Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 4 (2023)

A Review on Big Data Analytics on Cyber Security

Dr. Savya Sachi[1]
[1]Assistant Professor, Department of Information Technology
Lalit Narayan Mishra College of Business Management, Muzaffarpur, Bihar, India

Dr. Manish Kumar[2]
[2]Director
Lalit Narayan Mishra College of Business Management, Muzaffarpur, Bihar, India

Abstract- Big data analytics in the context of defence refers to the capacity to gather substantial amounts of digital data for analysis, visualisation, and conclusion-making that might aid in anticipating and thwarting cyberattacks. It strengthens our position in terms of cyber protection when paired with security technology. They enable companies to identify activity patterns that point to network dangers. In this research study, we'll examine how big data might improve information security. I concentrate on how Big Data might enhance information security in this essay. In order to analyse, visualise, and derive insights that might help forecast and thwart cyberattacks, big data analytics in security requires the capacity to collect enormous volumes of digital data. It improves our cyber defence stance together with security technology. They make it possible for businesses to identify patterns of behaviour that indicate network dangers.

Keywords— Big data, Threat Detection, Cyber Attacks, Cyber Security, Data Base Security, Cyber Protection.

INTRODUCTION

The phrase "Big Data" refers to data sets that are incredibly huge or complicated, making it difficult or impossible for traditional data set processing application software to handle them. Big data differs most significantly from traditional data in terms of scale, velocity, and variation. Volume, velocity, and variance all refer to distinct forms of structured and unstructured data, respectively. Volume refers to the total amount of data created. These days, big data is a popular issue for research in practically every sector, including cyber security. The main sources of this information are smart gadgets and social networking websites. Data is currently being created. This enormous number of malware infections in just one industry exemplifies the severity of the danger to the world economy every year. It is obvious that in light of the fast growth of cyberattacks, the cyber-security of IT systems, business networks, and online applications may not be enough. Big Data is a term used to describe data sets that are so massive or complicated that typical data set processing application software is insufficient for or unable to handle them. Big data differs significantly from conventional data in terms of volume, velocity, and variance. Volume denotes the quantity of data generated, Velocity the pace at which the data is produced, and Variation the categories of organised and unstructured data. Big data is becoming a hot issue for research across practically all disciplines, notably cyber security. Social media websites and mobile devices are the primary sources of this data creation. Since data is being generated at such a rapid rate, many people are concerned about the security of the newly created data. It is crucial to keep this data secure because it contains sensitive information like credit card numbers and bank account numbers. Additionally, improvements in big data analytics offer ways to collect and use this data, making privacy infractions simpler. As a result, in addition to creating Big Data technologies.

LITERATURE SURVEY

In a last decade, several researches have proposed many solutions for mitigating security threats on big data environment based on different techniques. introduced a non-machine learning technique for DoS detection change-point detection and activity profiling. The author provided a mechanism to identify DoS flooding attacks which promised in limited testing, but not help in solving the detection problem. It proposed a collaborative detection technique using change ag- gregation trees for improving the detection rate. it presented two information metrics, i.e., information distance metrics and generalized entropy metrics. Both are used to identify

low rate security attacks. But their solution requires to control all router network which is very difficult to achieve. Most of the past researchers have focused on differentiating of various attacks and proposed solutions one by one. It is very difficult to detect the approaches mentioned above if an enterprises having many applications have been attacked by various different security tools. To overcome such problems many researches has started focusing on machine leaning technique. Many machine learning theories have been proposed over last few decades, the flow based methods such as Navie Bayes, K-means, SVM, KNN etc had been already used in traffic classification. Many packet processing tool for web traffic analysis are available such as WireShark, Nmap, netsniff-ng, tcp dump and snort. In general, the above mentioned packet processing tools have restriction that they run a single machine with limited number of storage and computing resources. Also as it used on a single machine thus has a chance of high fault tolerance services such as a node failure which regularly happen when read or write operations are performed repetitively on disks. Researchers have progressively moving from SIEM based method to Big Data approaches like in developing DDoS attack system against Big Data center using correlation analysis. In this the author make use of the correlation information of training data to enhance the classification accuracy and decrease the overhead caused by the weight of training data. In, authors have proposed the first packet processing technique for Hadoop which uses map reduce to examine packet trace files by investigating packets across multiple HDFS blocks.Researchers have developed a new traffic monitoring system which performs flow analysis on a web traffic of terabytes in a very scalable manner.They build a Map Reduce algorithm in a different format which is capable to handle libpcap files in parallel manner. However there's limitation related to this technique.

