A REVIEW OF BIG DATA ANALYTICS IN CYBER SECURITY

Abstract

The linking of people's personal data poses a serious danger to their right to privacy and civil liberties. To meet these problems, it's crucial to develop new security solutions and tactics for better security operations, threat detection, and attack analysis. The methods must be improved in light of the current cyber security risks. The internet is experiencing a massive increase in data. Because of the large amount of data, cyberattacks will likewise grow exponentially. In order to recognize specific dangers and assaulting patterns, it is crucial to interpret and visualize it. Big data analytics aids in monitoring a broad range of user behaviors to prevent various risks connected with them. This helps prevent numerous data breaches. In this article, the significance of big data is examined, emphasizing how it may be used as a tool in cyber security to help certain security operations. Additionally, a thorough overview of Big Data Analytics in the subject of cyber security is discussed.

Keywords: Big data Analytics, Big data query Data analytics, Privacy, threat, Big data security, cyber security, visualization, anomaly

Author

Dr. Susheel George Joseph
Associate Professor
Department of Computer Application
Kristu Jyoti College of Management and Technology,
Changanassery, Kerala, India
susheel@kjcmt.ac.in
susheelgj@gmail.com

I. INTRODUCTION

In the recent years, a variety of applications have produced data at a rapid rate, resulting in the generation of Big Data, or extremely large amounts of data. The phrase "big data," which refers to amounts of data that are generated in various forms at a tremendous rate, was adopted in recent years as a result of developments in internet services and other communication systems. Big data analytics has the ability to process these massive amounts of data. Utilizing data gathered from networks, the cloud, computers, and other devices can assist in identifying system exposure and responding appropriately.

Big Data is the term used to describe a vast amount of data that includes both structured and unstructured data. Big data is typically used to improve customer service and offer more protection for client data. Additionally, employing big data enables an organisation to make business choices more quickly and thoroughly [1]. Major concepts of Big data are generalised into five V's are: volume, velocity, variety, veracity, value.

- Volume is the amount of data that is been generated.
- Velocity can be defined as the rapidness of data from various origins.
- Variety refers to the diversity or it can be defined as different types of available data such as structured, unstructured and semi-structured data.
- Veracity defines the accuracy of data. It is not at all about the quality of data but about its reliability.
- Value can be referred as the benefit that big data can provide.

Big data is defined as too complex to process and analyse. It enables many people to overcome the problems that are associated with small samples of data. There is a need to find new possibilities for accommodating these data because it is growing day after another in an exponential manner. Two major modes of Big data analytics are: verification and identification. In verification, data analyst already has an assumption about certain property of services that he wants to verify by means of data analytics

In identification, data analyst gathers a large dataset potentially from multiple sources and tries to identify interesting facts hidden within the dataset. Two key concepts of aggregation of datasets within the big data content must be defined –

The first is the aggregation of schematically identical datasets. For example, joining the service access logs of two different online services that are saved on the same web server implementation. mostly wed to attain more information within an existing content. Second type of aggregation is about linkage created from joining two datasets from disjunction contents, based on some key information shared in both datasets to be aggregated. A key challenge of big data analytics consists in identifying linkage – a link can be a user email addresses, postal codes/combinations of IP addresses and timestamps. The identity of service plays a major role. This linkage via user identify bear some very challenging pitfalls in the field of privacy. The skill of handling, storing, and gathering massive amounts of data is known as bigdata analytics.

The primary focus of big data is the identification of abnormalities and attacks. It allows studying organised and unstructured data like documents, photographs and videos which are utilised as digital evidence in computer forensic process.

It becomes much harder to safeguard data as it grows in volume. When thinking about how to protect massive data, confidentiality is by far the most crucial factor. Big data analytics is utilised as a tool for all business and organisational possibilities and for every type of data.

Hadoop is one of the crucial instruments that enhances method processing. They are managing the characteristics of enormous amounts of company data using this technique. Combining Hadoop with Revolution Analytics offers benefits that help businesses meet their need for making strategic decisions. Hadoop divides and stores data in various devices, saving a copy of each dataset in each device. To put it another way, Hadoop is the name given to the massive amount of data that is spread into large data sets across a large number of cheap servers. It is operated in parallel.

Cyber security is the process of protecting user's data from unauthorised access, attacks or damages. Cyber security has now gone beyond the traditional way. Big data has unfolded new ways for cyber security sector.

The cyber security where big data analytics can contribute: - Forensic focuses on the analysis, preservation and interpretation of computer data. This field deals with a large dataset, we use various conceptual models for forensic analysis in order to remove redundant data. By applying visualisation technique we can reduce the time and improve the effectiveness to find suspicious files [2].

Big data solutions provide two essential approaches so that the analyst can make his search in abundant data easier [3]. First one is an integrate information from different sources and second has customised visualisation tools.

