A Project Report on

AI BASED SOAR

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Engineering

in

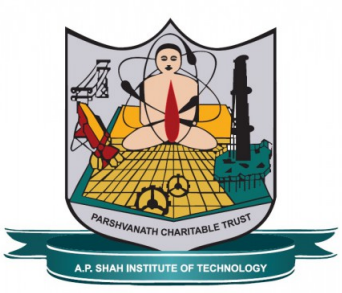
Information Technology

by

Rahul Vast(17104042)
Shruti Sawant(18204001)
Aishwarya Thorbole(18204002)

Under the Guidance of

Mr. Vishal Badgujar



Department of Information Technology
NBA Accredited
A.P. Shah Institute of Technology
G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615
UNIVERSITY OF MUMBAI

Academic Year 2020-2021

Approval Sheet

This Project Report entitled ***"A.I. based SOAR"*** Submitted by ***"Aishwarya Thorbole" (18204002)*** is approved for the partial fulfillment of the requirement for the award of the degree of ***Bachelor of Engineering*** in ***Information Technology*** from ***University of Mumbai***.

(Mr. Vishal Badgujar)
Guide

Prof. Kiran Deshpande
Head Department of Information Technology

Place:A.P.Shah Institute of Technology, Thane
Date:

CERTIFICATE

This is to certify that the project entitled ***"AI BASED SOAR"*** submitted by ***"Aishwarya Thorbole" (18204002)*** for the partial fulfillment of the requirement for award of a degree ***Bachelor of Engineering*** in ***Information Technology***,to the University of Mumbai,is a bonafide work carried out during academic year 2020-2021.

(Mr. Vishal Badgujar)
Guide

Prof. Kiran Deshpande
Head Department of Information Technology

Dr. Uttam D.Kolekar
Principal

External Examiner(s)

1.

2.

Place:A.P.Shah Institute of Technology, Thane
Date:

Acknowledgement

We have great pleasure in presenting the report on **AI BASED SOAR** We take this opportunity to express our sincere thanks towards our guide **Mr.Vishal Badgujar** & Co-Guide **-** Department of IT, APSIT thane for providing the technical guidelines and suggestions regarding line of work. We would like to express our gratitude towards his constant encouragement, support and guidance through the development of project.

We thank **Prof.Kiran B.Deshpande** Head of Department,IT, APSIT for his encouragement during progress meeting and providing guidelines to write this report.

We thank **Prof.Vishal S.Badgujar** BE project co-ordinator, Department of IT, APSIT for being encouraging throughout the course and for guidance.

We also thank the entire staff of APSIT for their invaluable help rendered during the course of this work. We wish to express our deep gratitude towards all our colleagues of APSIT for their encouragement.

Aishwarya Thorbole:
18204002:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

_____

_____

(Aishwarya Thorbole and 18204002)

Date:

Abstract

Cyber security is becoming very crucial in today's world where technology is now not limited to just computers, smart phones,etc.It is slowly entering into things that are used on a daily basis like home appliances,automobiles,etc.Thus, opening a new door for people with wrong intent.With the increase in speed of technology dealing with such issues also requires quick response from security people.Thus, dealing with a huge variety of devices quickly will require some extent of automation in this field. Generating threat intelligence automatically and also including those which are multilingual will also add plus point to prevent well known major attacks. Here we are proposing an AI based SOAR system in which the data from various sources like firewalls, IDS, etc.is collected with individual event profiling using a deep-learning detection method.For this the very first step is that the collected data from different sources will be converted into a standardized format i.e. to categorize the data collected from different sources.For standardized format Here our system finds out about the true positive alert for which the appropriate/ needful steps will be taken such as the generation of Indicators of Compromise report and the additional evidence with the help of Security Information and Event Management system. The security alerts will be notified to the security teams with the degree of threat.

