Detecting Unknown Attacks using Big Data Analysis

[1]Miss Bhagyashree S Jawariya, [2]Prof. P.N. Kalvadekar
[1] ME II year, [2] PG Co-ordinator
[1] Computer Department,
[1] Sanjivani College of Engineering, Kopargaon, India

***Abstract* - Nowadays threat of previously unknown cyber-attacks are increasing because existing security systems are not able to detect them. Previously, leaking personal information by attacking the PC or destroying the system was very common cyber-attack. To provide security to network I use Big Data Analysis System for Detecting known attack and unknown attacks with low false alarm. To provide this security I am using Big Data Analysis System. This system consist of four phases viz Data Collection, Clustering, Data Analysis and Detection. Here the packets are send from client to server, then all attributes are extracted for each packet in a network. Then I am comparing attributes of each packet with KDD dataset. If the attributes match then known attack found. If attacks are Unknown i.e they are not matched with the KDD Dataset then the pop up message or an alarm is generated.**

***Index Terms* - Computer crime, Alarm systems, Intrusion detection, Data mining, Big Data**

I. INTRODUCTION

Self-propagating malware, such as worms, have prompted cyber attacks that compromise regular computer systems via exploiting memory-related vulnerabilities which present threats to computer networks. A new generation worm could infect millions of hosts in just a few minutes, making on time human intrusion impossible. The new worms are spread over the net work on regular basis and the computer systems and network vulnerabilities are growing enormously. Here we also facing the problem of automatically and reliably detecting previously unknown attacks. Previously, leaking personal information by attacking the PC or destroying the system was very common cyber attacks .Existing defence technologies to counter these attacks are based on pattern matching methods which are very limited. Because of this fact, in the event of new and previously unknown attacks, detection rate becomes very low .Wikipedia defines zero-day virus as a previously unknown computer virus or other malware for which specific antivirus software signatures are not yet available. However, according to various reports, intrusion detection systems and intrusion prevention systems are not capable of protecting systems against APT attacks because there are no signatures. Therefore to overcome this issue, security communities are beginning to apply heuristic and data mining technologies to detect previously unknown attacks. In this paper, a new model is proposed based on big data analysis technology to prevent and detect previously unknown APT attacks.

APT Attacks. APT attack penetrate into the target system and persistently collect valuable information by using social engineering, zero day vulnerabilities and other techniques. It can damage national agencies or enterprises. They are also used as a cyber-weapons. Instead of Targeting ordinary desktops or servers they target industrial control systems. APT attack is usually done in four steps:

Intrusion, Searching, Collection and Attack

Intrusion Step - In the intrusion step of an APT attack, the hacker probes for information about the target system and prepares the attack.

Searching - To get the access to the system, the attacker searches for users with high access privileges such as administrators and use various attack techniques such as SQL injection, phishing, farming and social engineering to hijack their accounts searching is done after the hacker gained access to the system. Hacker analyses system data such as system log for valuable information and look for security vulnerabilities than can be exploited for further malicious behaviors.

Collection - In this next step, after the hacker has located valuable information in the system such as confidential documents etc, then, he installs malwares such as rootkits, backdoors to collect system data and maintain system access for the future.

Attack - In this final step, the hacker leaks data and destroys target system using acquired privileges. Leaked information can be used for developing other additional security vulnerability exploits. Because APT exploits use zero-day vulnerabilities and obfuscation methods, Anti-Virus program, IDS and IPS are difficult to detect such exploits

