Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 2; January-March, 2015 pp. 112-116

http://www.krishisanskriti.org/acsit.html

A Survey of Big Data Analytics for Network Traffic Monitoring to Identify Cyber Attacks

Amreesh Kumar Patel[1] and D.S. Bhilare[2]

[1]School of computer science& IT DAVV, Indore, M.P., INDIA
[2]School of computer science & IT DAVV, Indore, M.P., INDIA
E-mail: [1]amreesh21@gmail.com, [2]bhilare@hotmail.com

Abstract—*The term Big Data refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies. The rapid growth of the Internet has brought with it an exponential increase in the type and frequency of cyber attacks. The security related data is increasing in the form of Volume, Velocity and Variety. Facing these problem traditional approaches of network monitoring and traffic measurement fails. In this paper, we present the role of Big Data Analytics in network management and traffic monitoring. We also analyze the challenges of Network Monitoring and Traffic measurement for Big Data with low rate of false positives to detect cyber attacks.*

1. INTRODUCTION

Internet is one of the fastest-growing areas of technical infrastructure development. In today's business environment, troublesome technologies such as cloud computing, social computing, and next-generation mobile computing are primarily changing, how organizations utilize information technology for sharing information and conducting commerce online. Today more than 80% of total commercial transactions are online, due to which this field required a high quality of security for transparent and best transactions. The trend of Cyber Security extends not only to the security of IT systems within the enterprise, but also to the large digital networks upon which they rely including cyber space itself and critical infrastructures[1]. Traditional cyber security techniques or tools are not capable to detect cyber attack efficiently in Big Data era. To present the capability of Big Data analytics and process, retrieve the useful information

from large amount of raw data and detect the unknown attack.

Big data -"Big data in general is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" By Gartner[16].

Data is the new Oil. Data is just like crude. It's valuable, but if unrefined, it cannot really be used**:** By- Clive Humby.

Big data analytics - The large-scale analysis and processing of data in recent years, has pay attention on interest of security community for its assuring ability to analyze and correlate security-related data efficiently. Big Data analytics can be employed to analyze log files and network traffic to identify anomalies and suspicious activities. It also correlate multiple sources of information into a coherent view. Analytics applied data sources communally will provide a 360 view of network traffic, e.g., by singling out an abnormal behavior in the access pattern of a given user[2]. Need of big data analytics for cyber security , show in the fig.1. Traditional security operation and technology work efficiently at limited amount of data and limited variety of data and generate a many false positive alarm but now volume, velocity and variety of data are growing rapidly in the form of network flow, business process data, banking transaction. Big Data is a challenge of cyber security point of view because maximum data are unstructured form. By using, big data analytics one can convert challenges in the form of opportunity.

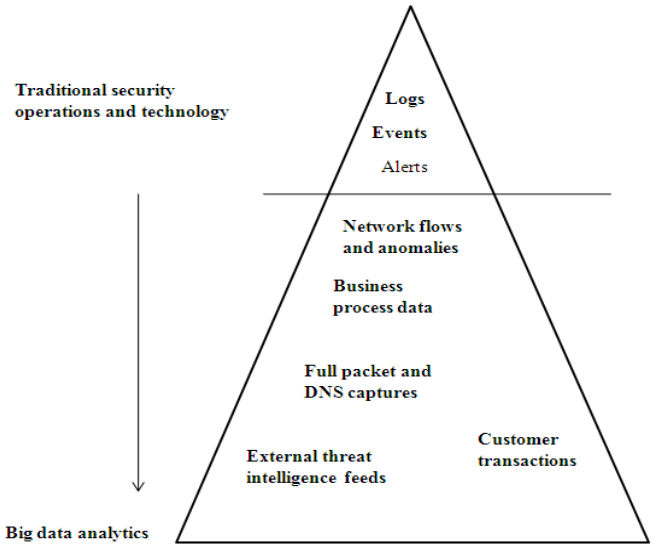

Fig. 1: The Variety And Volume Of Data Is Increased.