Contents

List of Figures

List of Tables

List of Abbreviations

IDS:	Intrusion Detection System
IPS:	Intrution Prevention System
AI:	Artificial Intelligence
SOAR:	Security Orchestation Automation And Response
SIEM:	Security Information and Event Management
TI:	Threat Intelligence
STIX:	Structured Threat Information eXpression

Chapter 1

Introduction

Technology is growing day by day. In this modern world, it has now become one of the integral parts of our body. It has slowly but steadily started to became everywhere around us.Even in day-to-day life we can see various examples of it.Our smart phones, smart watches,tabs, etc. are there beeping and glowing all around us.There are fitness trackers which keeps track of our fitness and our health conditions.There are even beds that can track sleep patterns.Automated home system that track the patterns of day to day life of the people living in it.This is just to name few.Technology has even entered in our household things and in our home appliances. The main thing to note here is all these technologies are producing different types of data.Health related data, sensitive data, confidential data, etc.No matter what kind of data it produces it is important to protect all of these data.If not then Such kind of data can actually make one vulnerable to harmful intentions of other people.That is the area where Cyber Security comes into the picture.Cyber Security is the field which aims to protect users all around the world to protect from Cyber attacks. This seems to be simple when we hear it first.An attack will happen and cyber security people will prevent it and everything will be fine.Well it's not that easy.A cyber security person always needs to be few steps ahead of attacker.They should be quick when deciding what to do and should not get panic themselves.Because they can often find something useful which can trace back to the attacker,it also can lead to information such as why the attack happened,how the attack happened, what were the intentions of hacker/attacker when performing such attack.Often such information on attack help us to find the answer of other questions as well such as what will be the impact of such attacks on the organization,how to avoid such attacks in future, where are we weak in terms of security and what steps needed to carry out to fill up those loop holes.As today's world has well understood the importance of Cyber Security in their organization and there are various measures to deal with such attacks in future.There are also various tools designed by organizations to protect themselves. Such as firewalls,IPS, IDS, Antivirus, Anti malware,SIEM systems, etc.There are also practice of keeping information about the attacks happened on organization previously which is generally known as Threat Intelligence.Threat intelligence often help organizations to deal with the attacks previously happened and now they might have developed the prevention strategy for the same attack and thus attacker fails.Sharing such information with other organization can help them too to get over those attacks but it does not often happen due to privacy reason of a particular organization.

Chapter 2

Literature Review

Consistent tuning and daily updates according to modern threats is requiring human assistance.In literature[4], the author suggested a new threat intelligence technique which will be evaluated by analyzing honey pot log data that will identify behavior of attackers and by finding attack patterns.They deployed their honeypot on AWS cloud for collecting incident related data. This log data is then analyzed using ELK (Elastic search, Logstash, Kibana) stack. These systems generate alerts and prevent Cyber Attacks based on the learned attack patterns.The potential drawback of this Model is that it requires data related to same incident in huge amount.Most of the data collected is also similar to each other which can decrease the performance of model on new data.In literature [5], the author suggested a mechanism to represent raw cyber threat-data in Structured Threat Information Expression format in an automated manner.The method also takes care of privacy preservation.The Standard Format required for an organization to share these data with other Organizations is provided with these systems.
They Improve the Privacy of The Data Because Sensitive information is removed as a CTI (Cyber Threat Intelligence) is generated.This helps the organizations to understand the threat and also make it ready for the advance analytics.This data can also be shared on a Threat Intelligence sharing platform.Thus,the system aims to bring complete automation of the security to deal with modern day attacks efficiently.In literature [6], the authors suggested to create a neutral network that takes in threat intelligence available indifferent languages using which it translates that it in desired languages and thus help in making available threat intelligence in different languages.This proposed system uses Russian And English word embeddings created from cyber security data.It uses an LSTM based neural machine translation architecture.
It also uses encoder-decoder architecture which maps Russian words to their English words. Their results show that their system easily out performs other third-party translation engines as well as successfully detect cyber security terms better.The system is able to run independently in secluded operational settings.The System requirement of a cyber security rich data to train the model is quite high though.The system aims at making threat intelligence data available in different languages to be available globally in their natives.

Chapter 3

Project Design

SOAR stands for Security Orchestration, Automation, and Response.The main focus of this technology is to automate various security processes like network security audit,privileged password management and coordination and execution of tools between various tools and security groups. It does this by using various playbooks that has required steps need to be performed developed by experts themselves.