ANALYTICS OF BIG DATA

"Big data" is frequently used to describe "high-volume, high-velocity, and high-variety information assets that require cost-effective, creative information processing for increased insight and decision making," according to Gartner. Big data might mean many different things, but to me, Gartner's definition made the most sense. One noteworthy omission from this description is visualisation. The significance of visualisation to me comes from the reality that our capacity to interpret data is frequently outpaced by data processing. This is brought on by the dearth of useful knowledge. The majority of big data analytics techniques have been around for a while, including machine learning, statistics, predictive analysis, behavioural analysis, and others. These methods have historically been used to structured data collections with sizes between a few MB and a few GB. They can now handle petabyte-sized amounts of data, both organised and unstructured. The reasons for this quick transition include the following: Data quantity (Number) - Size: Because it affects how much data is created, the amount of datasets is an important consideration. Data size (Number): The size of datasets is crucial since it determines how much data will be handled. Speed (pace of data production and transmission)-Complexity (structure, behaviour, and permutations of datasets)-Complexity (rate of data generation and transmission). Massive amounts of data that are exchanged and stored in computer systems are referred to as "big data." Three factors set big data apart from conventional technology:

(i) The volume of the data Size- The volume of datasets, or the amount of data collected, is an important consideration.

(ii) Data complexity- the organisation, behaviour, and permutations of datasets have a crucial role in determining the velocity (or pace) of data collection and transmission.

(iii) The many forms of organised and unstructured data Technologies- The tools and methods that have been employed to analyse large or complicated datasets are an important consideration.

LANDSCAPE OF THE CYBER ATTACK

More than 430 million new malware pieces were found in 2015, according to a security report (Internet Security Threat Report, 2016), and what's even more amazing about this is that the experts weren't surprised by their discovery. According to the security assessment, targeted, sophisticated, and persistent assaults on government agencies and companies of all sizes are on the rise and pose a severe danger to the country's economy and security. According to a security study (Internet Security Threat Report, 2016), over 430 million new malware pieces were discovered in 2015. What's more astonishing is that the experts were unsurprised. The security study continues by stating that targeted, intricate, and persistent assaults on institutions of higher learning and

businesses of all kinds are an increasing issue. Simple viruses to more sophisticated malware, such as cyber weapons, are all examples of different forms of cyberattacks. Virvilis and Gritzalis (2013) list the following qualities of an APT: They often have an initial attack vector, such as malicious office documents or removable devices; they are typically targeted at specific and high-value targets, and hence for certain operating systems or platforms; The encryption of network traffic is one of their evasive techniques, and they have a toolkit of evasion strategies to get around anti-virus software and intrusion detection systems (IDS), which implement command and control mechanisms. They also use stolen but genuine digital certificates to fool the targeted systems into believing they are secure. Examples of external attackers include crime gangs, state-sponsored hackers, hacktivists, and lone hackers. Highly complex malware, which is difficult to detect even with equally sophisticated security systems, is one of the attackers' attack methods. As an illustration, it was suggested that the malware employed in the Sony assault may have evaded most network safeguards.

BDA-BASED THREAT DETECTION

The three sorts of threats posed by big data are monitoring, disclosure, and discrimination. The term "surveillance" describes the experience of being watched as a result of the collection, aggregation, and/or use of one's data. Being aware that you are being observed could be a problem in and of itself, comparable to mental anguish. It could also be a problem since people may start to second-guess what they do, read, or search for as a result of these sentiments. Information that has been made public outside of the circumstances under which it was obtained. A inquisitive employee who searches a company database for people he knows is one potential cause of information exposure. Another scenario is an identity thief who successfully hacks into a database. This is how insecurity problems. According to the definition of discrimination, someone is treated unfairly depending on data collected about them. People are threatened by discrimination in many different ways. The most apparent instance would be making assumptions about someone's membership in a protected group like race or religion and then practising discrimination as a result. A protected feature may also be the basis for objections from certain persons to any discrimination. The conventional methods for identifying and thwarting cyber-attacks include antivirus software, network and host IDS/IPS, network device incidents, logging, FIM and white listing, and SIEM. These systems have numerous advantages, but they also have certain drawbacks. is proven to be mostly ineffectual against the sneaky cyber-attacks of today. This is because these systems create a lot of data that is difficult and time-consuming to examine without the right technology, making it easy to overlook important cyber-attack events. These systems also operate independently of one another. This suggests that the proper adoption of the appropriate instrument that can quickly sort through data might increase the effectiveness of these diverse networks.

BIG TECHNOLOGY TRENDS

Along with analytics and cloud-based technology, big data is receiving a tonne of attention from industry, the media, and even consumers. All of them are a component of the contemporary eco-system that technological megatrends have produced. Big data has emerged as a dominant issue or theme in the technology media. It has also been used into several compliances and internal audits. According to 72% of participants in EY's Global Forensic Data Analysis Survey 2014, developing big data technologies can be crucial in the prevention and detection of fraud. However, just a small percentage of respondents about 7% knew about any specific big data technologies, and only a very small percentage about 2% were really employing them. FDA (Forensic data analysis) solutions are available to assist businesses keep up with the pace of rapidly growing data quantities and organisational complexity. The top ten developing technologies are assisting users in coping with and handling Big Data in a cost-effective manner. Big Data is vast and incorporates numerous trends and new technology advances.

