1.2.3-1build2 [18.1 kB]
```

The version of java was checked using *'java -version'*. As shown below the version is 11.0.18.

```
michelle6391@michelle6391-virtual-machine:~$ java -version
openjdk version "11.0.18" 2023-01-17
OpenJDK Runtime Environment (build 11.0.18+10-post-Ubuntu-0ubuntu122.04)
OpenJDK 64-Bit Server VM (build 11.0.18+10-post-Ubuntu-0ubuntu122.04, mixed mode, sharing)
```

'sudo nano /etc/environment' was used to open the environment so that the environment variable could be defined. This was done by adding the line *'JAVA_HOME="/usr/lib/jvm/java-11-openjdk-amd64"'*.

```
michelle6391@michelle6391-virtual-machine:~$ sudo nano /etc/environment
```



'source /etc/environment' was used to load the environment.

```
michelle6391@michelle6391-virtual-machine:~$ source /etc/environment
```

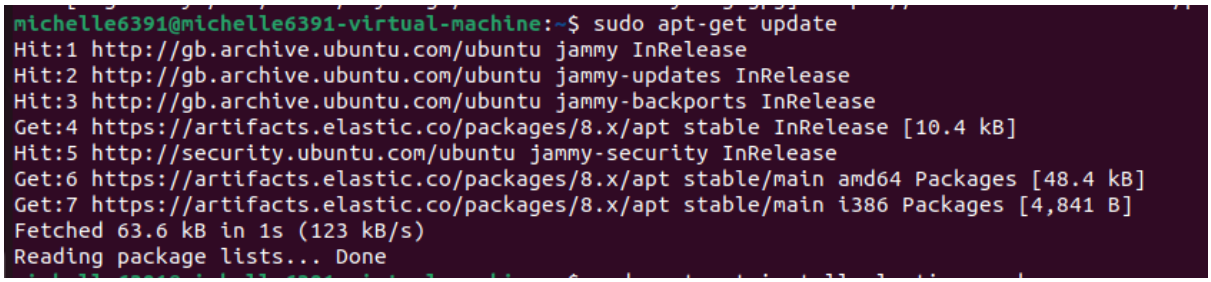
JAVA_HOME variable was verified using *'echo $JAVA_HOME'*.

