Big Heterogeneous Data for Intrusion Detection

Rupali V. Molawade
ME Student,
Computer Engineering Dept.,
Saraswati college of Engineering ,
Kharghar,India.

Vijaya S. Waghmare
Prof.Computer Engineering Dept,
Saraswati College Of Engineering,
Kharghar ,India

ABSTRACT
Intrusion Detection has been heavily studied in both industry and academia, but cyber security analysts still desire much more alert accuracy and overall threat analysis in order to secure their systems within cyberspace. Improvements to Intrusion Detection could be achieved by embracing a more comprehensive approach in monitoring security events from many different heterogeneous sources. Correlating security events from heterogeneous sources can grant a more holistic view and greater situational awareness of cyber threats. One problem with this approach is that currently, even a single event source (e.g., network traffic) can experience Big Data challenges when considered alone. Attempts to use more heterogeneous data sources pose an even greater Big Data challenge. Big Data technologies for Intrusion Detection can help solve these Big Heterogeneous Data challenges. In this paper, we review the scope of works considering the problem of heterogeneous data and in particular Big Heterogeneous Data

Keywords
IDS;BigData;Heterogeneity;Corelation;

1. INTRODUCTION
Intrusion Detection frequently involves analysis of Big Data, which is defined as research problems where mainstream computing technologies cannot handle the quantity of data. Even a single security event source such as network traffic data can cause Big Data challenges. Another Big Data challenge that larger organizations can face is having an incredible amount of host log event data. Large volumes of data are "overwhelming" and they even struggle to simply store the data. Enterprises dealing with such Big Data issues at this scale cannot use existing analytical techniques effectively, and so false alarms are especially problematic. Additionally, it can be very difficult to correlate events over such large amounts of data, especially when that data can be stored in many different formats. Relational database technology can commonly become a bottleneck in Big Data challenges. While traditional Intrusion Detection Systems (IDSs) are a critical component of Intrusion Detection, more focus should be placed on gathering security data from a wider variety of heterogeneous sources and correlating events across them to gain better situational awareness and holistic comprehension of cyber security.

2. TYPES OF INTRUSION DETECTION
Intrusion Detection Systems are broadly classified into two types. They are host-based and network-based intrusion detection systems. Host-based IDS employs audit logs and system calls as its data source, whereas network-based IDS employs network traffic as its data source. A host based IDS consists of an agent on a host which identifies different intrusions by analyzing audit logs, system calls, file system changes (binaries, password files, etc.), and other related host activities. In network-based IDS, sensors are placed at strategic position within the network system to capture all incoming traffic flows and analyze the contents of the individual packets for intrusive activities such as denial of service attacks, buffer overflow attacks, etc. Each approach has its own strengths and weaknesses. Some of the attacks can only be detected by host-based or only by network-based IDS.The two main techniques used by Intrusion Detection Systems for detecting attacks are Misuse Detection and Anomaly Detection. In a misuse detection system, also known as signature based detection system; well known attacks are represented by signatures. A signature is a pattern of activity which corresponds to intrusion. The IDS identifies intrusions by looking for these patterns in the data being analyzed. The accuracy of such a system depends on its signature database. Misuse detection cannot detect novel attacks as well as slight variations of known attacks.

An anomaly-based intrusion detection system inspects ongoing traffic, malicious activities, communication, or behavior for irregularities on networks or systems that could specify an attack. The main principle here is that the attack behavior differs enough from normal user behavior that it cannot be detected by cataloging and identifying the differences involved. By creating supports of standard behavior, anomaly-based IDS can view when current behaviors move away statistically from the normal one. This capability gives the anomaly-based IDS ability to detect new attacks for which the signatures have not been created. The main disadvantage of this method is that there is no clear cut method for defining normal behavior. Therefore, such type of IDS can report intrusion, even when the activity is legitimate.

3. BIG HETEROGENEOUS DATA
Big Data is currently defined using three data characteristics: volume, variety and velocity [1]. It means that some point in time, when the volume, variety and velocity of the data are increased, the current techniques and technologies may not be able to handle storage and processing of the data. At that point the data is defined as Big Data.