(i) Column oriented database- Traditional, row oriented database are excellent for the online transaction processing with the high update speeds, but they fall short in the query performance as more data volume grows and as data becomes unstructured.

(ii) Schema less database or No Sql database - there are various database types that fit into this category, such as key value storage and document stores, which focus on storage and retrival of large volume of data which is either unstructured, semi-structured, or even structured data.

(iii) Map Reduce - This is a programming paradigm that allows for massive job execution scalability against thousands of servers or clusters of servers. Any Map Reduce implementation consists of two tasks: The "Map" task, where an input dataset is converted into a different set of key/value pairs, or tuples. The "Reduce" task, where several of the outputs of the "Map" task are combined to form a reduced set of tuples

(iv) Hadoop – c Hadoop is the best and the most popular implementation of map reduce, being an entirely an open source platform for handling of big data. It is flexible enough to be able to work with multiple data sources. It has several different applications, but one of the top use cases is for large volumes of constantly changing data, such as location- based data from weather or traffic sensors

(v) Hive - It is a SQL-LIKE bridge that allows conventional BI application to run queries against a Hadoop cluster It was developed originally by Facebook, but has been made open source for some time now, and it's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store.

(vi) Pig- PIG was developed by Yahoo .PIG is bridge that tries to bring Hadoop closer to the realities of developers and business users, similar to Hive. Unlike Hive, however, PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, instead of a SQL-like language.

(vii) WibiData- Wibi data is a combination of web analytics with hadoop it is been built on the top of Hbase which itself a database layer on hadoop.

(viii) Sky Tree- It is a high performance machine learning and data analytics platform focussed specially on the handling of big data. machine learning is a very important part of big data, since the data volume make manual exploration.

CYBER BIG DATA ANALYTICS SECURITY

Fraud Detection Using Big Data Analytics The two main categories of fraud detection methods are statistical methods and artificial intelligence methods. Techniques for data pre-processing that identify, validate, fix errors, and fill in gaps in data are examples of statistical data analysis techniques. the computation of a variety of statistical characteristics, including averages, quintiles, performance measures, probability distributions, and so on. models and probability distributions for different business operations, either in terms of different parameters or probabilities of the creation of user profiles. examination of time series involving time-dependent data. finding patterns and correlations between groupings of data via clustering and classification. comparing algorithms to find discrepancies between user or transaction activity and established models and profiles. Techniques are also required to detect false alarms, calculate risks, and forecast future behaviour of present users or transactions. Management of fraud requires a lot of expertise. Analogy-based intrusion is detected via big data analytics Algorithms for anomaly detection are fairly easy to set up and work automatically. Thresholds are set once some key performance indicators are selected for an event. An incident is flagged for further inquiry if a threshold is surpassed. The selection of the monitored indicators, the analysis time, and the threshold value settings all have an impact on the method's performance. Algorithms for anomaly detection are fairly easy to set up and operate automatically. The selection of the monitoring parameters, the analysis time, and the threshold value settings all have an impact on the method's efficacy. Information on security They can produce actionable security responses and shorten the time needed to correlate data for forensics purposes.

CONCLUSION

Big Data analytics for security seeks to gather current, useful intelligence. Big Data may significantly impact your present organisation in three different ways. You will benefit in the following ways from it: Although they make it possible to retain data in a single "window" of events, they need a lot of computing effort to evaluate these "windows." Additionally, systems built on the actor model required more memory because each entity's data had to be repeated, although using less computing resources. Solutions built on modified MapReduce played a middle ground role as a consequence. Cybersecurity and real-time big data analytics have become major concerns.The amount of data being generated rises exponentially every day, making it increasingly susceptible to cyberattacks.This study explored several large data threats, their storage methods, and real-time analytics techniques for cyber security.There is still more work to be done, despite the numerous innovative solutions developed by researchers for detecting security vulnerabilities in big data environments. Extraction of the necessary characteristic from huge data, which is often in semi-structured or unstructured format, is a highly

challenging operation.Data in social networks is enormous and subject to graph-based assaults.Social network data is anonymised via the addition of dummy vertices and edges to protect against such assaults.However, adding or removing fake data has an impact on the social network's neighbourhood attributes.As a result, create noise in the social network data, which degrades the social network's originality and causes information to be lost.Later, writers could concentrate on minimising the amount of noise by including the optimal number of fake edges in order to reduce information loss.

REFERENCES

[1]. Wang, L. and Jones, R., 2020. Big data analytics in cyber security: network traffic and attacks. Journal of Computer Information Systems,
[2]. Kantarcioglu, M. and Xi, B., 2016, October. Adversarial data mining: Big data meets cyber security. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security (pp. 1866-1867).
[3]. Angin, P., Bhargava, B. and Ranchal, R., 2019. Big data analytics for cyber security.
[4]. Mahmood, T. and Afzal, U., 2013, December. Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools. In 2013 2nd national conference on Information assurance (ncia) (pp. 129-134). IEEE.
[5]. Mahmood, T. and Afzal, U., 2013, December. Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools. In 2013 2nd national conference on Information assurance (ncia) (pp. 129-134). IEEE.
[6]. Alguliyev, R. and Imamverdiyev, Y., 2014, October. Big data: Big promises for information security. In 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-4). IEEE.
[7] J. Gantz and D. Reinsel, “The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east,” IDC iView: IDC Analyze the future, vol. 2007, no. 2012, pp. 1–16, 2012.
[8] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, “Big data: The next frontier for innovation, competition, and productivity,” 2011.
[9] D. Ganesh Chandra, “Base analysis of nosql database,” Future Gener. Comput. Syst., vol. 52, no. C, pp. 13–21, Nov. 2015.
[10] C.-M. Chen, Y.-H. Ou, and Y.-C. Tsai, “Web botnet detection based on flow information,” in Computer Symposium (ICS), 2010 International. IEEE, 2010, pp. 381–384.
[11] S. Gupta, A. Singhal, and A. Kapoor, “A literature survey on social engineering attacks: Phishing attack,” in Computing, Communication and Automation (ICCCA), 2016 International Conference on. IEEE, 2016, pp. 537–540.
[12] J. H. Abawajy, M. I. H. Ninggal, and T. Herawan, “Privacy preserving social network data publication,” IEEE communications surveys & tutorials, vol. 18, no. 3, pp. 1974–1997.
[13] N. K. Singh and D. S. Tomar, “Privacy preservation of social media services,” Exploring Enterprise Service Bus in the Service-Oriented Architecture Paradigm, p. 236, 2017.
[14] I. Aljarah and S. A. Ludwig, “Mapreduce intrusion detection system based on a particle swarm optimization clustering algorithm,” in 2013 IEEE Congress on Evolutionary Computation, June 2013, pp. 955–962.
[15] A. W. Moore and D. Zuev, “Internet traffic classification using bayesian analysis techniques,” SIGMETRICS Perform. Eval. Rev., vol. 33, no. 1, pp. 50–60, Jun. 2005.
[16] G. Carl, G. Kesidis, R. R. Brooks, and S. Rai, “Denial-of-service attack- detection techniques,” IEEE Internet Computing, vol. 10, no. 1, pp. 82–89, Jan 2006.
[17] Y. Chen, K. Hwang, and W. S. Ku., “Collaborative detection of ddos attacks over multiple network domains,” IEEE Transactions on Parallel and Distributed Systems, vol. 18, no. 12, pp. 1649–1662, Dec 2007.
[18] Y. Xiang, K. Li, and W. Zhou, “Low-rate ddos attacks detection and traceback by using new information metrics,” IEEE Transactions on Information Forensics and Security, vol. 6, no. 2, pp. 426–437, June 2011.
[19] J. Erman, M. Arlitt, and A. Mahanti, “Traffic classification using clustering algorithms,” in Proceedings of the 2006 SIGCOMM workshop on Mining network data. ACM, 2006, pp. 281–286.

[20] A. Este, F. Gringoli, and L. Salgarelli, “Support vector machines for tcp traffic classification,” Comput. Netw., vol. 53, no. 14, pp. 2476–2490, Sep. 2009.
[21] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, “Class- of-service mapping for qos: A statistical signature-based approach to ip traffic classification,” in Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement, ser. IMC ’04. New York, NY, USA: ACM, 2004, pp. 135–148.
[22] P. Xiao, W. Qu, H. Qi, and Z. Li, “Detecting ddos attacks against data center with correlation analysis,” Comput. Commun., vol. 67, no. C, pp. 66–74, Aug. 2015.
[23] Y. Lee, W. Kang, and Y. Lee, “A hadoop-based packet trace processing tool,” in Proceedings of the Third International Conference on Traffic Monitoring and Analysis, ser. TMA’11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 51–63.
[24] J. Ortiz-Ubarri, H. Ortiz-Zuazaga, A. Maldonado, E. Santos, and J. Grulln, “Toa: A web based network flow data monitoring system at scale,” in 2015 IEEE International Congress on Big Data, June 2015, pp. 438–443.
[25] T. F. Yen and M. K. Reiter, “Are your hosts trading or plotting? telling p2p file-sharing and bots apart,” in 2010 IEEE 30th International Conference on Distributed Computing Systems, June 2010, pp. 241–252.