```
michelle6391@michelle6391-virtual-machine:~$ echo $JAVA_HOME
/usr/lib/jvm/java-11-openjdk-amd64
```

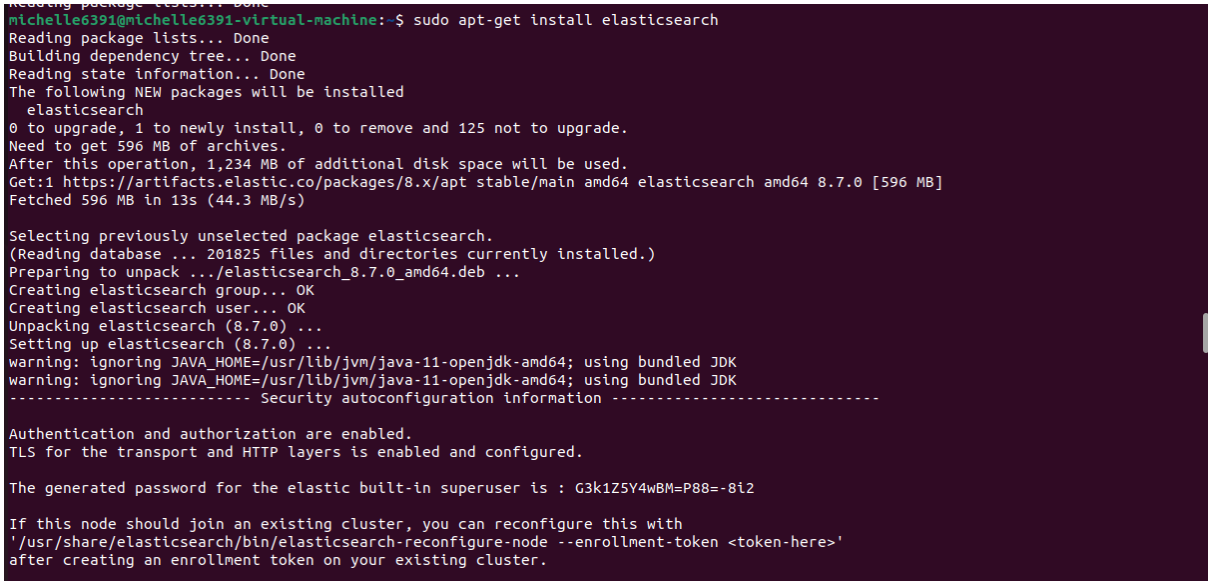
The public signing key was downloaded and installed by using the command *'wget -qO - https://artifacts.elastic.co/GPG-KEY-elasticsearch | sudo gpg --dearmor -o /usr/share/keyrings/elasticsearch-keyring.gpg'*

'echo "deb [signed-by=/usr/share/keyrings/elasticsearch-keyring.gpg] https://artifacts.elastic.co/packages/8.x/apt stable main" | sudo tee /etc/apt/sources.list.d/elastic-8.x.list' was used to save the repositories.

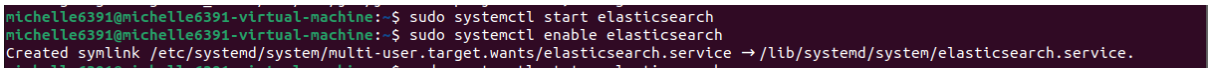
Another update was performed.



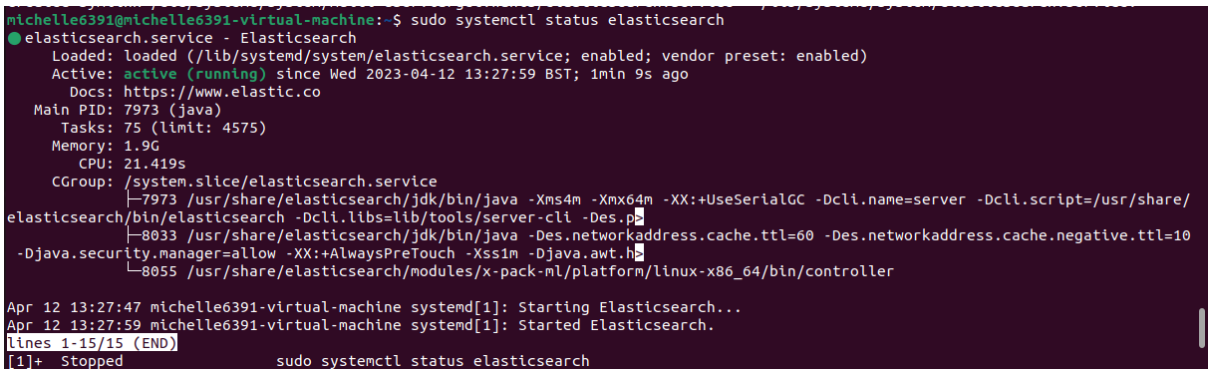
Elasticsearch was installed with the command *'sudo apt-get install elasticsearch'*



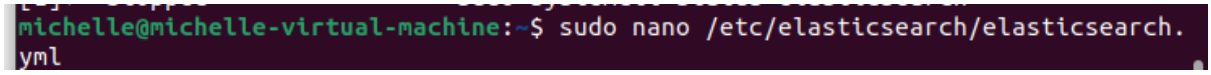
Elasticsearch was started with the command *'sudo systemctl start elasticsearch'* and enabled with *'sudo systemctl enable elasticsearch'*.