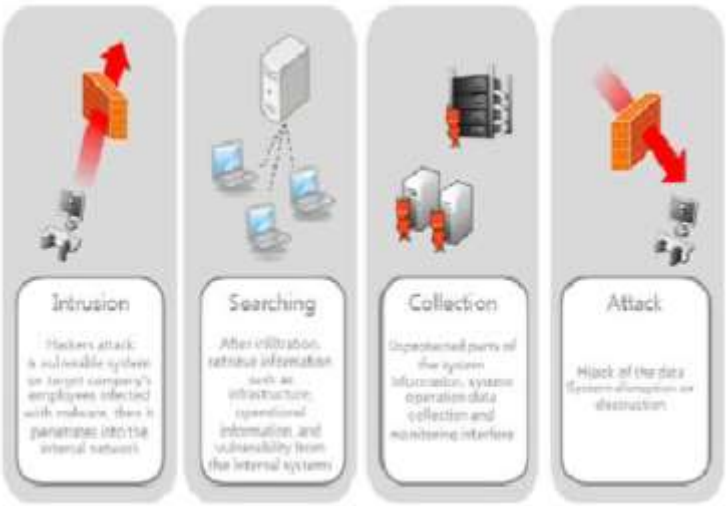

Fig 1 Sequence Diagram for APT attacks

II. LITERATURE SURVEY

Big data analysis system concept for detecting unknown attacks [1]in this a new model based on big data analysis techniques that can extract information from a variety of sources to detect future attacks is given
Big Data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data [2]

Advanced Persistent Threats: A Decade in Review [5] In this the term Advanced Persistent Threat (APT) in the context of cyber threats and cyber attack is explained
Cloud Model based Outlier Detection Algorithm for Categorical Data [4] This algorithm is based on data driven idea and does not require the user to specify parameters
NICE: Network Intrusion Detection and Countermeasure Selection in Virtual Network Systems [3]

III. SYSTEM OVERVIEW

Figure 2 shows the Big Data Analysis System Model for Detecting Unknown Attacks . As illustrated in the design, from various sources the data is being collected . The extracted data is taken as input and is provided to the system for pre-processing. After preprocessing the data it is analysed . The Analysis is done on the basis of Behaviour Matching . Genetic Algorithm is used for behavior matching . If any unknown behavior is found then an alert will be generated by the system

It consists of following phases

Data Collection and Creation of Network

Data collection step collects event data . The Event data is collected from firewalls and log, Servers, application , behaviour, status information (date,time, inbound/outbound packet, daemon log, user behaviour, process information etc.) from anti-virus, database, network device and system. Data appliance is used to store the collected data . The Network is been created by client server application . Through this the data will be send through .

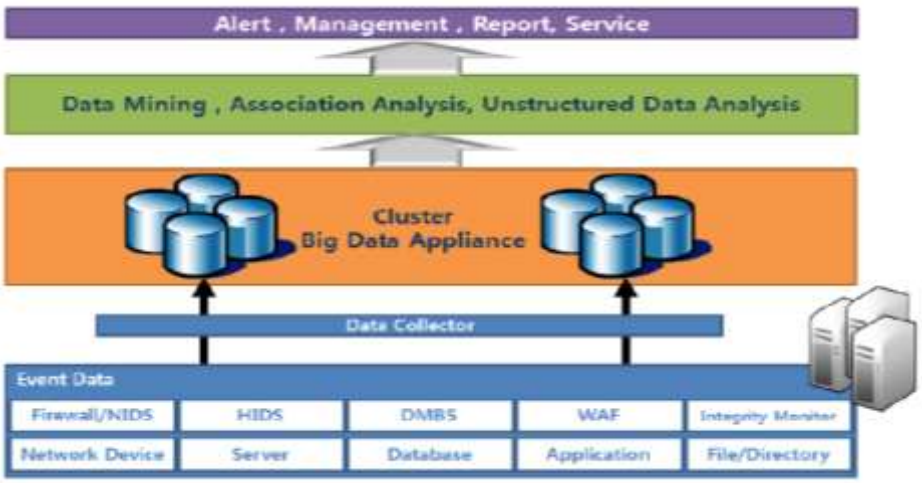

Fig .2: System Overview

Analysis of Data

Pre-processed data from previous step is analysed using Data Mining Technique . The analysis is done by matching the input packet attributes with the KDD Dataset attribute. In this firstly the system perform integrity monitoring(checks its own system

for any viruses or attack) After this the collected data which was stored is then Normalized and distributed in various clusters The clustering is done on the basis of K-Means algorithm . Then this clustered data is check for any attack

Detection
If attack or abnormal behaviours are detected, it alarms the administrator and give pop up message . The mail is also sent to the administrator about the attack

Objectives:

- To detect the attacks which are known and unknown
- Ideally, the system should be able to detect the unknown attacks , with a high degree of accuracy.
- The system should work better than existing Information Security Technologies like Firewall, IDS , WAF.

