EXPLORATORY ANALYSIS ON HONEYPOT DATA

Kuldip Gadapa
Master's in Data Analytics Engineering
George Mason University
Fairfax, Virginia
kgadapa@gmu.edu

Satya Sai Jayanth Devineni
Master's in Data Analytics Engineering
George Mason University
Fairfax, Virginia
sdevine3@gmu.edu

Yeshwanth Reddy Bommu
Master's in Data Analytics Engineering
George Mason University
Fairfax, Virginia
ybommu@gmu.edu

Vineel Vishwanth Busi
Master's in Data Analytics Engineering
George Mason University
Fairfax, Virginia
vbusi@gmu.edu

***Abstract* - This study aims to analyze and predict the AWS Honeypot data, which comprises of various country source, host, prototype, and time related data to know the exact attacking patterns of the hackers. Also, the paper draws some regression analysis about further attacks, that could be pre-assumed (known in advance). Several patterns like type of protocol used, locale, peak time of attack and top source address can be addressed specific to a country and target chosen. So, this study helps us to understand the critical information and apply higher protection to those sources that are most targeted. Using concepts of machine learning like regression and analysis techniques we retrieved several out of bound patterns and visualized to produce interesting outcomes from a huge observation. Our contribution can be analyzed about understanding threat intelligence and investigate major attacks in various countries. As a result, the outcomes can be used before preparing a secure infrastructure.**

Keywords - AWS Honeypot, regression analysis, machine learning, visuals, secure infrastructure, patterns.

I. INTRODUCTION

According to Online Digital industry experts, a Cyberattack is defined as a trial to obtain unauthorized and unaware access to one's personal and official assets (data, money) through network which is finally lead to misuse and destroy their assets. It is an offensive maneuver that effects computer information systems, computer networks, infrastructure, and personal computer devices. It can be implemented by nations, states, individuals, hackers, groups, or organizations. In our project, the focus is on Amazon Web Services (AWS) honeypot attack data and visualizing those cyberattacks [1]. Honeypot aim at cyber criminals by gathering their information and targets. The tend to imitate a system to be true and rectify the existing vulnerabilities. The behavior of the hackers, malicious IP addresses, network protocols and other vital data are collected by honeypots. Honeypots are reliable, difficult to detect, and fast. They are platform independent, compatible and have low cost. They are helpful as they can route or intercept network traffic at any point of time.

Honeypot is an effective way of tackling the impact of cyberattack on computer infrastructure. It mimics a system to cause a cyberattack. It can be used to identify, deviate, and know functionality of cyber criminals but it cannot stop attacks completely. They could be utilized as traps for cybercriminals as they think it is a legitimate target because it has system applications and data. AWS Honeypot resembles as Amazon network to outsiders which is maintained by Amazon IT teams. They monitor traffic for suspicious systems, track the attacks and operations to analyze what they want. Finally, they diagnose their security measures, firewall performances and perform necessary updates [1].

II. LITERATURE REVIEW

Amazon Web Service (AWS) honeypot is nothing but a trap point and security mechanism deliberated to tempt the attempted attack and if any source accesses the honeypot, the IP addresses will be recorded. Generally, a honeypot is the distraction for the attacker from their actual attack attempt and it will collect the information of the attacker by observing their request responses and the target hosts. Nowadays the cyber-attacks are

immeasurable and more sophisticated to the companies, individuals, industries, and government [2].

In 1986, the system admin of UC Berkeley named Clifford Stoll was involved in a process to track the charge for $0.75 of a Unix system at the lab. He used two honeypots to track the attacker. The actual target for the attacker was the nuclear defense secrets and later Clifford Stoll created a fictional department working on “Star Wars” to attract the attacker and he was later arrested. Since then honeypots became standard and the deception toolkit was launched in 1997. The honeynet project remained as the active security community resource [3]. To prevent the attacks from the attackers there should be some measures to be considered as the following by installing the antivirus software, regular updates and disabling all the unwanted services. If there is any poorly implemented honeypot then it is harmful as experienced attacker can attack the honeypot. In a large system or organizations, a honeypot can also be used as a decoy where it attracts the attacker. This leads the attacker to distract the attention and in this way the prevention of the attacks are possible. It can also prevent attacks to the systems are has sensitive data by alerting the presence of the attackers.

It is not the last technology as it is already implemented in the past. Honeypots capture what comes to them as it is different with other secured logs. The honeypot is divided into two types which are research honeypots and production honeypots. The production honeypots mainly aim to detect intrusions which provides security to the organizations. Our Research data is used to identify the vulnerabilities and improves the security. Generally, a honeypot aims to attract the attacker as it observes the activities. Generally, a honeypot is implemented in a large system environment as the data will be collected and it is used to study the attack patterns. It serves as an input to the signature training of a Detection System (Intrusion). A larger network is deployed with the Detection System (Intrusion) and Honeypot collaboration which will prevent the attacks that are caused by the attackers and distracted with the honeypot. Now this will able to react and alert the presence of those intrusions and it provides the reactive role [4].

There was a record of 451,581 attacks in a 6 months duration on AWS honeypots. AWS honeypot deals the attackers in a simple method by attracting the attackers with honeypot then the attacker will encounter the honeypot instead of our servers. The top 10 popular AWS data centers include Sydney, Sao Paulo, California, Mumbai, Frankfurt, London, Paris, Ireland, Singapore, and Ohio were placed with the cloud server honeypots by an enterprise security company. Most of the honeypot projects are open source and there is a honeypot project that has extension tools where it will also analyze the data that will be collected by the honeypot [4].

Types of Honeypots:

The following are the different types of honeypots:
A low interaction honeypot is a Virtual Machine that just represents the frequently common attacks registered. The low interaction honeypot is very simple to create and maintain it, but the attacker will be able to know that it is not a genuine platform.

A high interaction honeypot will use virtual machines to keep the systems isolated. In this type, several honeypots can run in a single physical machine. For several honeypots it makes easier to scale up. By using high interaction honeypot, the researchers will be able to learn the tools that the attackers will use to attack a data which is private data and confidential.

A pure honeypot is a physical server which is configured in way to attract the attackers and there is a special software which monitors the connection between both the network and the honeypot platform. There are also some disadvantages with these type of honeypots as it requires intensive manpower to configure and maintain it [3].

III. PROBLEM DEFINITION

The key problem is to identify malicious activity that organizations tend to fortify. A honeypot is used for such purpose that will deliberately configure with known vulnerabilities at a location to make more tempting or obvious target for attackers. As honeypot has no production data or do not participate in legitimate traffic on your network and that is how this will be recorded and identifies the cybercrime. The definition covers a diverse array of systems, from simple virtual machines which offer a few vulnerable systems to build fake networks spanning multiple servers. The goals of honeypot are diverse as they can be used as defense in depth to academic research [3].

The major problem is honeypot data prediction and it is hard for the cyber professionals in upgrading the honeypot tools and mechanisms. This research helps us to analyze and explore the cyber-attacks caused on honeypots and take necessary actions to avoid such practices. In this project we have drawn some insights and model fitting based on the hypothesis tests. The aim of this study is to handle the growth of cybercrime by studying the attacker's behaviors and observe the drawbacks in the hosts that are frequently attacked. As a result, the organizations

can avoid the hackers to break the system and save data loss.
A honeypot is always isolated and monitored depending on the importance of the resources like attacker's data, security mechanism, etc. It is like a trap that is set to observe, detect, mislead and divert an unauthorized user from his attempts on reaching an information system. Several companies must follow this bating procedure to safeguard the data from various threats and challenges like unidentified access during off time, using bad IP addresses for breaching, cross-site scripting attacks, distributed denial of service attacks and HTTP flood attacks [5].

Our main aim is to analyze and investigate the profiles of attackers, prone areas and the target. It helps to prepare a better defense by identifying important information and understand most attacks from a country and type of protocol used. Situations like cyber terrorist attacks on a nation data, research on data from honeypot on critical data, etc., The domain can also be identified to be protected prior to attack and rectify the method used by attackers.

Research honeypots allows close analysis of how hackers do their dirty work. The hacker's techniques on using infiltrate systems, escalate privileges, etc. are scrutinized. They are set up by security companies, academics, and government agencies to examine the threat landscape. However, once the honeypot is detected its value diminishes and it is used by spamming industries to identify spam-catching honeypots [3].

IV. DATASET DESCRIPTION

Our dataset contains the attack data of the Amazon web services (AWS) containing the following data which include datetime, host, src, proto, type, spt, dpt, srcstr, cc, country, locale, locale abbr, postal code, latitude and longitude. Using this dataset, the following can be visualized which includes the geolocation of the attacked places, presenting the top attackers, detecting attacks by the host, and highly active IP addresses. It has only [7].

The Dataset consists of 451,581 observations with 16 feature variables where there were over 88,000 missing values in overall column fields. There are columns like type, country code, locale, postal code , and X that cannot be used as they has very few observations and most of them have potential outliers.
Then the fields that were having high significance are chosen and all the null observations have been omitted. We have filtered and grouped the geographical coordinates i.e. Latitudes and longitudes. Using the lubridate library we have organised the date structure. We have organised the affected port sources country wise and factorised all the fields. We have then gathered all the frequent occurring IP addresses' source and added a column that gives the count of its occurrence in each country.

Nominal	**Description**
Datetime	Packet Arrival Date (YYYY-MM-DD)
host	Honeypot Server
src	Packet Source
proto	Packet Protocol Type
type	Packet Type
spt	Source Port
dpt	Destination Port
srcstr	Source IP Address
cc	Source Country Code
country	Source Country
locale	Source Location
localeabbr	Locale Abbreviation
postalcode	Postal Code

Ordinal	**Description**
latitude	Source Latitude
longitude	Source Longitude

This raw data will be then processed into a CSV file containing refined data about AWS honeypot. The dataset will have rows and 15 attributes.

V. METHODOLOGY & RESULTS:

From our dataset, the observed data measurements (interval, nominal, ratio, ordinal) of our attributes. The quantitative research will be performed which is tested and objective. It also has both independent and dependent variables. For our hypothesis, the following frequencies tables, cross tables, and chi squared tests are used. The bar charts and line graphs are used to view the data in a simple way.

The analysis is made from the Honeypot attack dataset to identify network attacks, security threats to understand and analyze some patterns, trends, and characteristics. During our research, there are some related articles and dominant cyberattacks are fetched. From the analysis, there are some observations on the reports that matched our attribute data and included some interesting observations.

A. Data Cleaning

They were several missing values in the original data set that is used in the analysis. The data is filtered like geo-locations, the co-ordinates data was not full like postal, spt and dpt and some data is omitted that were not much important for analysis. As the data was huge it took a lot of time to re-examine the attributes and contents. The challenging task was he had over 9,55,000 data point with NA values in a 4,51,500 data row count and some data were wrongly recorded. The time format is converted on an hourly basis by splitting into 4 parts of the day.

B. Data Regression and Data Analysis

Firstly, some quantitative co-relations are performed on the common factors like type, src addresses, proto count, etc. Some generalizable conclusions are performed. Some visuals like pie charts, interactive histograms, cross-tables, etc., and performed correlation analysis can be seen.

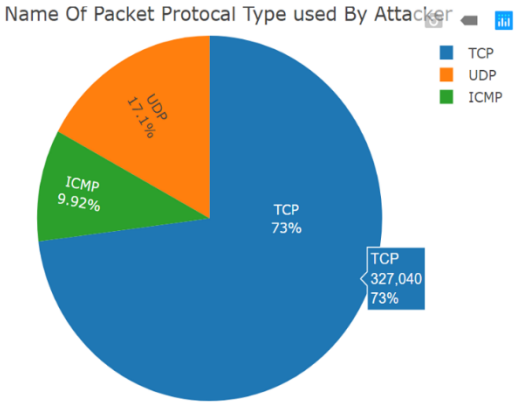

Fig.1 Name of Packet Protocol type used by attacker

From the above fig.1, it indicates the name of the packet protocol type used by the attacker. The three protocols used by the attackers are TCP, UDP and ICMP. Transmission control protocol (TCP) is used by most of the attackers as it registered 73% of the attackers following User Datagram Protocol (UDP) with 17.1% and Internet Control Message Protocol (ICMP) with 9.92% of the attackers.

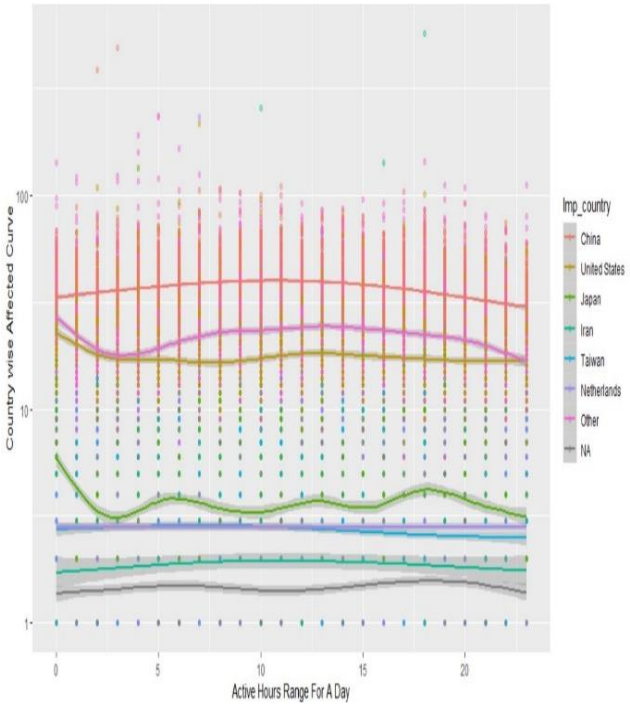

Fig.2 Active Hours Range for a day

The above plot shows that china has the most attack rate and is maintained throughout the day. From the plot, united states and Taiwan has most attacked cases in the morning and remains fluctuated.

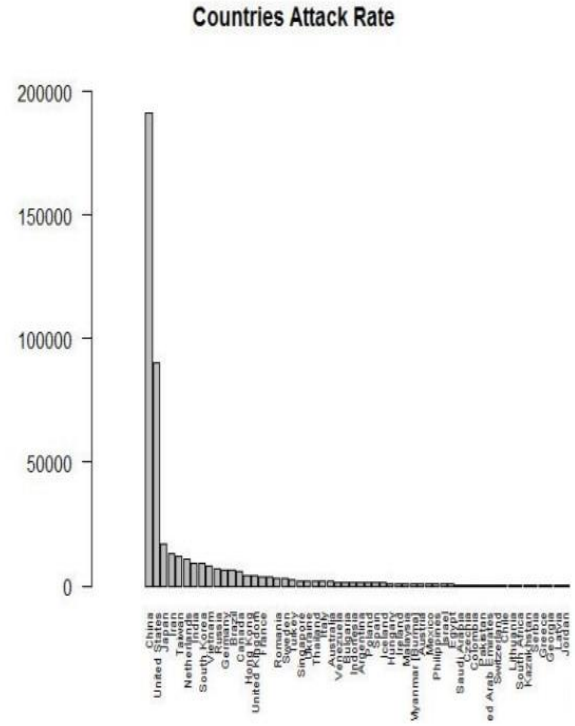

Fig.3 Countries Attack Rate

From the above plot, most of the false IP addresses and attacks are from China and united states.

C. Methods for correlation analyses

Pearson correlation ®, measures a linear dependence between two variables. It's also known as a parametric correlation test because it depends to the distribution of the data.
It can be used only when x and y from a normal distribution. The plot of y = f(x) is linear regression curve. Kendall tau and Spearman rho are rank-based correlation coefficients (non-parametric). Pearson correlation is the most practiced method [7].
Correlation Formula: In the formula below,

x and **y** are two vectors of length **n**
mxmx and mymy corresponds to the means of x and y, respectively.

Pearson Correlation Formula
$r=\sum(x-mx)(y-my)/\sum(x-mx)2\sum(y-my)2$
mxmx and mymy are the means of x and y variables.
The p-value (significance level) of the correlation can be determined:

1. by using the correlation coefficient table for the degrees of freedom: df=n−2, where n is the number of observations in x and y variables.
2. or by calculating the **t value** as follow:

$$t=r*\sqrt{n-2}/\sqrt{1-r2}$$

In the case 2) the corresponding p-value is determined using **t distribution table** for df=n−2. It the p-value is <5%, then the correlation between x and y is significant [8].

According to the chi square results, the p value is less than 0.05 and with 95% confidence.
There is a relationship between time of attack, protocol type, target host and attacker country.

sn$host	sn$prot TCP	UDP	ICMP	Row Total
groucho-eu	17405	4380	2169	23954
	0.003	9.686	18.199	
	0.039	0.010	0.005	
groucho-norcal	16421	4823	3322	24566
	113.281	67.394	320.773	
	0.036	0.011	0.007	
groucho-oregon	84179	7755	2142	94076
	3676.656	4566.163	5542.776	
	0.186	0.017	0.005	
groucho-sa	17112	4429	2775	24316
	17.074	8.247	54.337	
	0.038	0.010	0.006	
groucho-singapore	61024	6165	10962	78151
	319.951	4091.331	1326.191	
	0.135	0.014	0.024	
groucho-sydney	17177	4479	2800	24456
	19.320	10.595	57.391	
	0.038	0.010	0.006	
groucho-tokyo	72809	37285	16095	126189
	3874.440	10593.667	1019.573	
	0.161	0.083	0.036	
groucho-us-east	24663	4643	2473	31779
	108.343	146.398	146.836	
	0.055	0.010	0.005	
zeppo-norcal	17201	4820	2073	24094
	5.105	90.501	42.264	
	0.038	0.011	0.005	
Column Total	327991	78779	44811	451581

Fig.4 Cross Table for host and protocols

D. Methods for Regression and Clustering

1. Random Forest Regression- Learns a random forest* (an ensemble of decision trees) for regression. Each of the regression tree models is learned on a different set of rows (records) and/or a different set of columns (describing attributes), whereby the latter can also be a bit/byte/double vector descriptor (e.g. molecular fingerprint). The output model describes an ensemble of regression tree models and is applied in the corresponding predictor node using a simple mean of the individual predictions. Displays an interactive histogram view with different viewing options. The interactive histogram supports hilting and the changing of the x axis and aggregation column on the fly.

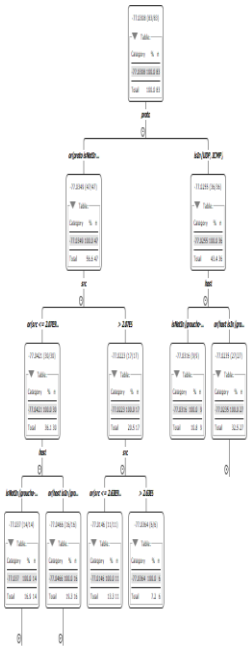

Fig.5 Classification Tree

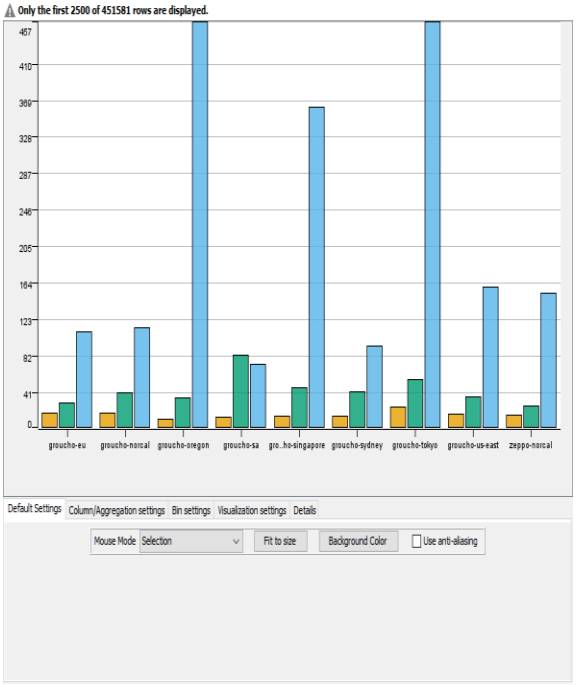

Fig.6 Interactive histogram for host differentiating by protocols

The above decision tree is the outcome of the applied regression using random forest method. It characterizes the dependency starting from the type of protocol used and source addresses. From the analysis, most of the country and host locations depend on the source of the attacker.
Learns a **random forest*** (an ensemble of decision trees) for regression. Each of the regression tree models is learned on a different set of rows (records) and/or a different set of columns (describing attributes), whereby the latter can also be a bit/byte/double vector descriptor (e.g. molecular fingerprint). The output model describes an ensemble of regression tree models and is applied in the corresponding predictor node using a simple mean of the individual predictions.

KNIME was used to perform K fuzzy means for clustering and displayed on scatter plot for host wise attacks. Random forest was also performed on the data and a interactive histogram was drawn along with a high predicted classification tree with a prediction of 85% for the best classification tree

The below graph is an interactive histogram that is produced from the random forest sample. This is performed using KNIME software.

2. FUZZY C-MEANS- The fuzzy c-means algorithm is a well-known unsupervised learning technique that can be used to reveal the underlying structure of the data. Fuzzy clustering allows each data point to belong to several clusters, with a degree of membership to each one. Make sure that the input data is normalized to obtain better clustering results. The list of attributes to use can be set in the second tab of the dialog. The first output data table provides the original data table with the cluster memberships to each cluster. The second data table provides the values of the cluster prototypes. Additionally, it is possible to induce a noise cluster, to detect noise in the dataset, based on the approach from R. N. Dave: 'Characterization and detection of noise in clustering'. Creates a scatterplot of two selectable attributes. Then each datapoint is displayed as a dot at its corresponding place, dependent on its values of the selected attributes. The dots are displayed in the color defined by the Color Manager, the size defined by the Size Manager, and the shape defined by the Shape Manager. The **fuzzy c-means** algorithm is a well-known unsupervised learning technique that can be used to reveal the underlying structure of the data. Fuzzy clustering allows each data point to belong to several clusters, with a degree of membership to each one.

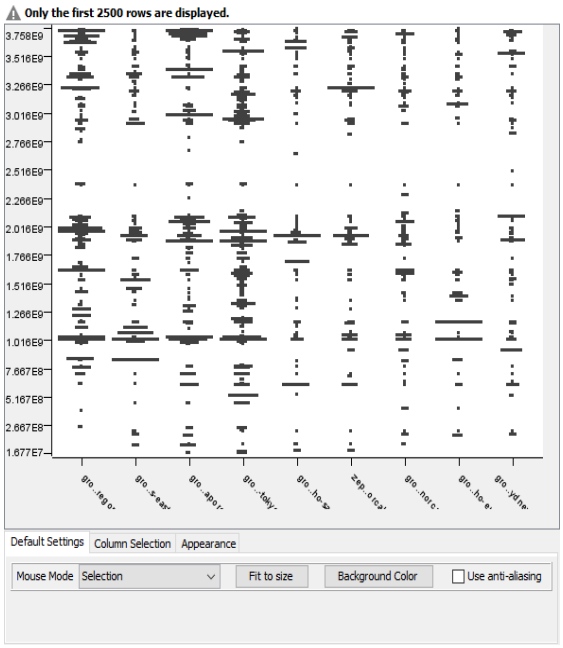

Fig.7 Scatter Plot

KNIME, the Konstanz Information Miner, is a free, has integration platform for reporting and open source data analytics tool. The concept of modular data pipelining is adapted via ML and data mining components. It uses Nodes as GUI that blend several data sources, preprocessing (Extraction, Transforming and Loading) for mining, modelling, analysis and visualization with minimal programming. KNIME, open for innovation can be easily understood and downloaded to build workflows, projects and develop insights [9].

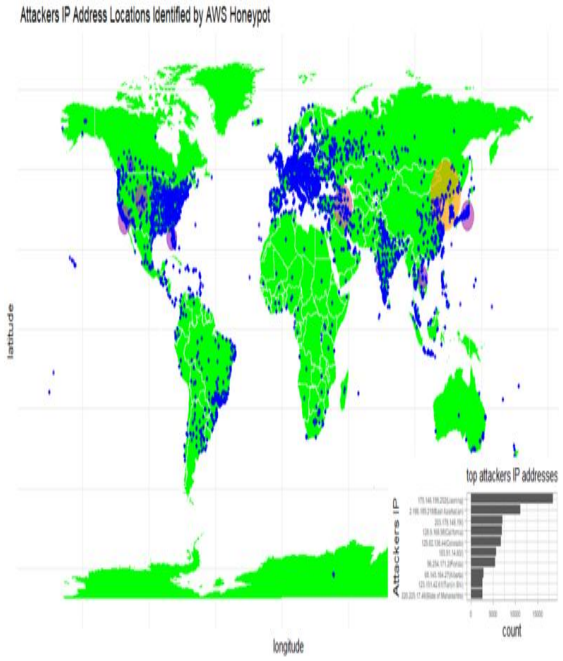

Fig.8 Attackers IP Address Locations Identified by AWS Honeypot

The above plot gives the information of the Attackers IP address locations identified by the AWS Honeypot data. It also includes the top attackers IP address with respect to the count. The blue colour indicates the attackers IP address around the world. The observed IP addresses gives the information that each IP address is used several times on the host location.

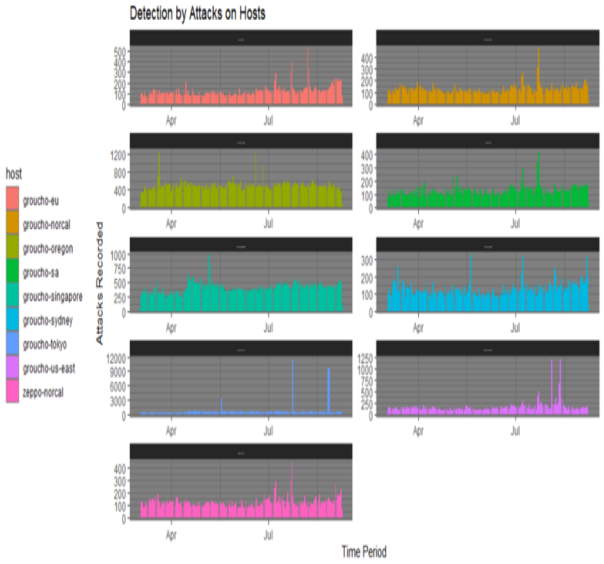

Fig.9 Detection by Attacks on Hosts

From the above plot, the detection of the attacks on the hosts can be observed in the following time for each host location. The highest attacked host location is Tokyo where it has witnessed around 12000 attack records in the month of July. All the host locations have observed constant attack records.

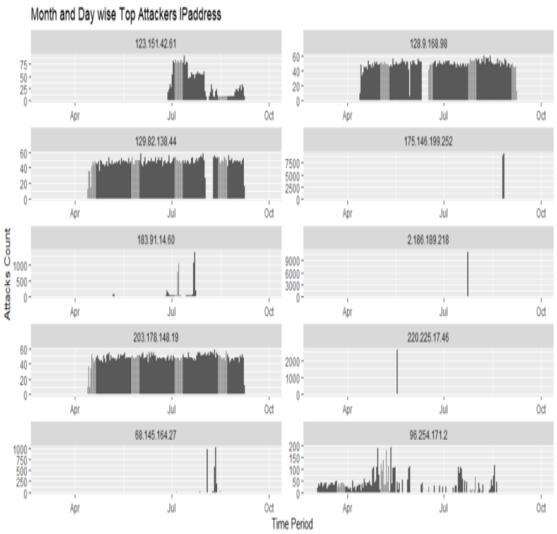

Fig.10 Month and Day wise Top Attackers IP Address

From the Fig.10, the plot gives detail records of the top attackers IP address used for attacking the honeypots. Some attackers use the IP address only once for the attack and the others use the same IP address for multiple attacks.

```
> topattacks
           srcstr       country longitude latitude              locale count
1  175.146.199.252         China  123.4328  41.7922           Liaoning 18472
2    2.186.189.218          Iran   46.2919  38.0800      East Azarbaijan 11116
3   203.178.148.19         Japan  139.6900  35.6900                      7086
4    128.9.168.98 United States -118.4351  33.9777           California  7046
5   129.82.138.44 United States -105.4648  40.4555             Colorado  6772
6    183.91.14.60       Vietnam  106.0000  16.0000                      5676
7    96.254.171.2 United States  -82.7048  28.3031              Florida  5413
8   68.145.164.27        Canada -114.0833  51.0833              Alberta  2834
9   123.151.42.61         China  117.1767  39.1422         Tianjin Shi  2663
10  220.225.17.46         India   72.8258  18.9750 State of Maharashtra  2605
                        fullIP
1          175.146.199.252(Liaoning)
2      2.186.189.218(East Azarbaijan)
3                  203.178.148.19()
4           128.9.168.98(California)
5            129.82.138.44(Colorado)
6                   183.91.14.60()
7             96.254.171.2(Florida)
8            68.145.164.27(Alberta)
9          123.151.42.61(Tianjin Shi)
10 220.225.17.46(State of Maharashtra)
```

Fig.11 Top Attacks on Honeypots

The above Fig.11 explains the top attacks which includes the country, longitude, latitude, locale, IP address and count. The highest attacked country is china (Liaoning) with a count of 18472 and the least attacked country is India (Maharashtra) with a count of 2605.

```
> head(Uniquehost)
[1] "groucho-tokyo"     "groucho-singapore" "groucho-sydney"    "groucho-oregon"
[5] "groucho-us-east"   "groucho-sa"
> sample(uniqueScr,20)
 [1] "United States-California-1211394072-8/6/13 11:40"
 [2] "United States-Texas-2927204522-8/14/13 6:53"
 [3] "Myanmar [Burma]-Yangon Region-2063099406-5/10/13 10:17"
 [4] "United States-California-2382653301-4/4/13 15:15"
 [5] "Hungary-Pest megye-2989517593-7/19/13 12:51"
 [6] "China-Zhejiang Sheng-2061508829-3/17/13 3:26"
 [7] "United States-California-2148116578-5/18/13 15:19"
 [8] "Russia--3273845291-3/21/13 10:51"
 [9] "Poland--3574725442-6/17/13 13:01"
[10] "China-Guangdong-1959653494-3/18/13 1:41"
[11] "Taiwan-Taipei-1878140219-6/5/13 6:23"
[12] "China-Hubei-2002896033-8/17/13 2:01"
[13] "China-Tianjin Shi-2073504317-7/10/13 15:57"
[14] "Thailand--1847780954-4/29/13 6:00"
[15] "China-Beijing Shi-3689613428-6/14/13 7:41"
[16] "China-Heilongjiang Sheng-1885683791-5/18/13 16:25"
[17] "Turkey-Izmir-1492819901-4/7/13 23:10"
[18] "China-Hebei-3720417886-8/22/13 0:49"
[19] "India-State of Karnataka-1975526721-6/5/13 10:50"
[20] "China-Shandong Sheng-2078826646-3/7/13 5:28"
```

Fig.12 Most Attacked Honeypots Location

From the Fig.12, the observations indicate the honeypots that are mostly attacked in these countries around the world and these are the most frequently attacked countries as well. The topmost attacked host locations are Tokyo, Singapore, Sydney, Oregon, and Saudi Arabia.

VI. DISCUSSION

On an overall exploratory data analysis, we found these Insights: Most of the attackers tend to attack during the night-time while the monitoring is idle. The locale belongs to the U.S., Middle East, and some parts of Asian countries. It is understood in the map data and the high amount of cybercrime on country wise analysis is observed. On overall analysis most of the parts that have amazon firms have most crime rate recorded.

The data regarding what is the source of aim of attackers is not specified as they generally tend to cross the data security barriers to get valuable data of the organisation. The top attackers source IP addresses are recognised and those are mentioned. It is observed that most of the intruders get active during the night times and afternoon. It is observed that there are some common IP addresses having common series of attacks and the country of attack is specified. It is observed that 71% of the attackers used TCP protocol and the top attackers use the UDP protocol for security breaches. The summary of the processed data gives the import factors like host significance, time, country analysis, and count of periodic attacks. The hypothesis tests give us the relation between the legends in the data. The time related analysis and location coordinates gives us the attackers timeline, spatial analysis and the count factor gives us the most attacked occurrences. It is seen that the U.S, China, Japan, and India are the most cyberattacks source locations specific to AWS host data.

VII. CONCLUSION

The conclusion for this project is to deduce the data of the honeypot that RANDOM FOREST with classification tree is the best fit with 85% accuracy and LM gave around 60%. Some methodologies are applied and used chi squared test. Also, random forest and fuzz-c-means are performed using KNIME. The obtained information from the results gives valuable information which includes the following as observed that there are some protocols which are used effectively in some systems and these systems have to be configured accordingly and maintained. Depending on the time of day and protocol the attacking countries vary with these factors, the IP addresses from those countries are monitored. Some prevention measures should be taken for the protocols which are being attacked on

specific days and months. These are the essential measures observed from our results. The main objective is to collect the data and analyze them to learn the information about attackers' patterns and their methods, tools, and objectives. The other important issue is the Analysis of the data from different the honeypots [10].

Generally, one can consider an intelligent approach to mutate or change a honeypot from a detectable back to undetectable form [11].

In the future work, an approach we employ an array of the anomaly detectors should be monitored and classify all traffic to the protected network. The traffic deemed anomalous is processed by a shadow honeypot as we are trying to protect a protected instrumented instance of the application in the organization [12].

If an administrator is responsible to keep the honeypot systems secure as the attacker would not have been possible, therefore as the administrator will be jointly responsible for any attack which has occurred [13].

References

[1] "AWS Honeypot Data: Visualizing The Threat of Cyberattacks," 2020. [Online]. Available: https://www.sisense.com/whitepapers/gofigure-aws-honeypot-data-visualizing-the-threat-of-cyberattacks/.

[2] "Cybersecurity," 2018. [Online]. Available: https://wiki.smu.edu.sg/1718t3isss608/Group01_Report.

[3] "What is a honeypot? A trap for catching hackers in the act," 2019. [Online]. Available: https://www.csoonline.com/article/3384702/what-is-a-honeypot-a-trap-for-catching-hackers-in-the-act.html.

[4] "APPLYING DATA-MINING TECHNIQUES IN HONEYPOT ANALYSIS," APPLYING DATA-MINING TECHNIQUES IN HONEYPOT ANALYSIS , 2015. [Online]. Available: file:///C:/Users/jayanth/Downloads/Veerasamy_2006.pdf.

[5] "Cybercriminals attack cloud server honeypot in 52 seconds," 2019. [Online]. Available: https://www.cio.com/article/3515424/cybercriminals-attack-cloud-server-honeypot-in-52-seconds.html.

[6] "eVitamins Case Study," 2017. [Online]. Available: https://aws.amazon.com/solutions/case-studies/evitamins/.

[7] "AWS Honeypot Attack Data," 2018. [Online]. Available: https://www.kaggle.com/casimian2000/aws-honeypot-attack-data.

[8] "Correlation Test Between Two Variables in R," 2020. [Online]. Available: http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r.

[9] "End to End DATA sCIENCE," 2020. [Online]. Available: https://www.knime.com/.

[10] "Data Collection and Data Analysis in Honeypots and Honeynets," Data Collection and Data Analysis in Honeypots and Honeynets, 2015. [Online]. Available: http://spi.unob.cz/papers/2015/2015-19.pdf.

[11] "Data Science Central," Data Science Central, 2016. [Online]. Available: https://www.datasciencecentral.com/profiles/blogs/honeypot-turing-test.

[12] "Detecting Targeted Attacks Using Shadow Honeypots," Detecting Targeted Attacks Using Shadow Honeypots, 2005. [Online]. Available: https://www.usenix.org/legacy/event/sec05/tech/full_papers/anagnostakis/anagnostakis_html/replay.html.

[13] "A Survey on Honeypot Software and Data Analysis," A Survey on Honeypot Software and Data Analysis, 2016. [Online]. Available: http://researchspace.csir.co.za/dspace/bitstream/handle/10204/3127/Veerasamy_2006.pdf;sequence=1