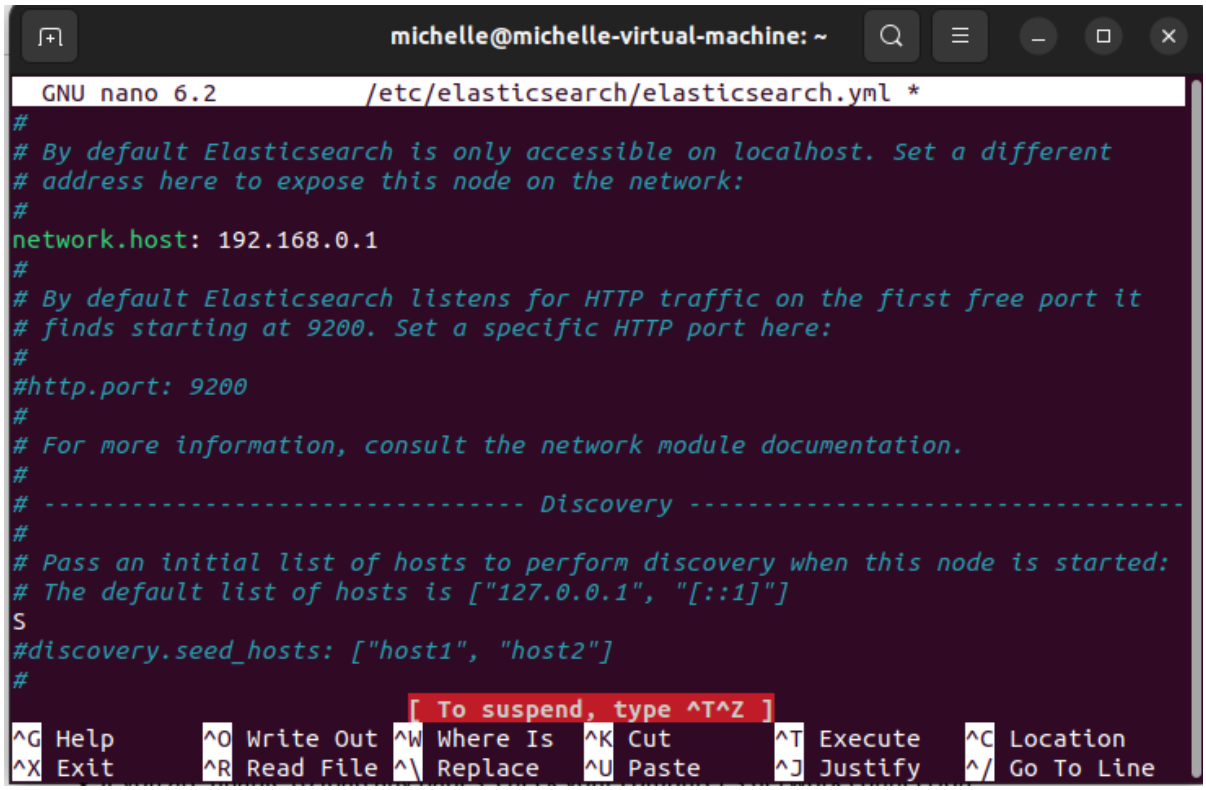
The status of Elasticsearch was checked with *'sudo systemctl status elasticsearch'*. As shown below it is working.

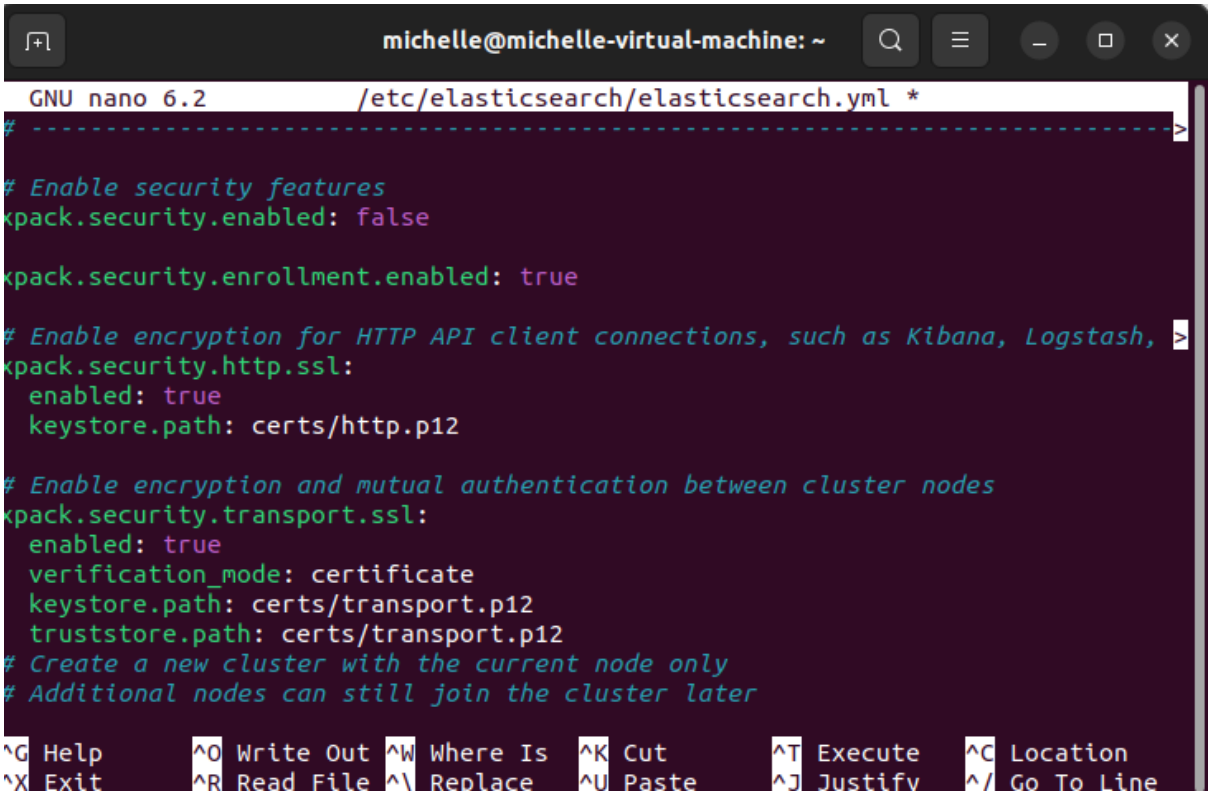


'sudo nano /etc/elasticsearch/elasticsearch.yml' took me to the configuration file

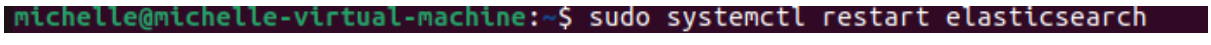


Within the configuration file network.host: 192.168.0.1 was uncommented and changed to 0.0.0.0. discover.seed_host: [ ] was added in above . discover.seed_host: ["host1", "host2"]. Further down xpack.security.enabled: true was changed to false.





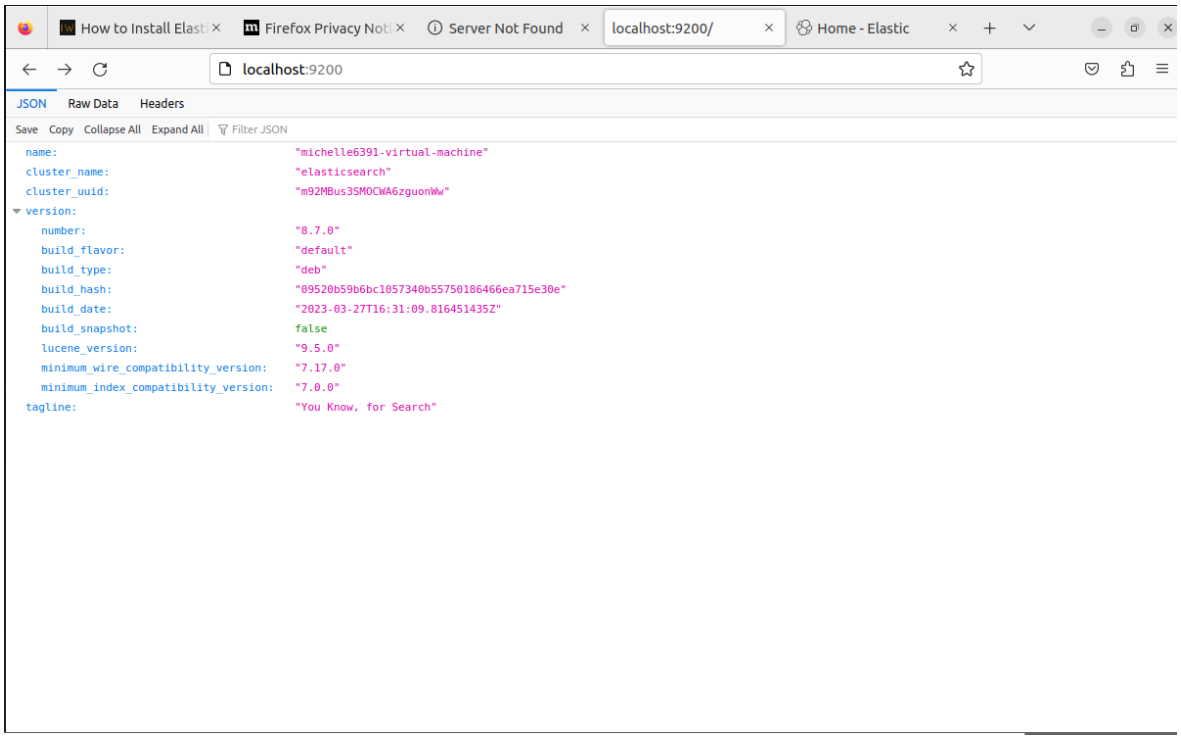
Once the changes had been changed Elasticsearch was restarted with *'sudo systemctl restart elasticsearch'*.



Elasticsearch was then tested using the curl command *'curl -X GET "localhost:9200"'.* What this does is send HTTP requests.