Scope :

- To provide security against Known and Unknown attacks.
- Distinguishing Between the Known attacks and Unknown attacks
- To Provide security to Big Data

IV. BREAKDOWN STRUCTURE

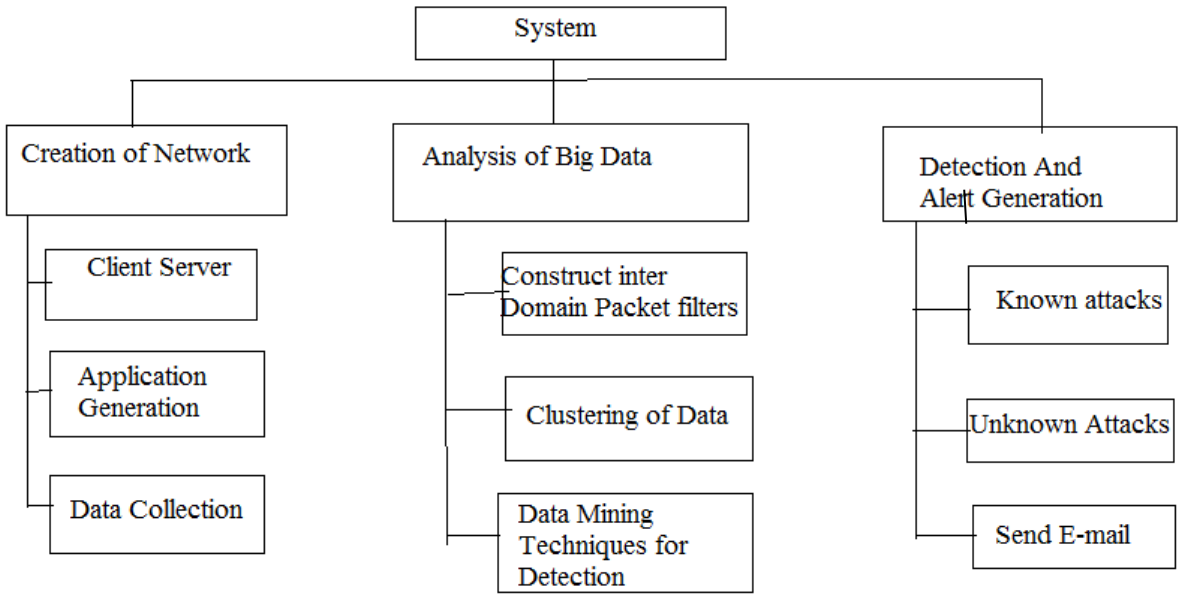

Fig.3 Breakdown Structure

Module 1 . Creation of a Network

a. Client Server
1. In this the client server is created
2. The packets are Send from client to server

b. Data Collection
1. In the the data is collected from various sources

Module 2 .Analysis of Big Data

a. Inter Domain Packet Filters
1. In this the system itself checks for any attacks

b. Clustering of Data
1. The data collected from module 1 is clustered in different clusters
2. For clustering K-Means algorithm is used

c. Data Mining Techniques for Detection
1. Data mining data techniques are used for detection
2. Behaviour Matching using Genetic Algorithm . Here Genetic Algorithm is used for Behaviour Matching . The Behaviour of the received packet is matched with the already known behaviours . If the behaviour is not Matched then it is Considered as Unknown