2. RELATED WORKS

Cyber security is the activity of protecting information and information systems (networks, databases, data centers and applications) with appropriate procedural and technological security measures. Firewalls, antivirus and other technological solutions for protect personal data and computer networks are essential but not sufficient to ensure security because of data, this traditional technology is not able to process unstructured data sets[3].

The term Big Data refers to large amount of data management and analysis technologies which exceeds the capability of traditional data processing technologies. Big Data is dissimilar from traditional technologies in three ways: the amount of data (volume), the rate of data generation and transmission (velocity), and the types of structured and unstructured data (variety). The amount of data generated is about 2.5 quintillion bytes per day. The rate of data generation has increased 90% of the data in today's world is generated in the last two years alone. This acceleration in the production of information has created a need for new technologies to analyze massive data sets[4].

Today, we are in a flood of data. Statistics show that 90% of the world's data was generated in the last two years itself, and it is growing exponentially. To tackle such a massive amount data, we need to depart the traditional batch processing behind and adopt the new big data analytical tools. The data generated everyday exceeds 2.5 quintillion bytes, which is an unimaginable figure. The growth of data has affected all fields, whether it is the business sector or the world of science. To process such huge amounts of data various new tools are being introduced by companies like Oracle and IBM, while on the other handle Open Source developers continue their work in the same field[8].

Recent unknown attacks easily bypass existing security solutions by using encryption and obfuscation. Therefore, new detection methods for reacting to such attacks are in need. To proposed big data system model for reacting to previously unknown cyber threats and researched on the deduction of practical technologies [9].

This document describes how incorporation of Big Data is changing security analytics by providing new tools and opportunities for leveraging large quantities of structured and unstructured data. It highlights the differences between traditional analytics and Big Data analytics, and briefly discusses tools used in Big Data analytics. It also reviews the impact of Big Data analytics on security and provides examples of Big Data usage in security contexts. It describes a platform to perform experiments on anti-virus telemetry data[12].

To present the effectiveness of BOCISS (Behavioral Observation for Critical Infrastructure Security Support). The Classification techniques used to present a demonstration on how our system is able to support security by applying big data analysis techniques to identifying anomalous behavior caused by cyber-attacks taking place[13].

Big Data and security not only intersect at the protection of Big Data infrastructures, but also at the leveraging of Big Data analytics to improve the security of network. Big Data analytics in real-time security monitoring, consists of two main angles: (a) monitoring the Big Data infrastructure itself (b) using the same infrastructure for data analytics Real-time security monitoring is a challenge because of the number of alerts generated by security devices. These alerts (correlated or not) lead to a massive number of false positives, which are often ignored due to limited human capacity for analysis, this problem might be increase with Big Data. Big Data analytics used to monitor anomalous connections to the cluster and mine logging events to identify suspicious activities and reduce the trustworthiness of the data sets used to train Big Data analytics algorithms. It is practically impossible to imagine the next application without it consuming data, producing new forms of data, and containing data-driven algorithms. The use of new infrastructures such as NoSQL database are capable to process the unstructured data seats in run-time environment[14].

Analytics applied data sources collectively will provide about 360 view of network traffic, e.g., by singling out an abnormal behavior in the access pattern of a given user. Appropriate prevention techniques can apply, e.g., lock accounts, quarantine, modify network settings, multiple authentications, alert on an on-going fraud etc. The data analytics converts a vast amount of raw data into useful information and subsequently into actionable knowledge[15].

Big Data storing in distributed environment and parallel processing thus driven by the big data 3v's faces new security and privacy challenges. The 3v's challenges flourishing of big data and hinges on fully understanding and managing newly arising security and privacy challenges. Collected data maybe personal and sensitive, we must have choice of physical protection methods as well as information security techniques to ensure data privacy. Therefore, we need to ensure the confidentiality of stored data in both physical and cyber ways. The key component of big data analytics is big data processing. The privacy requirements of big data processing becomes more challenging. We never forgo big efficiency for big privacy, big data need to process in distribute environment or cloud computing environment, they not only protect individual privacy but also ensure efficiency at the same time[17].