```
michelle6391@michelle6391-virtual-machine:~$ curl -X GET "localhost:9200"
{
  "name" : "michelle6391-virtual-machine",
  "cluster_name" : "elasticsearch",
  "cluster_uuid" : "m92MBus3SMOCWA6zguonWw",
  "version" : {
    "number" : "8.7.0",
    "build_flavor" : "default",
    "build_type" : "deb",
    "build_hash" : "09520b59b6bc1057340b55750186466ea715e30e",
    "build_date" : "2023-03-27T16:31:09.816451435Z",
    "build_snapshot" : false,
    "lucene_version" : "9.5.0",
    "minimum_wire_compatibility_version" : "7.17.0",
    "minimum_index_compatibility_version" : "7.0.0"
  },
  "tagline" : "You Know, for Search"
}
```

It was then tested in a web browser by using http://localhost:9200.



Next Logstash was installed by using *'sudo apt-get install logstash'*.

```
michelle6391@michelle6391-virtual-machine:~$ sudo apt-get install logstash
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed
  logstash
0 to upgrade, 1 to newly install, 0 to remove and 124 not to upgrade.
Need to get 327 MB of archives.
After this operation, 579 MB of additional disk space will be used.
Get:1 https://artifacts.elastic.co/packages/8.x/apt stable/main amd64 logstash amd64 1:8.7.0-1 [327 MB]
Fetched 327 MB in 8s (39.3 MB/s)

