2nd International Conference on "Innovative Trends in Science, Engineering and Management"
YMCA, Connaught Place, New Delhi ICITSEM-16
05 November 2016, www.conferenceworld.in ISBN : 978-93-86171-10-8

APPLYING BIG DATA ANALYTICS ON INTEGRATED CYBER SECURITY DATASETS

Dr. M. Mohammed Ismail[1], K. Murali[2], P. Rizwan Ahmed[3]

[1]*Associate Professor, Department of Computer Science, Mazharul Uloom College, Ambur*

[2]*Research Scholar, Department of Computer Science, Mazharul Uloom College, Ambur*

[3]*Asst. Professor & Head, Department of Applications, Mazharul Uloom College, Ambur*

ABSTRACT

With the growing prevalence of cyber threats in the world, various security monitoring systems are being employed to protect the network and resources from the cyber attacks. The large network datasets that are generated in this process by security monitoring systems need an efficient design for integrating and processing them at a faster rate. In this research, a storage design scheme has been developed using HBase and Hadoop that can efficiently integrate, store, and retrieve security-related datasets. The design scheme is a value-based data integration approach, where data is integrated by columns instead of by rows. Since rowkeys are the most important aspect of HBase table design and performance, a rowkey design was chosen based on the most frequently accessed columns associated with use cases for the retrieval of the dataset statistics. Tests conducted on various schema design alternatives prove that the rate at which the datasets are stored and retrieved using the model designed as part of this research is higher than that of the standard method of storing data in HBase. Network datasets representing DDoS attacks have been used for integration in this research. Use case requirements have been identified, which are related to the characteristics of attacker IP addresses from the integrated datasets, to generate statistical data. This statistical data was used to run the Logistic Regression (LR) classification algorithm for classifying the network traffic data into attack-related and non-attack related traffic. The Fuzzy k-Means (FKM) algorithm was also used to create clusters of attackers and non-attacks to segregate the attack-related traffic from the network datasets. The results obtained from the two algorithms show that both LR and FKM algorithms can successfully classify the network traffic datasets into attackers and non-attackers.

I. INTRODUCTION

With the advent of big data technology, many industrial problems and challenges that are related to large volumes of data are now being addressed. Many industries and companies are able to analyze and process volumes of data which was once beyond their capability. While many domains have benefited through the use of big data technologies, cybersecurity is one field that is just beginning to explore the advantages of big data analytics. The ability to detect and stop cyber attacks can make or break an enterprise (Harper, 2013). By means of big data, organizations may be able to rigorously detect threats, create more defense mechanisms and improve security.

Prior to the arrival of big data storage, most security systems have been dedicated to a single type of threat detection. SIEM (Security Information and Event Management) systems (Cardenas et al., 2013) do exist that are

capable of analyzing data from several log files, but such systems are limited to the amount of data they can handle. With systems such as Hadoop (Hadoop, 2005), cybersecurity data can now be stored in a dedicated repository which can not only accommodate more than three months of data but also combine and analyze real-time data together with historical data. Big data analytics can be run on long-term patterns and detect advanced persistent threats (APTs) that become manifest over time.

Big data analytics play an important role in detecting advanced threats and insider threats (Gartner, 2014). Monitoring systems can potentially minimize false alarms by providing smarter analytics. Data analytics can be used to assist systems in collecting internal data by merging with relevant external data to detect known patterns to stay ahead of malicious activities or intruders. Currently, 8% of major global companies (Gartner, 2014) have adopted big data analytics for one or more use cases related to security and fraud detection. Gartner predicted that within a year, this will be increased to 25% with a positive return on investment within six months of implementation. Data analysis should be intelligent and timely as anything that is delayed will lose its value, especially in the field of cybersecurity. Given that hackers are well aware of security measures and other fraud detection measures that are employed by enterprises, they are able to directly attack without any reconnaissance phase. Hence, to be always a step ahead, enterprises can use big data analytics to improve monitoring systems and detection systems with contextual data and apply smarter analytics. Data correlation techniques can be used among the high-priority alerts and monitoring systems to detect patterns and get a bigger picture on the state of security. Also, enterprises can opt for fast tuning of their rules and models to test against data streaming close to real time.

The Teradata report (Ponemon, 2013) states that the traditional methods that fall short in detecting and preventing threats can be enhanced with big data analytics. Many big data tools and techniques have emerged that can efficiently handle the volume and complexity of varied kinds of data, such as machine-generated and network-related data. Also, the results from the survey conducted by Teradata indicate, that the shortcomings of traditional solutions in detecting and preventing threats can be overcome by using big data analytics. Hence, big data systems are being part of a cyber defense strategy for every enterprise to meet the needs of complex and large-scale analytics.