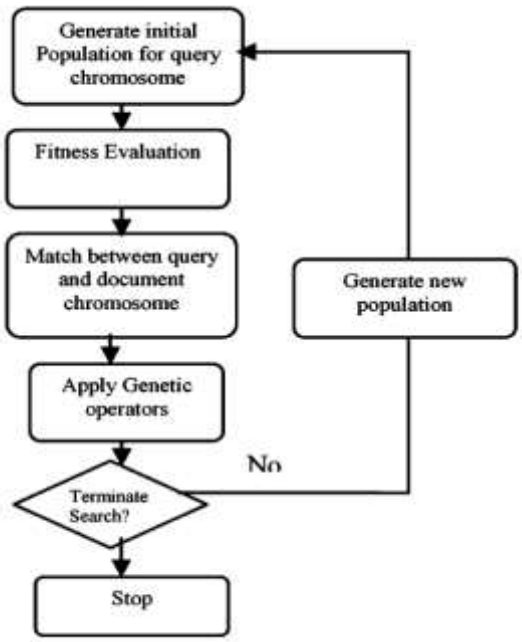

Fig 4 Genetic Algorith

Initially all the known attacks set is created . If any attack comes , first it is checked whether it is known or unkown i.e it is checked whether it matches with the known attacks set or not . If a match is found with the known attack set then it will get prevented as solutions are already present for them . But if matching does not found then an alert will be generated by the Detection Engine and reported to administrator

Module 3 : Alert Generation
1.Alert is indication for detection of attack.
2.Alert is generated, when known or unknown attack found.
3.Attack message display on system if attack found.
4.Alert can be in the form also

V. DATASET

1.KDD Dataset:
KDD Dataset is used for Detection of attacks as known or unknown

VI. RESULT ANALYSIS

Following tables shows the efficiency of the System

Table 1 Average Efficiency of the System

Day	Number of packets	Number of packets contain attacks	Number of packets detects attacks	True Positive	False Positive
Day1	29	29	24	83%	17%
Day2	29	29	21	72.41%	27.59%
Day3	29	23	18	78.26%	21.74%
Day4	24	25	19	76%	24%
Day5	20	16	12	75%	25%
Average Efficiency = 76.934%					

True positive=No.of attacks detected/Total no. of packets
for day1=24/29
=83%
False Positive=100-true positive
=100-83
=17%

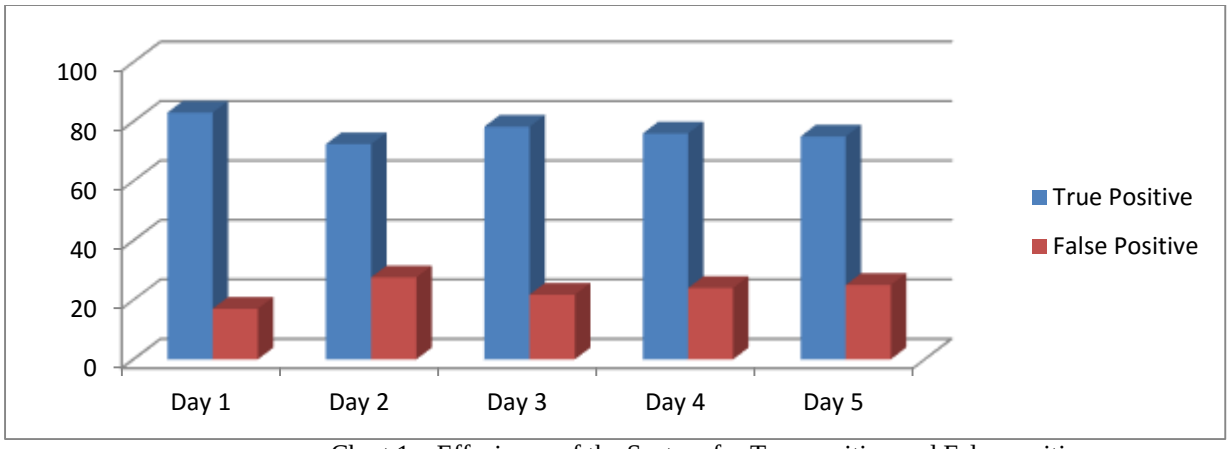

Chart 1 – Effeciency of the System for True positive and False positive

Table 2 Average false rate for the system

Day	Number of packet send	Number of attacks detected	False Alarm	False Alarm Rate
Day1	29	24	1	4%
Day2	29	21	2	9%
Day3	29	18	1	5%
Day4	24	19	1	5%
Day5	20	12	1	8%
Average false rate = 6.1 %				