When Big Data is present in heterogeneous forms, it can be considered Big Heterogeneous Data regardless of whether that data is input(s) or output(s) of the system. For example, this can arise due to the additive properties of Big Data. If one input is deemed Big Data and is added to another input which is not Big Data, the result will still be Big Data. This can be shown in Equation 1 below:

$$BD(\text{“}BigData\_\_)+NBD(\text{“}NotBigData\_\_)=BD(\text{“}BigData\_\_) \quad (1)$$

Similarly if some advanced data correlation for analysis is occurring and the Big Data is being combined with "Not Big Data" in a multiplicative manner, the result will still be Big Data. This can be shown in Equation 2 below (assuming "Not Big Data" is greater than one):

BD("BigData__)×NBD("NotBigData__)=BD("BigData__) (2)

Therefore, when Big Data is being combined with other data that is not classified as Big Data, the result will still be Big Data. At a high level, Big Heterogeneous Data can be described in terms of being input or output data. Big Heterogeneous Input Data can be further categorized into traditional Big Cyberspace Data and Big Industrial Data (i.e., data from industrial processes in the real physical world). Big Heterogeneous Output Data will be presented in the categories of Big Archival Security Data (which considers the long term storage aspects) and Big Alert Data .

1. Big Heterogeneous input data

It is important to consider that a great deal of heterogeneityamong the sources can be present within these categories. First, the traditional cyberspace input Big Data is presented. Then, Big Heterogeneous Industrial Data beyond cyberspace is discussed, and this section gives examples of BigData from the physical world outside of cyberspace (e.g., industrial process data) which can further improve situational awareness even in cyberspace

- **Big Heterogeneous cyberspace data**

Big Heterogeneous Cyberspace Data are the traditional input types of data which are commonly considered in Intrusion Detection literature, but here they are presented in the context of Big Data. Both network layer and host layer event sources are considered. The network layer coverage is essentially just the network traffic that traditional approaches like NIDSs (e.g., Snort) monitor with a focus on Big Data. The host layer coverage focuses on Big Data challenges with different host sources, and is equivalent to the traditional HIDS approaches where computer servers, workstations, devices, etc. are being monitored.Again, it is important to consider that a great deal of diverse heterogeneity can occur among event sources in this category.NIDS capable of handling BigData network streams suchas these by utilizing Big Data tools such as Hadoop and a network monitoring tool calledPacketPig According to the authors, PacketPig is capable of Deep Packet Inspection,deep network analysis, and even full packet capture when using it with Hadoop.To better cope with Big Data challenges organizationscan face with their log data, Yenet al. developed a system called Beehive which performs "large-scale log analysis for detecting suspicious activity in enterprise networks". They report that organizations are facing Big Volume challenges in terms of the logs being "very large in volume", and implemented their system at a large enterprise, EMC, for two weeks. At EMC, they describe their major challenges as the "Big Data problem" where 1.4 billion log messages are generated on average per day (about 1 terabyte). This also suggests Big Velocity challengesin dealing with such a high data rate as well

- **Big Heterogeneous industrial data**

Cyber threats can damage and even destroy real-world physical targets beyond cyberspace. Industrial and Utility operations are especially prone to this exposure given their evolution of integrating and automating their physical operations with Information Technology from cyberspace. Even when these systems are "air gapped" and physically disconnected from the public Internet and other networks, these cyber threats can still be catastrophic in nature to real-world objects. An example of a successful attack occurred against Iran's nuclear program with the Stuxnet virus, and some of Iran's nuclear centrifuges were destroyed in the attack.

Therefore, it can also be important to include heterogeneous sources from the physical world to better improve overall situational awareness for security. A good conceptual illustration for how to extend monitoring beyond cyberspace is given in Figure 1, and this shows different Host, Network, and Device IDSs harvesting information into a centralized SIEM system with the goal of improving Intrusion Detection by also analyzing data from Process Control System sensors.

This is an interesting concept in that cyberspace situationalawareness can be improved by correlating data from heterogeneous sources in the physical world beyond cyberspace, and that Intrusion Detection need not be merely limited to cyberspace sources. The authors indicate that important industries such as refining, pipelines, and electric power can benefit from this approach of utilizing more diverse heterogeneous sources, while cautioning that the stakes are especially high for detecting cyber-attacks against those platforms, as damage can also be physically harmful or even deadly, such as releasing hazardous materials into the environment.

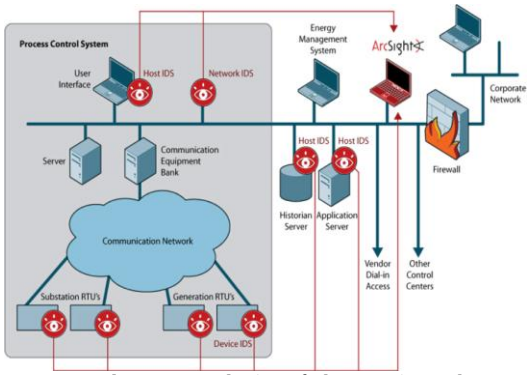

Figure 1: Control Center Level View of the Detection and Event Correlation Frame-work

2. Big Heterogeneous output Data

It can be output data as well, and this is classified as Big Heterogeneous Output Data. This section addresses the heterogeneity of output Big Data for Intrusion Detection in two main categories: Big Archival Data and Big Alert Data. Big Archival Security Data is output data which is being archived either for the purpose of forensics or Security Analytics, while Big Alert Data is output data either for further alerting analysis or for notifying an administrator or system component to take action. Both of these Big Heterogeneous Outputs can have very pronounced Big Data attributes in terms of Volume, Velocity, and Variety.

- **Big archival security data**

A very important aspect for Intrusion Detection is long-term storage of certain security data. Essentially, there are two main goals for the archival of security data. The first goal is to improve Intrusion Detection capabilities even in real-time with offline data mining operations and Security Analytics. This offline data mining operation on security data can further

try to identify previously unknown cyber threats, and then update the real time detection capabilities with additional new signatures or behavior traits. The second goal is to provide forensic capabilities with this data so that in the event of a security breach, forensic evidence is available to assist the investigation. This data can also be used as evidence in legal proceedings if properly maintained. Typically not every single piece of computing data will be kept in the offline repository store, and care must be taken to properly filter out what is not necessary.

- **Big alert data**

Intrusion Detection Systems and other security systems produce alerts to notify administrators of suspicious activity. Even an individual IDS can trigger many alerts, and the problem becomes even more prominent when dealing with heterogeneous sources such as a wide array of sensors or multiple IDSs. The basic problem is that a single security inspection event can trigger many alarms even if it is a single incident, or many false alarms can even be raised with normal traffic. A common technique which is used to stop a flood of alerts is called alert correlation. The basic concept of alert correlation is that when the same characteristic is causing the same alarm, the system should filter and aggregate multiple alarms into one alarm so that a flood of alarms of the same type does not occur (instead just a count of those same alarm types could be reported). An illustrative example of alert correlation is given in Figure 2 where alerts are initially correlated locally in a hierarchical fashion. They are subsequently correlated again at a more global level.

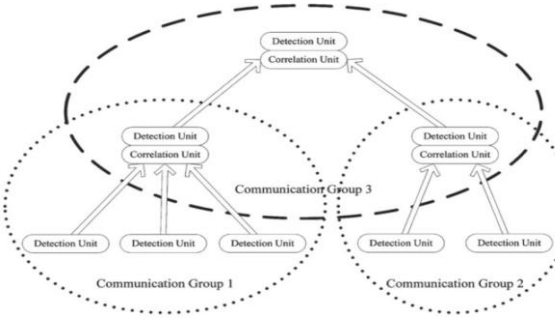

Figure2.Hierarchical CIDS architecture

The process of generating alerts certainly can involve Big Data challenges in terms of Volume, Velocity, and Variety. Big Volume and Big Velocity challenges for alert generation can involve correlation with other alerts, events, rules, or knowledge bases. These correlationactivities can involve massive processing power, storage requirements, and network traffic. Big Variety challenges for alert generation can involve correlation among alertgenerators such as IDSs that can have many different formats for their alert messages or event data. It is common for organizations to have security products with many different proprietary alert formats, even though efforts are still being made to standardize. Semantically,alerts can either be considered inputs or outputs as they can also serve as inputs for alert correlation purposes. Alerts always operate at least once in an output capacity,

But alerts do not always operate in an input capacity. Since alerts are typically considered outputs conceptually for notification purposes as well as for archiving and forensic purposes, they will be categorized as outputs for this study's organizational purpose

4. CONCLUSION

Both cybersecurity and physical security for organizations such as those in the utility and the industrial sector can even be enhanced by correlating traditional IT security events with those beyond cyberspace such as sensor devices measuring anomalous realworld quantities like gas leaks, electrical power/voltage/current, temperature, fire alarms, or many other sensors. Correlating security events from physical world sensors with cyberspace is becoming significantly more important as the utility and industrial sectorsare becoming increasingly computerized for automation, and thus exposing their physical infrastructures to new cyber threats such as malicious attackers or "cyber accidents".

While Intrusion Detection does not always face Big Data challenges, it does face Big Data challenges more often as time progresses and especially more so for larger private and government organizations. This trend of Big Data challenges will continue as a multitude of more heterogeneous sources are analyzed. Even medium and smaller organizations will need to assess whether their Intrusion Detection architecture or Security Analytics merit the deployment costs of Big Data technology.

5. REFERENCES

[1] Richard Zuech,Taghi M Khoshgoftaar(2015),Intrusion detection & Big heterogeneous data,Journal of Big data 2:3.

[2] Suthaharan S, Panchagnula T (2012) Relevance feature selection with data cleaning for intrusion detection system. In: Southeastcon, 2012 Proceedings of IEEE. IEEE, Orlando, FL, USA. pp 1-6

[3] Group BDW (2013) Big Data Analytics for Security Intelligence. https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_Intelligence.pdf. Accessed 2015-1-10

[4] Ismail Butun,Salvatore D.Morgera, A survey of Intrusion Detection Systems in wireless sensor networks,IEEE communications surveys & tutorials,vol.16.no1,First quarter 2014.

[5] Amit Kumar,Harish Maurya,A research paper on hybrid intrusion detection system,IJEAT,vol-2,Issue-4,April 2013.

[6] Mostaque Md,Morshedur Hassan,Current studies on Intrusion Detection System,Genetic Algorithm & Fuzzy Logic,International Journal of Distributed & Parallel System,vol-4,no-2,March2013.

[7] A. Kartit,A. Saidi ,F.Bezzazi,A new approach to intrusion detection system,Jornal of Thoratical & Applied Information Technology,vol-36,n0-2,2012

[8] Shatiullah Khan,Kok-keong Loo,Framework for intrusion detection in IEEE 802.11 wireless mesh networks,International Journal of Information Technology,vol-7,no-4,Oct 2010

[9] Peyman Kabiri and Ali A. Ghorbani, Research on Intrusion Detection and Response, International Journal of Network Security, Vol.1, No.2, PP.84–102, Sep. 2005

[10] Sitaram D, Sharma M, Zain M, Sastry A, Todi R (2013) Intrusion detection system for high volume and high velocity packet streams: A clustering approach. Int J Innovation Manag Technol 4(5):480–485

[11] Yen T-F, Oprea A, Onarlioglu K, Leetham T, Robertson W, Juels A, Kirda E (2013) Beehive: large-scale log analysis for detecting suspicious activity in enterprise networks. In: Proceedings of the 29th Annual Computer SecurityApplications Conference. ACM, New Orleans, LA, USA. pp 199–208

[12] XU X-b, YANG Z-q, XIU J-p, LIU C (2013) A big data acquisition engine based on rule engine. J China Universities Posts Telecommunications 20:45–49

[13] Brosche S, Cheng F, Menial C (2010) A flexible and efficient alert correlation platform for distributed ids. In: Network and System Security (NSS), 2010 4th international conference on. IEEE, Melbourne, Australia. pp 24–31