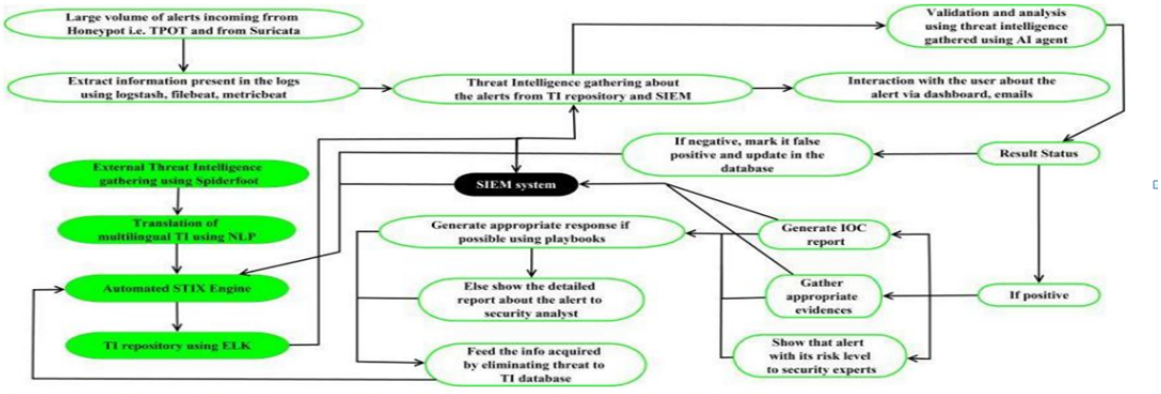


Figure 3.1: Proposed System Architecture - Artificial Intelligence based Security, Orchestration, Automation and Response

Our proposed system preserves the above-mentioned vital features and adding a flavor of AI on the SOAR system.The workflow of our system starts with two parts i.e. the honeypot and the common front- line defenses such as firewall, Intrusion Detection System, Intrusion Prevention System, etc. For honeypot we used open source multi honeypot TPOT and is deployed based on [4] and for frontline defenses we used Suricata Network Threat Detection Engine. While honeypot helps us in analyzing the attack pattern trends going on and in-turn helps AI to be familiarized with current trends and Suricata helps in identifying the basic required data such as Common Vulnerabilities and Exposures (CVE) the attacker is trying to utilized as well as the analysis of the commands and payloads the attacker is using. Furthermore for parsing and visualization of logs , reports, etc. is done using ELK and its beats by using methods described in [4]. Apart from this, for URL checking we are using the framework based on [1]. After this, all the Threat Intelligence will be gathered about the

alert with the help of Threat Intelligence Repository. This repository stores all the information about known threats. Now, if the threat is new i.e. it is not available in the repository then this data will be gathered from open source threat intelligence using an open source tool called Spiderfoot/ Virus Total API. Spiderfoot/Virus Total API will gather threat intelligence required for the analysis of alert. This will be passed through NLP based Translation of multilingual threat intelligence based on [6] only if the threat intelligence gathered is in different language(currently only supports English, Russian and German). Furthermore the Automated Standardization and Privacy Preservation Engine which is based on [2] and [5] will help us in converting this data in proper standardized format and will also take care of the privacy related issues before feeding it to SIEM which is based on [3].
SIEM will aggregate all the Threat Intelligence acquired about the alert. After this step, this data will undergoes various analysis to find out if there is any pattern hidden in all this data and based on this profiling of event is done. If any suspicious activity is detected, alert will be generated by SIEM to our AI agent. When the attack happens SIEM will try to search it in the repository first so that it will get all the detailed threat intelligence needed. This in-turn helps AI to mitigate the attack more efficiently. Now all this collected information about the alert will support AI in making its decision about the alert that whether the alert is true positive or false positive. If false positive then the related information will be passed to the Threat Intelligence Repository for future purpose. But if the alert is true positive then the Indicators of Compromise report generation, Evidence about the attack, the level of attack, etc information about the attack will be shown to the security analyst. Further, the playbooks, run books, etc present for the threat will also be used to speed up the response process. Else just the information about the threat will be shown to the analyst. Finally, at the end the threat related data about this true positive threat will also get feed in the repository. This will help the system in future for similar attacks as well as in SIEM's attack pattern analysing and event profiling [3]. Further, it is also possible to open source the data from threat intelligence repository as privacy issues will be already taken care by Automated Standardization and Privacy Preservation Engine

Chapter 4

Project Implementation

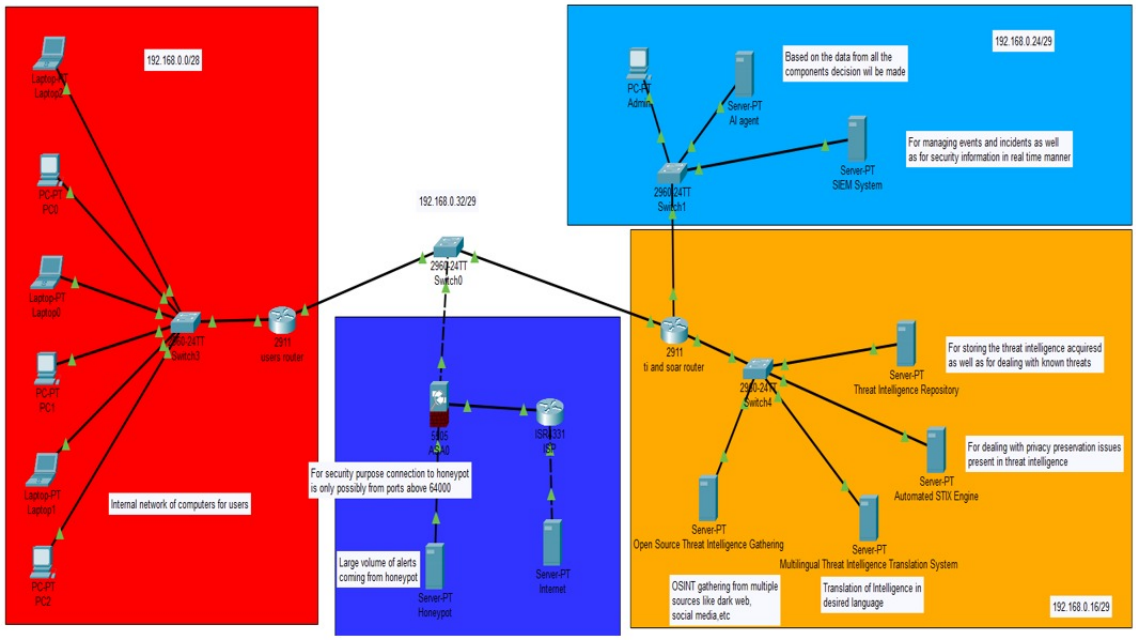


Figure 4.1: Cisco Packet Tracer Topology for AI SOAR.

The above diagram depicts the Cisco Packet Tracker Topology for AI based SOAR architecture,where in the different depicts the SOAR architecture.Here the red block represents the internal network for computer users,where the router is connected to the honeypot system accepting ports about 6400 connecting to honeypot server and ISP representing navy blue block.Also the sky blue block contains the SIEM System and followed by yellow block containing the threat intelligence system.

Figure 4.2: Being depicted by using histogram in which has Attacker SRC (Source) IP Reputation,attacks by honeypot where attacks Dioneaea,Cowrie,Tanner etc.

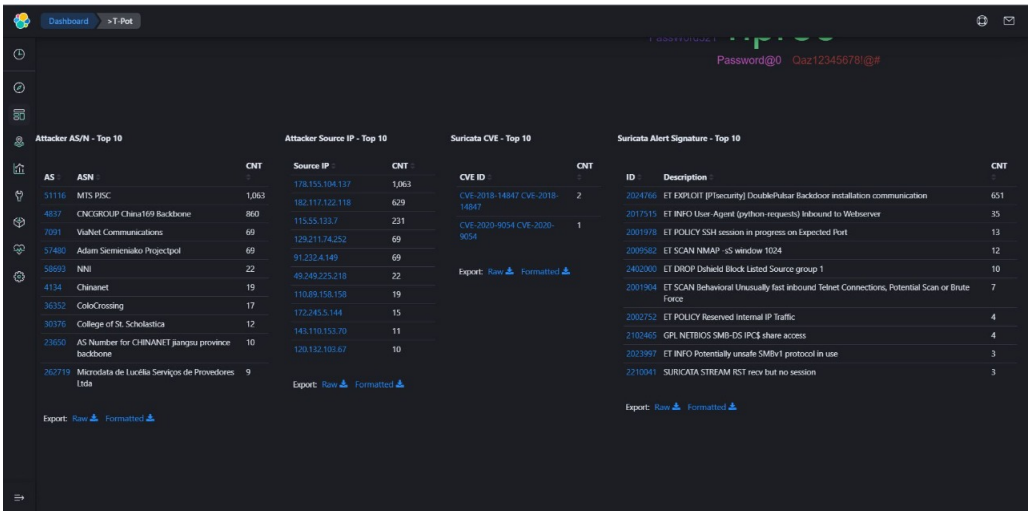


Figure 4.3: The above figure,depicts the different attacks there source IPs and its count.

Figure 4.4: Here the attacks are categorized by country and port where the different country's attack count can be seen

Figure 4.5: Above is the Honeypot Attack Map with the satellite view of countries with most attack,the red dot has the maximum attacks as seen.

Figure 4.6: The Above is the summarized in honeypot attack Bar graph,honeypot attacks by histogram,attacks by country histogram and so on.

Chapter 5

Testing

Testing being an important feature and to identify the project workflow works accordingly , here we represent the testing modules in accordance to the Wanna Cry Run model and Vulnerability response model .


```
'popular_threat_classification': {'popular_threat_category': [['trojan',
                                                               22],
                                                              ['ransomware',
                                                               17]],
                                  'popular_threat_name': [['wannacry',
                                                           9],
                                                          ['wanna',
                                                           8],
                                                          ['cztf',
                                                           4]],
                                  'suggested_threat_label': 'trojan.wannacry/wanna'},
'reputation': -99,
'sandbox_verdicts': {'Lastline': {'category': 'malicious',
                                  'malware_classification': ['MALWARE',
                                                             'TROJAN'],
                                  'sandbox_name': 'Lastline'}},
'sha1': 'f597a1cc16d42b7f02e077696e067cd3030a06d9',
'sha256': 'c05e2dab77349cd639aa837e7e121710b8a0718d8fc93fb4cc6458ae90e5c597',
'size': 5267459,
'ssdeep': '98304:+DqPoBhz1aRxcSUDk36SAEdhvxWa9P593R8yAVp2H:+DqPe1Cxcxk3ZAEUadzR8yc4H',
'tags': ['overlay',
         'exploit',
         'cve-2017-0147',
         'armadillo',
         'via-tor',
         'pedll'],
'times_submitted': 2507,
'tlsh': 'T192363394622CB2FCF0440EB44463896BB7B33C6967BA5E1F8BC086670D43B5BAFD0641',
'total_votes': {'harmless': 1, 'malicious': 6},
'trid': [{'file_type': 'Win32 Executable MS Visual C++ '
                       '(generic)',
          'probability': 38.8},
         {'file_type': 'Microsoft Visual C++ compiled '
                       'executable (generic)',
          'probability': 20.5},
         {'file_type': 'Win64 Executable (generic)',
          'probability': 13.0},
         {'file_type': 'Win32 Dynamic Link Library (generic)',
          'probability': 8.1},
         {'file_type': 'Win16 NE executable (generic)',
          'probability': 6.2}],
'type_description': 'Win32 DLL',
'type_extension': 'dll',
'type_tag': 'pedll',
```


Figure 5.1: Threat intelligence of Wanna Cry Run Model-includes the type of malware,size,its probability to threaten the system, if the malware is malicious or not and so on..

Figure 5.2: Showcases this the TI which contains the entities such as its prerequisites name,assigners,related-weaknesses,solution provided ,summary of the attack and so on.

Chapter 6

Result

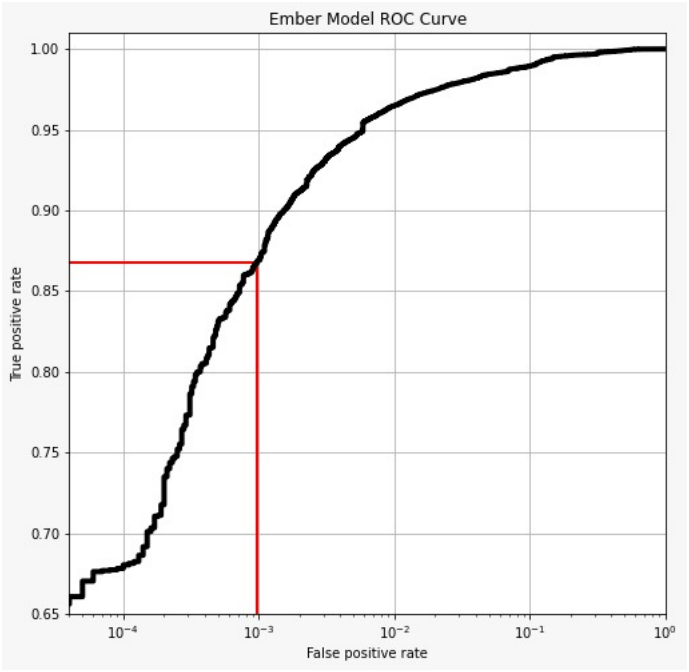


Figure 6.1: ROC Curve- the ROC Curve is represented using Ember Model which depicts the curve plot between to find out the True positive and False positive rate in terms of Alerts

The ROC Curve here represents the performance of our Malware Detection Model on thevsamples passed. Here, the true positives is depicting the recall score of our model and false positives represents the precision of our model.

Network Devices
Core Layer devices -
1. Router - PT : Router connected with a Serial DTE cable and Cooper Cross-Over to the other Router's and Servers.
Distribution Layer Device -
1. Router-PT - This Router is connected just outside of every LAN provided in the Network. 2. Switches - We have 4 Switches all of them 2950-24, for routers , switches , servers and PC's. 3. Server - We have an Inside Server and an Outside Host with a server belongs to the ISP.

Access Layer Devices -
1. Laptops and Computers - These Workstation are the main thing and we all see in the AI Based SOAR System.

Devices	**Count**	**Cost**
Router	3	6000
Switch	4	4500
Computers	6	200000
Server pt	9	152000
Total	22	3625000

Table 6.1: Bill of Materials-The table depicts the devices,count of the devices and the total and overall cost of deploying the components

Chapter 7

Conclusions and Future Scope

In our work, we are proposing a system that will bring a certain amount of automation with the help of emerging technologies like A.I.that will help to reduce the burden of day-to-day activities of security professionals like going through the millions of logs swiftly, carrying out necessary procedures and targeting the areas where human intervention is required the most.We are also planning to automate one of the important factors that plays major role in identifying and preventing previously held attacks i.e. Threat Intelligence collection.Also, we are trying to translate threat intelligence in different languages.
In future work,we will be focusing more on the zero-day exploit attacks,drive-by attacks and the eaves dropping attacks.Preventing these will be a need in the future as such attacks are getting increased day by day.Also the support of translation engine for multilingual threat intelligence will also be made for other languages such as Japanese and French languages. This will help us to increase the reach of our system.

Bibliography

[1] Farhan Sadique, Raghav Kaul, Shahriar Badsha, Shamik Sengupta,"Ana utomated Framework for Real-time Phishing URL Detection",in IEEE 10th Annual Computing and Communication Workshop and Conference (CCWC), Accepted For Publications, 2020.

[2] Farhan Sadique, Khalid Bakhshaliyev, Jeff Springer, ShamikSengupta, "A system architecture of cyber security information exchange with privacy (cybex-p)",in IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), Accepted For Publications, 2019.

[3] Jonghoon lee, Jonghyun Kim, Ikkyun Kim, and Kijun Han, "Cyber Threat Detection Based on Artificial Neural Networks Using Event Profiles", in IEEE Access, Accepted For Publications, 2019.

[4] Hamad Almohannadi , Irfan Awan , Jassim Al Hamar , Andrea Cullen, Jules Pagan Disso , Lorna Armitage, "Cyber Threat Intelligence from Honeypot Data Using Elastic search", in IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA), Accepted For Publications, 2018.

[5] Farhan Sadique , Sui Cheung , Iman Vakilinia ,Shahriar Badsha ,Shamik Sengupta, "Automated Structured Threat Information Expression (STIX) Document Generation with Privacy Preservation",in 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), Accepted For Publications,2018

[6] Priyanka Ranade, Sudip Mittal, Anupam Joshi, Karuna Joshi, "Using eep Neural Networks to Translate Multi-lingual Threat Intelligence IEEE International Conference on Intelligence and Security Informatics (ISI), Accepted For Publications, 2018.

Appendices

Appendix-A: Configuring TPOTCE

git clone https://github.com/telekom-security/tpotce cd tpotce/iso/installer/ ./install.sh –type=user

Appendix-B: How To Install Elasticsearch, Logstash,and Kibana (Elastic Stack) on Ubuntu 18.04

The Elastic Stack has four main components: 1.Elastic search:a distributed RESTful search engine which stores all of the collected data
2.Logstash: the data processing component of the Elastic Stack which sends incoming data to Elastic search
3.Kibana: a web interface for searching and visualizing logs
4.Beats: lightweight, single-purpose data shippers that can send data from hundreds or thousands of machines to either Logstash or Elastic search

Step 1 — Installing and Configuring Elasticsearch file,elasticsearch.yml. Here,we'll use nano: sudo nano /etc/elasticsearch/elasticsearch.yml
/etc/elasticsearch/elasticsearch.yml
. . . network.host: localhost
. . . Save and close elasticsearch.yml by pressing CTRL+X, followed by Y and then ENTER if you're using nano. Then, start the Elasticsearch service with systemctl: sudo systemctl start
Next, run the following command to enable Elasticsearch to start up every time your server boots:
sudo systemctl enable elasticsearch
You can test whether your Elasticsearch service is running by sending an HTTP request: curl -X GET "localhost:9200"

Appendix-C: I How to Install And Setup Suricata IDS on Ubuntu 20.04

Prerequisites: A fresh Ubuntu 20.04 VPS on the Atlantic.net Cloud Platform A root password is configured on your server

Step 1 – Create Atlantic.Net Cloud Server First, log in to your Atlantic.Net Cloud Server. Create a new server, choosing Ubuntu 20.04 as the operating system, with at least 2GB RAM. Connect to your Cloud Server via SSH and log in using the credentials highlighted at the top of the page. apt-get update -y

Step 2 – Install Required Dependencies
First, you will need to install some dependencies required to compile Suricata from the source. You can install all of them with the following command:
-apt-get install rustc cargo make libpcre3 libpcre3-dbg libpcre3-dev build-essential autoconf automake libtool libpcap-dev libnet1-dev libyaml-0-2 libyaml-dev zlib1g zlib1g-dev libcap-ng-dev libcap-ng0 make libmagic-dev libjansson-dev libjansson4 pkg-config -y

By default, Suricata functions as an intrusion detection system (IDS). If you want to include intrusion prevention system (IPS) functionality, then you will need to install some more packages in your system. You can install them with the following command:
apt-get install libnetfilter-queue-dev libnetfilter-queue1 libnfnetlink-dev libnfnetlink0 -y

Once all the packages are installed, you will need to install the suricata-update tool to update the Suricata rules. You can install it with the following commands: apt-get install python3-pip
pip3 install –upgrade suricata-update
ln -s /usr/local/bin/suricata-update /usr/bin/suricata-update

Step 3 – Install Suricata
First, download the latest version of Suricata from their official website with the following command: -wget https://www.openinfosecfoundation.org/download/suricata-5.0.3.tar.gz

Once the download is completed, extract the downloaded file with the following command: tar -xvzf suricata-5.0.3.tar.gz Next, change the directory to the extracted directory and configure it with the following command:
cd suricata-5.0.3 ./configure –enable-nfqueue –prefix=/usr –sysconfdir=/etc –localstatedir=/var
Next, install the Suricata with the following command: make make install-full

Publication

Paper entitled ***"Artificial Intelligence based Security, Orchestration,Automation and Response System"*** is published at ***"6th IEEE 12CT 2021"*** by ***"Rahul Vast", "Shruti Sawant", "Aishwarya Thorbole" and "Vishal Badgujar."***

Link: `https://ieeexplore.ieee.org/document/9418109`