Selecting previously unselected package logstash.
(Reading database ... 203156 files and directories currently installed.)
Preparing to unpack .../logstash_1%3a8.7.0-1_amd64.deb ...
Unpacking logstash (1:8.7.0-1) ...
Setting up logstash (1:8.7.0-1) ...
```

As with Elasticsearch I started and enabled Logstash.

```
michelle@michelle-virtual-machine:~$ sudo systemctl start logstash
michelle@michelle-virtual-machine:~$ sudo systemctl enable logstash
Created symlink /etc/systemd/system/multi-user.target.wants/logstash.service → /lib/systemd/system/logstash.service.
```

The status of Logstash was then checked. As shown below it is working.

```
michelle6391@michelle6391-virtual-machine:~$ sudo systemctl status logstash
● logstash.service - logstash
     Loaded: loaded (/lib/systemd/system/logstash.service; enabled; vendor preset: enabled)
     Active: active (running) since Wed 2023-04-12 13:51:38 BST; 4s ago
   Main PID: 10131 (java)
      Tasks: 22 (limit: 4575)
     Memory: 302.9M
        CPU: 8.960s
     CGroup: /system.slice/logstash.service
             └─10131 /usr/share/logstash/jdk/bin/java -Xms1g -Xmx1g -Djava.awt.headless=true -Dfile.encoding=UTF-8 -Djruby.compile.i
nvokedynamic=true -XX:+HeapDumpOnOutOfMemoryError -Djava.security.egd=>

Apr 12 13:51:38 michelle6391-virtual-machine systemd[1]: Started logstash.
Apr 12 13:51:38 michelle6391-virtual-machine logstash[10131]: Using bundled JDK: /usr/share/logstash/jdk
lines 1-12/12 (END)
[2]+  Stopped                 sudo systemctl status logstash
```

The last thing to be installed is Kibana. This is done by using the command *'sudo apt-get install kibana'*.

```
michelle6391@michelle6391-virtual-machine:~$ sudo apt-get install kibana
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed
  kibana
0 to upgrade, 1 to newly install, 0 to remove and 124 not to upgrade.
Need to get 240 MB of archives.
After this operation, 629 MB of additional disk space will be used.
Get:1 https://artifacts.elastic.co/packages/8.x/apt stable/main amd64 kibana amd64 8.7.0 [240 MB]
Fetched 240 MB in 6s (41.0 MB/s)
Selecting previously unselected package kibana.
(Reading database ... 217432 files and directories currently installed.)
Preparing to unpack .../kibana_8.7.0_amd64.deb ...
Unpacking kibana (8.7.0) ...
Setting up kibana (8.7.0) ...
Creating kibana group... OK
Creating kibana user... OK
Created Kibana keystore in /etc/kibana/kibana.keystore
```

Once the installation was completed Kibana was started and enabled.