1. **Malware detection:** we use big data for malware detection. These are the methods for classifying, combining Big data analysis with machine learning, binary instrumentation and dynamic instruct flow analysis.

2. **Security offence:** Security offence include cyber description threat hunting and attack detection. Nowadays it is motivated to use artificial intelligence, game theory and big data to enhance cyber security strategies against attackers. The main objective of the cyber description is to detect attacks such as:

3. **Threat hunting:** It is an active defence searching. It is an iterative activity to check through hardware and detect threats in advance instead of waiting for attack alerts. By using big data solution, processing of large amount of information generated by logs can be handled.

4. **Attack detection:** It is very important to detect attacks in the shortest time if possible. It will reduce the time between detection and attack response. Even though big data enhances security, on the other hand Big data gives a great chance not only for the development of an organisation but also for cyber criminals because they have much

more to achieve when they track such a huge volume of data. There are various algorithms and analytics used to find out information [4]. These algorithms are also applied based on the nature of the data. Some examples for this kind of algorithms are:

- Apriori Algorithm and Naive Bayes Classifier Algorithm. Apriori algorithm works on the principle of bringing frequent data variables, then extending them to larger as long as they are frequent in nature.
- Naive Bayes Classifier Algorithm based on Bayes Theorem. It is a classification algorithm with assumptions of independence among predictors. This model is easy to build and work very well for large datasets.

Data mining is also an important process when it comes to big data analytics. It processes large, pre-existing data. It is used for find measure detection and also anomaly detection.

II. RELATED WORK

The authors in [1] mentioned big data definition and how it is useful for the development of an organisation. The usage of Big Data Analytics for enterprise data which is the data generally shared by users of an organisation. Their main objective is to access unstructured data from all extreme, and to convert processed data to structured form so that the process of accessing is easier. For the easier protection and storage of Big data many organizations use tools like Hadoop which distribute and stores the huge data efficiently by using the method of parallel processing. This method is an efficient and best method for Big Data Analytics because it is less expensive since the data are distributed to inexpensive servers and it is less time consuming. Big data is described in a way that it increases data processing efficiency. Here various authors enumerate the major differences between traditional and big data Analytics. This technique is divided into Batch processing and stream processing. In this paper various authors mentioned the desire to build different platforms to store and analyse data. The process is partially enriched and partially illustrative. The authors [2] enumerate about the rapid growth of data. Contribution of smart devices, such as smartphones hand held computers, wireless networks and social media generating more data over past few years. In social media domains such as Facebook, more than 30 million users are updating posting and sharing their images and video per minute. Like in Instagram, also 300 million Instagram users share more than 60-million photos every day. More than 100 hours of video are uploaded in every minute. This huge enormous data is Big Data and there is a need to protect and secure these data from & unauthorized access. This Big data allows new possibilities in technology as well as in research field.

The importance of big data analytical techniques to overcome cyber security threats. They show various technique to interpret, mine and visualise big data from different sources so that it can be applied in cyber forensic, cyber security and threat intelligence [3].

Authors focuses [4] on big data applications and threat intelligence. They also show various research topics on big data for future research which includes anomaly detection for big data, big forensic data provenance, analysis of big data for cyber intelligence, advanced persistent threats detection, big data analytical technique for cyber defence, big data forensic data management and reduction.

Mazel et al. [5] shows various challenges related to big data analytics on privacy. The authors proposed that data erosion in terms of privacy and user's rights may due to the upcoming trend in big data analytics. He proposed various fields of research on privacy in big data analytics. The most challenging part of privacy in big data analytics is that to provide transparency of personal data of the individuals with respect to type of processing. It is always necessary to process information bound to an individual. Informed consent means that there are many types of big analytics based on complex data algorithm, so each Individual must be given an explanation of all these algorithms so that they can understand what is happening there, this is a big challenge to data analysis.

An individual decides to revoke the consent for processing personal data later. This is similar to getting a person used among various data collectors and data analysts that is not easier to stop processing on these data’s and to delete it. This has become a highly challenging issue. There are various types of attacks such as targeted identification attacks, correlation attacks and arbitrary identification attacks. Most threatening type of attack is targeted identification attack. It is to identify some more details of an individual. In order to create more unique database entries, we link a dataset of uniform data values to other sources. Correlation attacks consist of this kind of linking form datasets. There datasets contain more information per User ID. This helps in analysing more on individual. Arbitrary identification attacks show failures of a set of anonymized data. This type of attack link at least to one entry of the dataset to identify a human individual. A threat to big data analytics is if the information gathered is valid or not. Various types of results can be formed. It will depend on the type of query used by a big data analyst.