False alarm rate=false alarm/No.of attack detected
=1/24
=4%

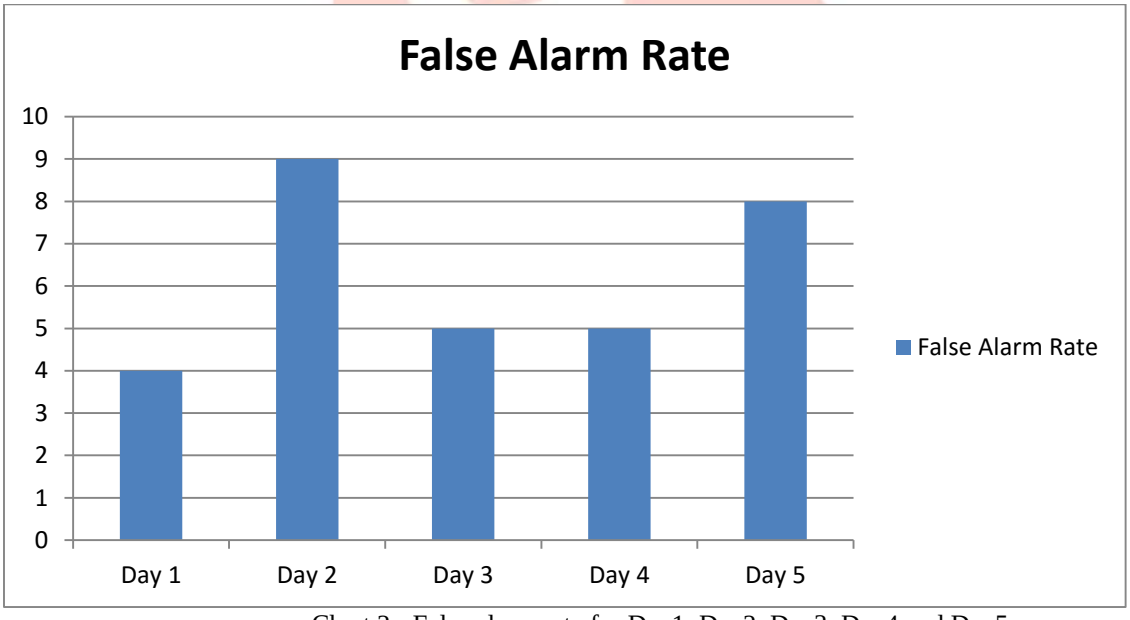

Chart 2 - False alarm rate for Day1, Day2, Day3, Day4 and Day5

Table 3:Number of Different categories of attacks found in 5days

Days	No. of different categories of attacks found			
	Normal	DDOS	R2L	Unknown
Day1	6	9	2	12
Day2	5	3	1	10
Day3	6	11	0	5
Day4	10	6	0	7
Day5	4	7	0	5