```
michelle6391@michelle6391-virtual-machine:~$ sudo systemctl start kibana
michelle6391@michelle6391-virtual-machine:~$ sudo systemctl enable kibana
Created symlink /etc/systemd/system/multi-user.target.wants/kibana.service → /lib/systemd/system/kibana.service.
```

The status was checked and showed that it is working.

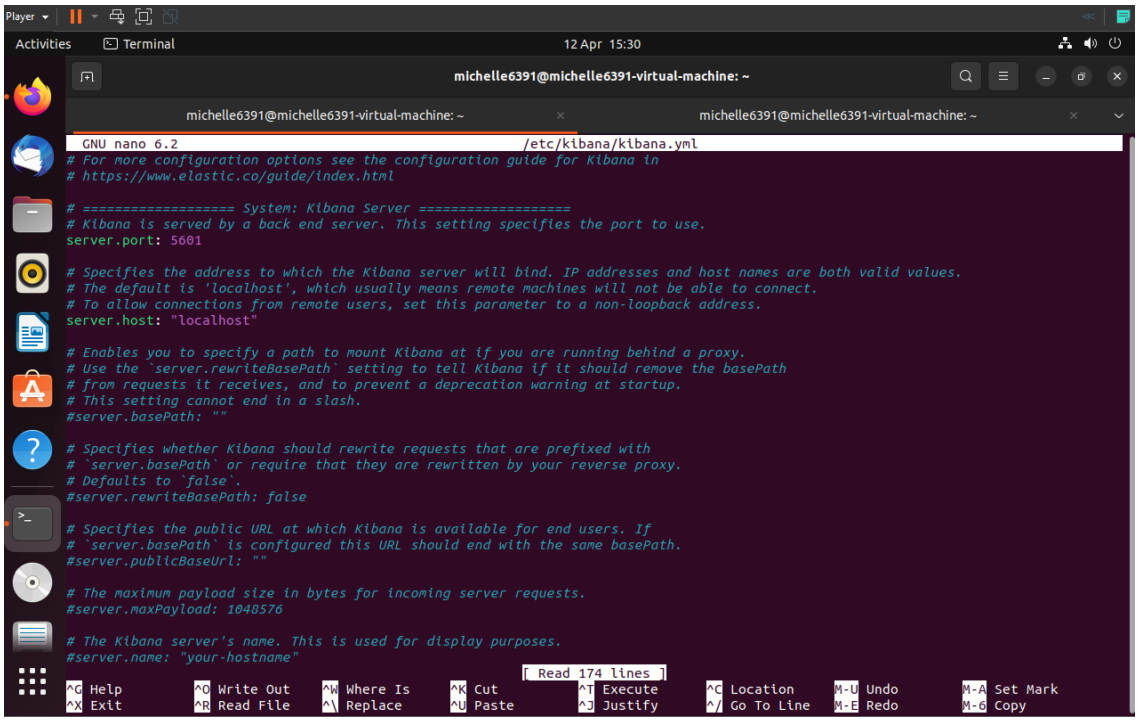
```
michelle6391@michelle6391-virtual-machine:~$ sudo systemctl status kibana
● kibana.service - Kibana
     Loaded: loaded (/lib/systemd/system/kibana.service; enabled; vendor preset: enabled)
     Active: active (running) since Wed 2023-04-12 14:00:12 BST; 1min 36s ago
       Docs: https://www.elastic.co
   Main PID: 13944 (node)
      Tasks: 11 (limit: 4575)
     Memory: 455.5M
        CPU: 16.691s
     CGroup: /system.slice/kibana.service
             └─13944 /usr/share/kibana/bin/../node/bin/node /usr/share/kibana/bin/../src/cli/dist