The Intrusion Detection Module can identify Port Scan attacks, Denial of Service and many others with carefully

designed attack patterns that are compare to the NetFlow records in the Database. DDoS attack are implemented by either forcing on targete computer(s) to reset, or consuming its resources so that it can no longer provide its intended service or obstructing the communication media between the intended users and the victim so that they can no longer communicate adequately. NetFlow monitoring has become a prevalent method for network traffic monitoring in high-speed networks. By focusing on the analysis of flows, rather than individual packets, it said to be more scalable than traditional packet-based traffic analysis[18].

A rule based anomaly detection technique that can help MSPs (Mobile Service Providers) to improve their system consistency and reduce the time to detect location based anomalies. A rule based approach using Big Data Analytics technique is promising for both detecting those anomaly problems with high accuracy and presenting flexible and robust ways to solve these problems according to operators needs and domains. The advantages of proposed anomaly detection method, is fast lifecycle in order to reach target results, better ways to extract relevant rules without needing a training phase and the ability to use modern cost-effective big data technique[19].

3. ATTACKS DETECTION STRATEGIES

3.1 Intrusion Detection Systems (IDS)

Intrusion detection systems that are reside on and monitor an individual host machine. Intrusion Detection Systems are categorized into two categories based on detection techniques they use[6].

3.1.1 Behavior Based Detection. Behavior based intrusion detection system assumes that an intrusion can be detected by observing a variation from normal or expected behavior of the system or the users [6].

3.1.2Signature Based Approach. A signature based IDS monitor packets on the network and compares them against a database of Signature or attributes from known malicious threats [6]**.**

3.2. Anomaly Detection

Anomaly detectors identify abnormal behavior on a host or network. They function on the postulation that attacks are different from genuine activity can be detected by systems that identify these differences. statistical method for anomaly detection is one of the oldest techniques applied in IDS research. In this approach, the normal user behavior is shows what is acceptable within the system usage policies[7].

4. LIMITATION OF TRADITIONAL APPROACHES

- Storing and retaining a large amount of data is not economically feasible. Most of the event logs and other recorded computer activity is deleted after a fixed retention period[2].
- Traditional techniques only detects known attack on the bases of signature matching[12].
- Big Data analytics and complex queries on large, structured and unstructured data sets were inefficient because traditional tools did not leverage Big Data technologies.
- Traditional network traffic monitoring tools take so much time to process the big data because they have a single system environment[12].
- Ttraditional tools was not designed to analyze and manage unstructured data. Traditional tools had rigid, defined schemas. Big Data tools (Piglatin scripts and regular expressions) can query data in flexible formats[18].
- Big Data systems use cluster-computing infrastructures. Those systems are more reliable and available, and provide assurance that queries on the systems are processed to completion[12].

As compare to traditional approaches, security analytics provides a “richer” cyber security context by separating what is normal and what is abnormal, i.e., separating the patterns generated by genuine users from those generated by suspicious or malicious users.

5. MODEL OF NETWORK TRAFFIC MONITORING

Previously unknown attacks are evolving to bypass existing security measures. These attacks are impossible to detect or prevent with traditional technologies. New security paradigm to react to these attacks is in need. The new paradigm requires big data analysis techniques as a core and integration of security technologies[5]. We propose a system model that uses big data analysis technology for extracting data from various sources to react to previously unknown attacks. The era of Big Data traditional tools is not capable to process vast amount and high velocity of security related data at efficient way. Network security monitoring system, data efficiency, effectiveness and scalability is very important. Therefore, through the standardization and integration of security incidents, security event data can manipulated in a convenient format based on the standard data, removing redundant and outdated data[5]. Big Data analytics model consist of four modules, Data Collection, Data Integration, Data Analysis and last one is Data Interpretation that’s are show in Fig. 2.

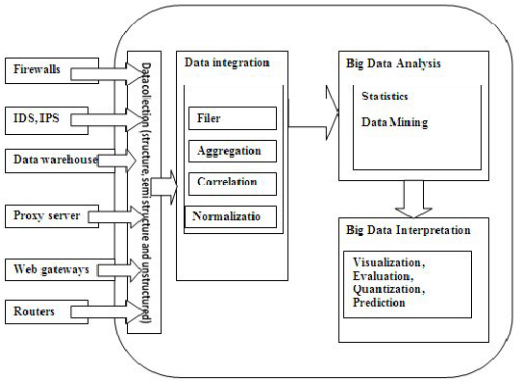

Fig. 2: Big Data Analytics Model

5.1 Big Data collection Module

The collection of data including traffic data, event logs, and security logs. Some aspects of network traffic monitoring system in Big Data environment need to be considered, Data collection includes entire network system including network devices (routers, switches, etc.), security devices (firewalls, IDS, anti-virus, etc.), the host server's logs and events information, alarm information. The fundamental purpose of Big Data collection is extract useful knowledge from raw data based on demands. Then, we apply it to detect the cyber attack[10].

Big Data Sources for Security Analytics is the concept of data for security analytics is expansive[19], data sources can include:

- Computer-based data, e.g., geographical IP location, computer health certificates, keyboard typing and click stream patterns, WAP data.
- Mobile-based data, e.g., GPS location, WAP data.
- Travel data, e.g., travel patterns, destinations, and itineraries.
- SIEM data, e.g., network logs, threat database and application access data.
- Social media data, e.g., Twitter, Facebook, internal office network etc.
- Biometric identification data, e.g., fingerprint, facial recognition, voice recognition, handwriting recognition.

5.2 Big Data Integration Module

Event Collector module collectes mostly messy data, various information sources report different data format to describe different, which gives the system detects network behavior inconvenience, we deal with this through the integration modules .This module will filter raw data according to the classification rules. Since unstructured data characteristics of Big Data, the work is a long way to go Big Data cleaning filters and quality management techniques[11].

- **Filer**: A storage filer is a file server designed and programmed for high-volume data storage, backup, and archiving. Storage filers are network attached storage (NAS) filers, storage file servers, or storage area network (SAN) filers.
- **Aggregation:** Data aggregation is any process in which information is gathered and expressed in a summary form, for purposes such as statistical analysis.
- **Normalization:** Database normalization is the process of organizing the fields and tables of a relational database to minimize redundancy.

5.3 Big Data Analysis Module

Hadoop distributed file system that is capable to process a large amount, high velocity and large variety of data sets .This module is the key to the whole system, the processed data removes the redundant and the entire monitoring system is still initial data. We need to analyze the core event data according to the events of the network environment, system services and event status. Data are divided in small chunk on the bases of port number and packet count and mapper distribute the data to data node and after process data collect by data collector (name node).The knowledge based for typical security loopholes, typical safety behavioral pattern and scene. After the incident, we can be mapped to the system structure and determine the nature of incident and the location of event. The difficulty in this module is the representation of knowledge association algorithm and background of knowledge base[20].

Map Reduce programming model, the computation takes a set of input key/value pairs, and produces a set of output key (port number.)/value (packet count) pairs. Map and Reduce are two basic functions in the MapReduce computation. Map takes an input pair and produces intermediate key/value pairs. The Hadoop MapReduce library will group the intermediate values according to the same key. Reduce that will merge the intermediate values for smaller values.

- **Mapper**: Net-flow mapper reads each flow record split by new lines. A flow record has attributes of timestamp, IP addresses, port, protocol, flag, packet count, and interface numbers. After reading a flow record, filter out necessary flow attributes for a flow analysis job. When the flow analysis job sums up packet counts per destination port number, we set key/value pairs (dst. port, packet count). The flow mapper will write its temporary results on the local disk [11].
- **Reducer**: The flow reducer will be called with the inputs as the intermediate values generated by flow mappers. As in the port-breakdown for example, a value list of packet count belonging to the same destination port number summed up. After merging packet count values associated with the destination port, the flow reducer writes the packet count value for each port number[11].

5.4 Big Data Interpretation Module

The analytic results give a visual output of statistical analysis and then we predict the development trend of network behavior, pattern matching and achieve real time analysis of network packets without violating the privacy of data.

6. CONCLUSION

After reviewing the above papers, we can say that Big Data Analytics is a strong approach that is used to network traffic monitoring for identifying a cyber attacks. Size of NetFlow data is increasing drastically and traditional approaches of analytics will take more time to detect an attack or intrusion. Big Data Analytics is a good approach to analyze the huge amount of Network Traffic data sets because big data analytics provides a distributed environment. So, we conclude of that a traditional security tools and techniques are not able to process efficiently 5v's. volume-of data increase exponent form those take a lot of time to process data and big data analytics used a distributed environment, velocity-data flow speed increased then traditional tool not able to monitoring the network efficiently and variety- the data generated are structure, semi structured and unstructured form. Big data analytics overcome the failure of traditional tools and techniques.

REFERENCES

[1] Ravi Sharma, "Study of Latest Emerging Trends on Cyber Security and its challenges to Society", *International Journal of Scientific &Engineering Research*, Volume 3, Issue 6, June-2012 1 ISSN 2229-5518 IJSER © 2012.

[2] Alvaro A. Cárdenas *et al*., "Big Data Analytics for Security", *University of Texas at Dallas @IEEE* nov./dec.-2013.

[3] Jamal Raiyn *et al*., "A survey of Cyber Attack Detection Strategies", *International Journal of Security and Its Applications Baqa Alqarbiah*, Israel*(2014).*

[4] Alvaro A. Cárdenas *et al*., "Big Data Analytics for Security Intelligence", *CSA, Las Vegas, US*, Tech. Rep., Sept. – 2013

[5] Liu Lan, Lin Jun *et al*., "Some Special Issues of Network Security Monitoring on Big Data Environments" *IEEE 11th International Conference on Dependable, Autonomic and Secure Computing*, 2013.

[6] Martin Tomasek & marek cajkovsky *et al.,* "Intrusion Detection System Based on System Behavior" *Technical University of Kosice*, Slovakia (2012).

[7] http://en.wikipedia.org/wiki/Anomaly_detection.

[8] Collins Michael, "Network Security through Data Analysis" *Published by O'Reilly Media, Inc*., 1005 CA 95472, USA.prt.III, ch.10, pp.191-210.

[9] Sung-Hwan Ahn *et al*., "Big Data Analysis System Concept for Detecting Unknown Attacks", *in ICACT*2014.

[10] Collins Michael, "Network Security through Data Analysis" *Published by O'Reilly Media, Inc*., 1005 CA 95472, USA.prt.1, pp.2-13.

[11] Youngseok, Hyeongu Son,"An Internet Traffic Analysis Method withMapReduce" *University, Korea @IEEE* 2010.

[12] Steve Schupp "Limitations of Network Intrusion Detection" @SANS Institute 2000 – 2002.

[13] William Hurst *et al*., "Big Big Data Analysis Techniques for Cyber-Threat Detection in Critical Infrastructures", in ICAINAW-2014.

[14] Cloud security alliance, "Expanded Top Ten Big Data Security and Privacy Challenges", *April* 2013.

[15] S. Curry, E. Kirda, E. Schwartz, W. H. Stewart, and A. Yoran, "Big Data Fuels Intelligence-Driven Security", RSA Security Brief, January, 2013.

[16] http://www.gartner.com/it-glossary/big-data.

[17] Rongxing Lu, Hui Zhu *et al*., "Toward Efficient and Privacy-Preserving Computing in Big Data Era", *IEEE Network, July/August* 2014.

[18] Rick Hofstede, Pavel Celeda *et al.,* "Flow Monitoring Explained: From Packet Capture to Data Analysis With NetFlow and IPFIX", *IEEE communication surveys & tutorial*, vol.16, no.4,fourth quarter 2014.

[19] Dr. Tariq ,Uzma Afzal *et al.,* "Security Analytics: Big Data Analytics for Cybersecurity"@ 2013 *2nd National Conference on Information Assurance* (NCIA).

[20] Yeonhee , Youngseok Lee ., "Toward Scalable Internet Traffic Measurement and Analysis with Hadoop"*ACM SIGCOMM Comp. Comm. Review,korea january*.2013.