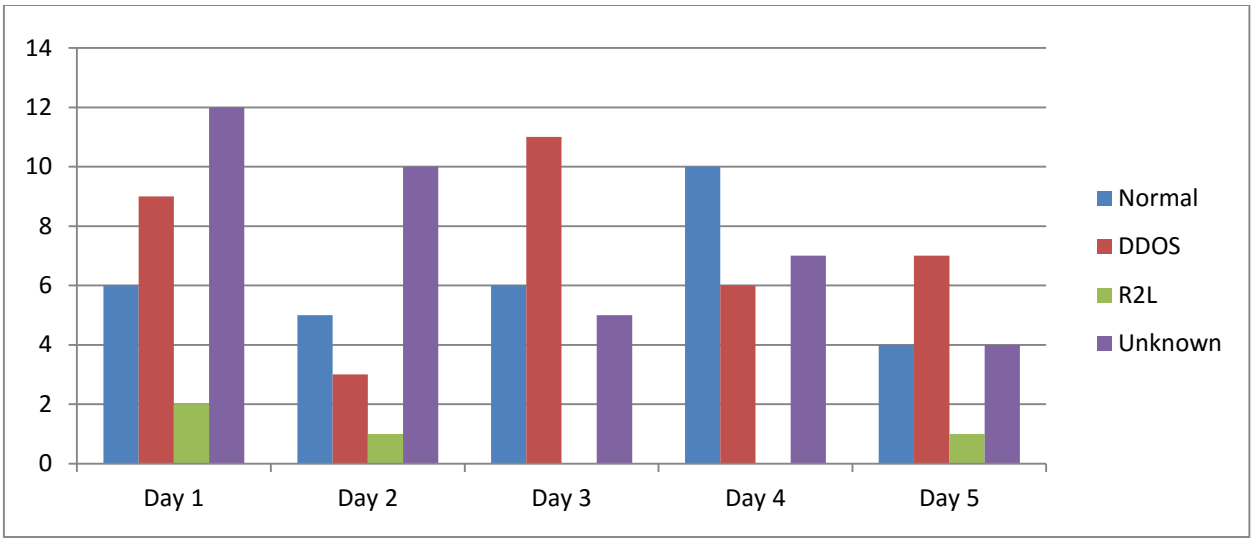

Chart 3 No. of Normal, DDOS,R2L and Unknown attacks found in 5 days.

VII. CONCLUSION

In this a Big Data Analysis Model for reacting to previously unknown cyber threats is been given . Unknown attacks can easily bypass through the existing security solutions by using encryption and obfuscation. Existing security technologies to counter these attacks are based on pattern matching methods which are very limited . The new detection methods for reacting to such unknown attacks are in need. Therefore this System gives a model ”Big Data Analysis System Model” for reacting to previously unknown cyber threats .

VIII. ACKNOWLEDGMENT

First and foremost, I would like to give thanks to God Almighty for His mercy and beautiful blessings in my life. Even in the hardest times, He has always been there to guide and comfort me and to make sure that I never stray too far from His narrow way.I would like to thank my family for their unconditional love and support. I would especially thanks my father and my grandmother. I also wish to thank my mother and my brother for being there for me, despite all the troubles and problems that grow in life.

I express my sincere gratitude to **Prof. D. B. Kshirsagar**, Head of Department (Computer), and **Prof. P.N. Kalavadekar**, (Guide/ PG Co-ordinator) SRESCOE for his unending support and encouragement during the years I have studied under his tutelage.

I am also grateful to my guide for this project, for his guidance and useful insight. This project would never have been possible without his instruction and teaching.

My sincere thanks go to all staff for their help and understanding.

REFERENCES

[1] Tai-Myoung Chung Sung-Hwan Ahn, Nam-Uk Kim, ”Big data analysis system concept for detecting unknown attacks” , Technical report, IEEE Transaction , February 2014

[2] Dr. Kiran Jyoti Bhawna Gupta, ”Big data analytics with hadoop to analyze targeted attacks on enterprise data” , Technical report, International Journal of Computer Science and Information Technologies, IJCSIT, 2014

[3] Tianyi Xing Jeongkeun Lee Chun-Jen Chung, Pankaj Khatkar and Dijiang Huang, ”Nice: Network intrusion detection and countermeasure selection in virtual network systems” , Technical report, IEEE Transac- tions on Dependable and Secure Computing ,August 2013.

[4] Liping Zhang2 Dajiang Lei1 and Lisheng Zhang,”Cloud model based outlier detection algorithm for categorical data”, Technical report, International Journal of Database Theory and Application , August 2013.

[5] Command Five Pty Ltd, ”Advanced persistent threats: A decade in review” , Technical report , June 2011.

[6] Philomina Simon , S. Siva Sathya , ”Genetic Algorithm for Information Retrieval” , Pondicherry University, Technical report, IEEE Transactions , 2009.

[7] Christopher J.C. Burges,”A Tutorial on Support Vector Machines for Pattern Recog nition”, Kluwer Academic Publishers, Boston. Manufactured in the Netherlands