Apr 12 14:00:32 michelle6391-virtual-machine kibana[13944]: [2023-04-12T14:00:32.033+01:00][INFO ][plugins.ruleRegistry] Installed r
esources for index .alerts-observability.metrics.alerts
Apr 12 14:00:32 michelle6391-virtual-machine kibana[13944]: [2023-04-12T14:00:32.034+01:00][INFO ][plugins.ruleRegistry] Installed r
esources for index .alerts-observability.logs.alerts
Apr 12 14:00:32 michelle6391-virtual-machine kibana[13944]: [2023-04-12T14:00:32.035+01:00][INFO ][plugins.ruleRegistry] Installed r
esources for index .alerts-security.alerts
Apr 12 14:00:32 michelle6391-virtual-machine kibana[13944]: [2023-04-12T14:00:32.075+01:00][INFO ][plugins.fleet] Task Fleet-Usage-L
ogger-Task scheduled with interval 15m
Apr 12 14:00:32 michelle6391-virtual-machine kibana[13944]: [2023-04-12T14:00:32.335+01:00][INFO ][plugins.ruleRegistry] Installed r
esources for index .preview.alerts-security.alerts
Apr 12 14:00:33 michelle6391-virtual-machine kibana[13944]: [2023-04-12T14:00:33.942+01:00][INFO ][plugins.fleet] Running Fleet Usag
e telemetry send task
Apr 12 14:00:36 michelle6391-virtual-machine kibana[13944]: [2023-04-12T14:00:36.671+01:00][INFO ][plugins.securitySolution.endpoint
:metadata-check-transforms-task:0.0.1] no endpoint installation found
Apr 12 14:00:38 michelle6391-virtual-machine kibana[13944]: [2023-04-12T14:00:38.902+01:00][INFO ][status] Kibana is now available (
was degraded)
Apr 12 14:00:38 michelle6391-virtual-machine kibana[13944]: [2023-04-12T14:00:38.911+01:00][INFO ][plugins.reporting.store] Creating
 ILM policy for managing reporting indices: kibana-reporting
Apr 12 14:00:39 michelle6391-virtual-machine kibana[13944]: [2023-04-12T14:00:39.990+01:00][INFO ][plugins.fleet] Fleet Usage: {"age
nts_enabled":true,"agents":{"total_enrolled":0,"healthy":0,"unhealthy">
lines 1-21/21 (END)
[3]+  Stopped                 sudo systemctl status kibana
```

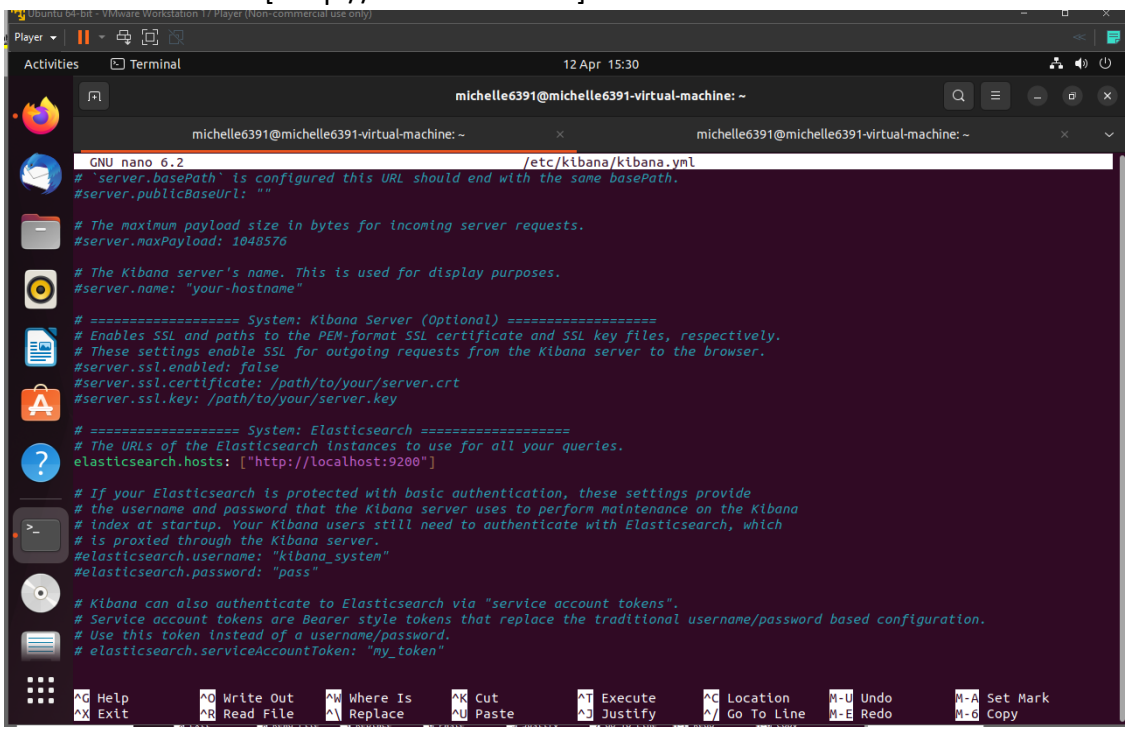
Next the configuration file was edited. To access the *'sudo nano /etc/kibana/kibana.yml'* was used.