Results from different big data query sometimes become a completely wrong final statement. A lot of threats to privacy can also arise from economic consideration in such data trading economic issues of the big data, paradigm is considered to be the fourth category of threats. So threats can be caused due to intentional attacks. It can also cause due to false data processing methodology or caused by interaction with concerned individuals. So field of privacy in big data faces a lot of challenges.

Author [6] explains about the spontaneous growth of the internet has resulted in the exponential increase of the number of cyber-attacks. Many organisations tried many popular cyber securities to prevent these attacks. Also, the introduction of Big Data made internet with enormous amount of data. To regale this issue, many researches are now focusing on Security Analytics, which is one of the important application of Big Data Analytics techniques to cybersecurity. This paper provides a survey on the art of Security Analytics which including its states such as its description, trends, technology and tools.

In the paper [7] presented challenges to privacy of Individuals. The paper discusses about various set of challenges that may threaten privacy of individuals. Another threat with respect to privacy in big data analytics is the ability to perform "re-identification attacks", also validity of the result gathered is also a threat. Another threat covers the economic issues of big data paradigm.

In the paper [8] proposed that high-precision, robust, lightweight and identification and understanding of technology is very important. It will be the direction of future research. Big data based on cloud computing technology will become a major trend. Difficulty of the new media big is because recognition and understanding of new media content is difficult. To create a healthy innovative new media environment, we need to research how we can safely provide, consume data and dig information faithfully from these data's.

In the paper [9] proposed that malicious and suspicious patterns can be identified by network managers particularly in the surveillance of real-time network streams. They show the survey on the art of security Analytics. Also the authors proposed that cyber application of analytics will become an imminent part in cybersecurity in the future. They mentioned different types of big data sources for analytics solution.

In the paper [10] proposed that big data consist of structured, semi-structured and unstructured data. They show the methods. to analyse the audio, video and text. They show different challenges faced by researches while performing big data analysis They also discussed various big data analytics methods and techniques.

III. CONCLUSION

This paper contains a detailed review on Big Data Analytics in Cyber Security sector. Big data is a new alternative to improve security operations. It has the ability process voluminous data in different format in short time. It is applied to monitor operations and detection of anomalies. Moreover, it is used in protective strategies such as threat hunting on cyber deception. It can also detect attack patterns by processing immense data from heterogeneous source.

Big Analytics is often used in cyber security lots of reasons. It facilitates the working of an organization easier by increasing security with the use of various algorithms and techniques. The main objective of Big Data analytics is to generate a safe environment for users to protect their data from unauthorised access attacks.

REFERENCES

[1] Poonam Vashisht , Vishal Gupta ,“ Big Data Analytics Techniques: A survey", International conference on Green Computing and Internet of things (ICGI0T), 2015

[2] Aviral Apurva, Pranshu Ranakoti, Saurav Yadav, Shashank Tomer, Nihar Ranjan Roy , “Redefining Cyber Security with Big Data Analytics", International Conference on Computing and communication technologies for Smart Nation (I c3TSN) 2017.

[3] Kim -Kwang Raymond C+ hoo ,Mauro Conti, Ali Dehghantanha, “special issue on Big Data Application in Cyber Security and threat intelligence – part 1”, IEEE transaction on Big Data , July – September 2019

[4] Kim -Kwang Raymond Choo, Mauro Conti, Ali Dehghantanha “ Special Issue on Big Data Application in Cyber Security and threat intelligence – part 2” ,IEEE Transaction on Big Data , October – December 2019

[5] Fontugne R Mazel I and Fuhada K. Hashdoop "A MapReduce framework for network anomaly detection “IEEE conference on work shops (2014)4] Meiko Jensen "Challenges of Privacy Protection in Big Data Analytics” IEEE International Congress on Big Data ,2013

[6] Dr. Tariq Muhammed, Uzma Afzal ," Security Analytics Big Data Analytics for Cyber Security", 2nd National Conference on Information Assurance (NCIA) ,2013

[7] M. Jensen, "Challenges of Privacy Protection in Big Data Analytics," 2013 IEEE International Congress on Big Data, 2013, pp. 235-238

[8] Esfahani, H., Tavasoli, K & Jabbarzadeh, A. (2019). Big data and social media: A scientometrics analysis.International Journal of Data and Network Science, 3(3), 145-164.

[9] D. B. Rawat, R. Doku and M. Garuba, "Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security," in IEEE Transactions on Services Computing, vol. 14, no. 6, pp. 2055-2072, 1 Nov.-Dec. 2021, doi: 10.1109/TSC.2019.2907247

[10] Neha Srivasta , prof. Umesh Chandra Jaiswal ,“ Big Data Analytics Technique in Cyber Security- A Review”, proceedings of third international conference on Computing Methodolgies and Communication (ICCMC 2019)