A concern with the cybersecurity monitoring process is that when multiple security monitoring systems are employed and each system generates numerous log files (such as security logs, network traffic logs), there is no well-established system that can identify the relationships among these log files and integrate them. These related log files could potentially be useful for identifying attack related patterns that help in early detection of APTs or any malicious attacks. The work in (Labrinidis, 2012) identifies the challenges in dealing with big data analysis, such as automating the whole process of locating, identifying and understanding the data. A suitable database design is required even when analyzing one single dataset. Similarly, mining requires data to be integrated, cleaned and efficiently accessible, which involves the use of effective mining algorithms and big data computing environments. Labrinidis (Labrinidis, 2012) also describes that significant research is required in order to

achieve automated integration of data sets as well as a suitable database design, even for simpler analysis of a single data set. It is also essential that effective mining techniques are used to extract information from the large datasets.

1.1 Background on Cybersecurity and Big Data Computing

Many big data systems are enabling the storage and analysis of large heterogeneous data sets at exceptional scale and speed (Big Data, 2013). These systems have the potential to provide significant advancements in security intelligence by reducing the time taken for data consolidation, contextualization of security event information, and correlation of historical data for forensic purposes. Initially, data is collected at a massive scale from many internal and external sources. Then, deeper analytics are performed on the data by providing a consolidated view of security-related information. Big data analytics can also be employed to analyze financial transactions, log files, and network traffic in identifying anomalies, suspicious activities and fraud detection. In this context, the more data that is collected, the greater value that can be derived from the data. However, there are several challenges such as privacy challenges, legal challenges and technical issues regarding data collection, storage and analysis that developers have to overcome for performing potential big data analytics.

HP labs has investigated big data analytics for security challenges by introducing large-scale graph inference and analysis of a large collection of DNS events, which consists of requests and responses. Large-scale graph inference identifies malware-infected hosts in a network and maps them with the malicious domains accessed by those hosts (Big Data, 2013). This information is again validated using an existing black list and white list to identify the likelihood for the host and domain to be malicious. This experiment was conducted on billions of HTTP requests, DNS request data and NIDS (Network Intrusion Detection Systems) data sets collected worldwide and finding that high true positive rates and low false positive rates are achieved, which can be used to train anomaly detectors. Large collections of DNS events are used to identify botnets or any kind of malicious activity in the network by deriving domain names, time stamps, and DNS response time-to-live values. Classification techniques such as decision trees and support vector machines were then used to identify infected hosts and malicious domains. Although, the graph inferences used here are well suited for handling complex types of data, this approach can use a lot of space and the operations performed on these large amounts of data are possibly slow (Sherman, 2014). The research presented in this thesis has focused on integrating datasets through a unique design using HBase instead of using the traditional way of storing the datasets in HBase. This approach helps in faster retrieval of data, as anything that is delayed will lose its value, especially in the case of attack detection.

Big data analytics has the ability to correlate data from a wide range of data sources across significant time periods. This helps in reducing false alarms and improving threat detection even when mixed with authorized user activities (Virvilis et al., 2013). Also, the analytics do not have to be performed in real-time. An organization can always perform the analysis within an acceptable time and provide warnings to the security professionals about potential attacks. The work in (Virvilis et al., 2013) also describes the importance of offline analysis along with the real time data in threat detection. Although analyzing the offline data causes a delay in the attack detection, it is equally important to consider the time the attackers spend in reaching their objective. For instance, after gaining the initial access, attackers take significant time to explore the network, navigate across subnets and identify their desired location. These steps are performed as stealthily as possible to avoid any detection. Here, big data analytics plays a key role in identifying the correlation of events across large time scales and from multiple sources which are very crucial for detecting sophisticated attacks. To achieve these needs, big data analytics support dynamic

collection, consolidation and correlation of data from diversified data sources. Unlike SIEM systems, the use of big data technologies does not have any limitations to perform correlation in a given time window. In fact, it increases the scope and quantity of data over which correlation can be performed. These data correlations result in a lower rate of false positives and increase the probability of detecting the threats.

1.2 HBase Schema Design Issues for Storing Datasets

Open source projects like Hadoop and HBase are common platforms for big data solutions, where Hadoop is a cross-platform distributed file system that allows computationally independent systems to process enormous amounts of data (Big Data, 2012). HBase is an open-source, NoSQL, highly-reliable, efficient, row-oriented and expandable distributed database system. HBase utilizes Hadoop HDFS as its own storage system and runs Hadoop MapReduce to process huge datasets. It can easily store large amounts of unstructured data and is great for processing large datasets, as the Hadoop Distributed File System (HDFS) provides a reliable low-level storage support for HBase (Zhao et al., 2014).

While HBase provides lot of features and many design choices to the user, the crucial feature for best performance lies in the schema design. In schema design or table design, the emphasis is particularly given on rowkey design as the lack of secondary indexes in HBase forces the use of the rowkey for column name sorting (George, 2012). Choosing sequential keys for sequential reads are the best but provide poor performance where writes are concerned. Similarly, random keys are good for performing writes, but provide poor performance in read operations. Based on the access pattern, sequential rowkeys, random rowkeys, or even the combination of both can be chosen. Choosing a good rowkey will improve the read and write performance. HBase provides many options to choose the rowkey, such as salting, hashing, randomization and key field swapping techniques, to prevent hot-spotting and other issues, with each of them having their own pros and cons (HBase, 2006). Hot spotting is a problem in which most of the clients' requests are directed to a single node or a small set of nodes of a large cluster, keeping other nodes idle and wasting its resources. The problem of hot spotting can be eradicated by distributing the data to create a well load-balanced cluster that can be achieved by changing the rowkey design. It is possible that a particular rowkey design which provides best write performance might give the worst read performance and vice versa. Therefore, it is necessary to choose good rowkey design based on the requirements.

II. CONCLUSION

This research has explored DDoS attack datasets and developed a unique column-based design approach using Hadoop and HBase for efficiently integrating, storing, and retrieving the datasets. Use case requirements related to these datasets have been identified and the statistical data related to each unique IP address is obtained from the integrated datasets as per the use cases, which were used for running machine learning algorithms, such as the Logistic Regression (LR) classification algorithm and the Fuzzy k-Means (FKM) clustering algorithm. The LR algorithm accurately classified attackers and non-attackers from the datasets and the FKM algorithm successfully created clusters of attackers and non-attackers.

Various attacks, worms, and viruses related to cybersecurity datasets have been explored and DDoS attack datasets were finalized after the initial research. Major schema design alternatives in HBase were tested to develop an efficient HBase rowkey design scheme for storing the integrated datasets. These datasets contained

purely attack related traffic. Hence, regular network traffic from a TTU desktop machine connected to the TTU network has been captured using Wireshark and merged with attack-related traffic. The statistical data obtained from the integrated datasets contain a mixture of attack and non-attack traffic, which was successfully classified and clustered by LR and FKM algorithms as attackers and non-attackers.

BIBLIOGRAPHY

[1] Big Data: A Workshop Report. Workshop. Washington D.C: National Academic Press, 2012. Report.

[2] Bezdek, James C. Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers, 1981.

[3] Big Data Working Group. "Big Data Analytics for Security Intelligence." Cloud Security Alliance, 2013.

[4] Cardenas, Alvaro, Pratyusa Manadhata and Sreeranga Rajan. "Big Data Analytics for Security." IEEE Security and Privacy 2013. Document.

[5] Chandola, Varun, et al. Data Warehousing and Data Mining Techniques for Computer Security. Springer, 2006.

[6] Chapple, Michael J., Nitesh Chawla, and Aaron Striegel. "Authentication anomaly detection: A case study on a virtual private network." Proceedings of the 3rd annual ACM workshop on Mining network data. ACM, 2007.

[7] Chien, Eric. CodeRed Worm - Symantec Enterprise. 13 February 2007. Report.

[8] Conficker, Worm. Microsoft Safety and Security Center - Protection from Conficker Worm. November 2008. Report.

[9] Danyliw, Roman and Allen Householder. CERT - Code Red Worm Exploiting Buffer Overflow In IIS Indexing Service DLL. 19 July 2001. Report. 17 January 2002.

[10] Dark Reading. Dark Reading: Security Monitoring. 9 March 2012. Case Study.

[11] Frederick, Karen Kent. Abnormal IP Packets: Symantec Connect. 12 October 2000. Article. 3 November 2010.

[12] Gao, Neng, Deng-Guo Feng, and Ji Xiang. "A data-mining based DoS detection technique." Jisuanji Xuebao(Chinese Journal of Computers) 29.6, 2006.

[13] García, Enrique, et al. "Drawbacks and solutions of applying association rule mining in learning management systems." Proceedings of the International Workshop on Applying Data Mining in e-Learning (ADML), Crete, Greece. 2007.