```
michelle6391@michelle6391-virtual-machine:~$ sudo nano /etc/kibana/kibana.yml
```

server.port: 5601 was uncommented along with server.host: "localhost" which was changed to 0.0.0.0.



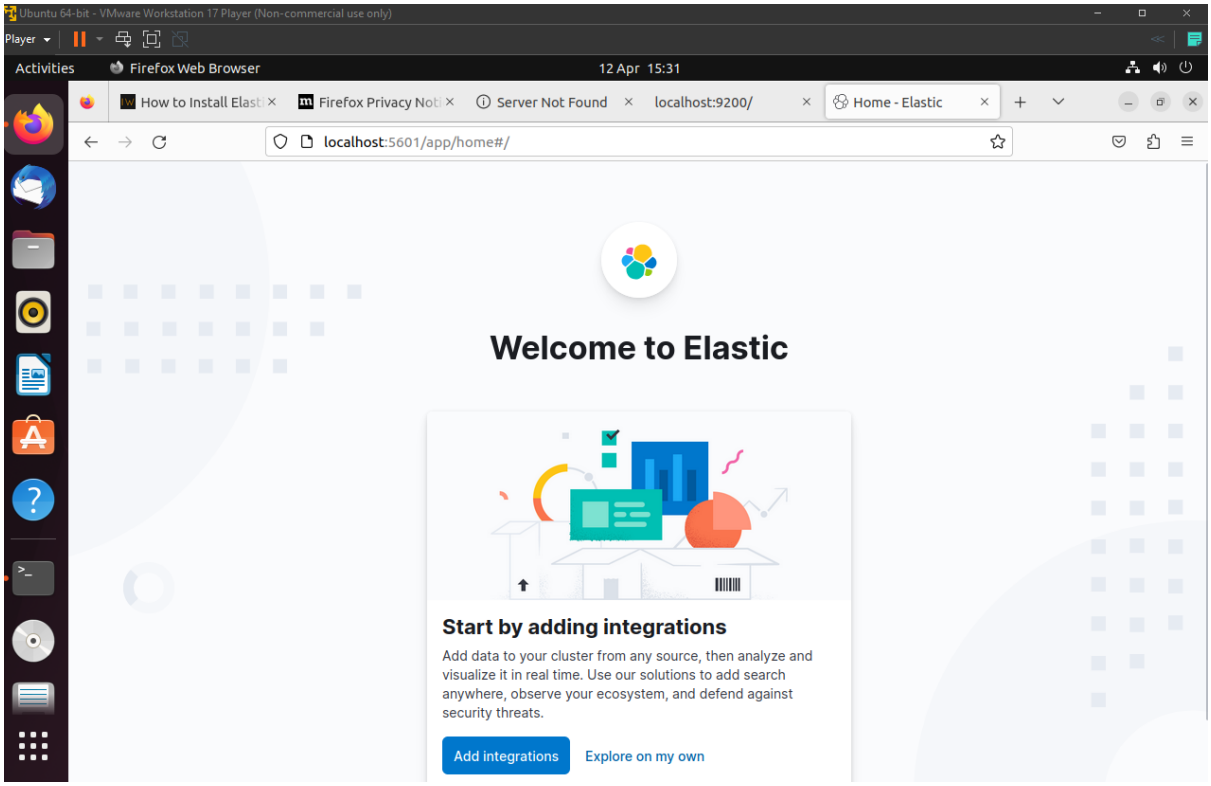
elasticsearch.hosts: ["http://localhost:9200"] was uncommented as well.



After the configurations were edited and saved kibana was restarted.

```
michelle6391@michelle6391-virtual-machine:~$ sudo systemctl restart kibana
```

Once that had finished *'http://localhost:560'* was put into a web browser. As shown below the Elasticsearch interface shows up meaning it is working.



(Mamidwar, S. (2022)).

Appendix B

MSc PROJECT (COMP11024)

PROJECT PROCESS DOCUMENTATION TEMPLATE

Student: Michelle Pantelouris **Supervisor: Graham Parsonage**

Meeting Number: 01 **Date/Time: 19th June 2023**

Agenda for meeting:

- **Discussion of project scope**
- **Discussion of project progress**

Discussion of agenda items:

What was discussed was what is expected to in the project and what should and should not be done and the minimum words expected for this project. Discussion surrounding the progress of the project in relation to timescales.

Summary of agreed action plan:

The agreed deadline for the project was the 3rd of July and what the minimum word was expected to be, which was 10000.

Notes: