INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2037 – 2042
RECEIVED: 20-03-17. PUBLISHED: 09-04-2017

April 09, 2017

LITERATURE REVIEW ON BIG DATA TECHNIQUES FOR SECURITY

[1]G.S.Gunanidhi, [2]R.Sethil,
[1]Research Scholar, Hindustan University,
[2]Assistant Professor, Valliammai Engineering College

ABSTRACT

The data is growing day by day to a larger extent. Recently, Big Data Analytics has become a hot topic in academics, industry and everywhere. Big Data Analytics is the process of examining large amounts of data (big data) to discover hidden patterns or unknown correlations. A key part of big data analytics is the need to collect, maintain and analyse enormous amounts of data efficiently. Due to increase in number of sophisticated targeted threats and rapid growth in data, the analysis of data becomes too difficult. Today's attacks are prepared by advanced technologies are not detected until the damage has been occurred. Big Data Security Analytics is important to mitigate the security threats to secure the data more efficiently. In Big Data Analytics, Data Security is a challenging task to implement and calls for strong support in terms of security policy formulation and mechanisms. In this paper we have discussed the analysis of the Big Data Analytics concepts and some existing techniques and tools, like Hadoop, for data Security.

Keywords: Analytics, Big Data, Data Security, Hadoop, Threats.

1. INTRODUCTION

In the current scenario, Web and its associated entity, Internet, a shadow has been cast on the same with the data explosion that has taken place in the last couple of years considering the interaction that has been taking place between people and systems associated at multiple touch points. This huge entity which is taking place at every touch point as mentioned above in its wholesome behavior is known as Big Data. Some decades earlier, Kilobytes and Megabytes used to be entities, which used to combine the entire definition of data existing on the planet, and due to continuous interactions between people and systems that have been taking place which has lead to exponential growth of data due to which new terms such as Gigabytes, Terabytes, Petabytes, Exabytes & Zettabytes have graced the steps of the computing world. Theorists and Researchers have propagated this that as Moore's law was to growth of transistors inside the circuits, Data in Internet would exceed the entire brain capacity of the living species. Technological advances have been taking place continuously across all the domains and the major reasons for it are advances in digital sensors, computation, communications and storage that have created humongous collection of data. As explained above, data is generated through various sources which will be used by multiple organizations to run and understand the various business scenarios which help them understand and run their business. All the above data when analyzed through various sources and methods of data analysis help organizations in studying customer behavior, interpreting market trends and taking strategic and financial decisions. When we define the term, as we have done above, we often forget to define that the same consists of the Big Datasets, which cannot be managed efficiently by common database

management systems often denoted by Relational Database Management Systems (RDBMS) and these datasets often range from Exabytes to Petabytes to Zettabytes.

The huge amount of data that we have been speaking about, that is created from multiple interactions, across Cellular phones, credit cards, social networking platforms and RFID (Radio Frequency Identification) devices and it is not necessary that all this data may be used and hence most of this data resides at unknown servers, in unstructured and unutilized form for many years. However, in the current scenario that we have been speaking of and with the evolution of the Big Data the same data can be accessed and analyzed to generate useful insights. According to the Big Corporations, that have been the at the platform of Information Technology, we create quintillions of data in a single day alone, or as explained earlier, 90% of the data that is existent in the world today, has been created in the last two years. There is no single source of the data, as it comes from multiple sources and to name a few, we can easily target it to Cell phone and associated GPS signals, digital pictures and videos, transactional records of purchase and selling, Social media activities, sensors used to gather climate information. In the current scenario of Data existence, it is everywhere and anywhere and in every possible format i.e. numbers, images, videos and text. Data, for which we have been speaking about earlier, in all its beauty, has had an exponential pace of growth, but this humongous collection of data has numerous critical issues associated with it and challenges which can often be put in nomenclature as transfer speed, diverse data, security issues and rapid data growth.

What Is Big Data: When we speak about Big Data, as we have done above, we often identify it as a jargon, catch phrase which means a exponential volume of unstructured and structured data that contains so many huge datasets which cannot be processed by traditional database management techniques and associated software techniques. With the size of the big data and simply the capacity of the data that it encompasses, it carries in itself the potential that will help companies, in making far better, intelligent and data driven decisions and help in improving operations. For most of the organizational scenarios, it can be easily identified either the data is in excess of the current storage and processing capacity, or the volume of the data is too big or it moves too fast. To give insights using the same data, that we have spoken about earlier, it has to help us in giving insights which would help us in gain competitive advantage, increasing revenues and customer retention and for that we need to capture the data, clean the data, format, manipulate, store and analyze the same. Big Data is a concept and a concept can have various interpretations, for which the same topic can have multiple definitions: Big Data is the amount of data beyond the ability of technology to store, manage and process efficiently (Manyika et.al, 2011). Big Data is a term which defines the hi-tech, high speed, high-volume, complex and multivariate data to capture, store, distribute, manage and analyze the information (TechAmerica Foundation, 2014). Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization (Gartner, 2014; Gürsakal, 2014).

Big Data Technologies are new generation technologies and architectures which were designed to extract value from multivariate high volume data sets efficiently by providing high speed capturing, discovering and analyzing (Gantz and Reinsel, 2011). Hashem et.al. Define Big Data by combining various

definitions in literature as follows: The cluster of methods and technologies in which new forms are integrated to unfold hidden values in diverse, complex and high volume data sets (Hashem et.al. 2015). As per the definitions, Data should be complex and increasing in multiplicity inclusive of its size. Simply considering the size of the data gives us enough oversight to understand that conventional methods would not be suitable in analyzing big data sets and to compensate for the same, new methods and technologies are needed. Aforementioned points should be taken into consideration, while going for the analysis of Big Data.

EXISTING BIG DATA ANALYTICS TECHNIQUES AND TOOLS FOR SECURITY

A. Nada Elgendy analyses some of the different analytics methods and tools which can be applied to Big Data (BD), as well as the opportunities provided by the application of Big Data Analytics (BDA) in various decision domains. Big Data tools, techniques, and governance processes can increase the prevention and recovery of fraudulent transactions by dramatically increasing the speed of identification and detection of compliance patterns within all available data sets. It also discussed about some of the different advanced data analytics techniques. BD,as well as its characteristics and importance has been discussed in . BD is largely untagged file-based and unstructured data, about which little is known .This means not only that large quantities of potentially useful data is getting lost. B. Bhawna Gupta proposes the use of BDA for analysing the enterprise data. The main focus is to gather the unstructured data from all the terminals, processed the data to convert into structured form so that accessing of the data would be easier. BDA describes the simple algorithm for large amount of data without compromising performance. Hadoop is one of the tools which are aimed to improve the performance of data processing. In this approach they are managing the Big Data characteristics of large volumes of enterprise data. If enterprise has an unmet business need for strategic decision making with a high degree of processing, a Revolution Analytics and Hadoop combination offers significant opportunity to gain advantage . Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel. C. Weiyi Shang et. al . describes a first step in assisting developers of big data applications BDA Apps for cloud deployments. It proposes a lightweight approachfor uncovering differences between pseudo and large-scale cloud deployments. Using injected deployment faults; they have shown that their approach is not only significantly reduces the deployment verification effort, but also provides very few false positives when identifying deployment failures. It proposes an approach for verifying the runtime execution of BDA Apps after deployment. The approach abstracts the platform's execution logs from both the small and large scale cloud deployments, groups the related abstracted log lines into execution sequences for both deployments, then examines and reports the differences between the two sets of execution sequences. The Authors specifies that the larger data and more complex environments lead to unexpected executions of the underlying platform. Such unexpected executions and their context cannot be easily uncovered by traditional approaches. In this paper, they propose an approach to uncover the different behaviour of the underlying platforms for BDA Apps between runs with small testing data and large real-life data in a cloud environment. To evaluate the approach, they have performed a case study on Hadoop, a widely used platform, with three BDA Apps. BDA Apps are a new category of software applications that leverage large-scale data, which is typically

too large to fit in memory or even on one hard drive, to uncover actionable knowledge using large scale parallel-processing infrastructures. D. Ulla Gain1 develops BD and symbolizes the aspiration to build platforms and tools to ingest, store and analyze data that can be voluminous, diverse, and possibly fast changing.This strategy is partly descriptive and partly improving. Through launching the term data-milling the Authors try to improve understanding of the phenomenon of BD, as well as, possibilities of data analytics. Launched the term data-milling to represent the searching of the information nuggets from the heterogeneous data. To justify the launched term data-milling, they made the literature review in which they searched the definitions of BDA. Their study shows that BDA is verbosely explained. They used only four statements from 19 to crystallize BDA. The literature review of BDA gave the description of current status of the phenomenon BD. The launched term data-milling improves the understanding of the phenomenon BD, as well as, possibilities of data analytics .There exist large amounts of heterogeneous digital data.This phenomenon is called BD which will be examined. The examination of BD has been launched as BDA. E. Alexander Ginsburg et.al describes the term BD to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies. BD is changing security analytics by providing new tools and opportunities for leveraging large quantities of structured and unstructured data. The Authors specifies the differences between traditional analytics and BDA, and briefly discusses tools used in BDA. They also proposes a series of open questions about the role of BD in security analytics. Big Data technologies can be divided into two groups: 1) Batch Processing, which are analytics on data at rest, and 2) Stream Processing, which are analytics on data in motion. The Authors proposes security to BD by resolving the BDA issues, such as , 1) Data Provenance , which provides the Authenticity and Integrity of data used for analytics. 2) Privacy which enhances a method for regulatory incentives and technical mechanisms to minimize the amount of inferences that BD users can make. 3) Securing Big Data stores ,which focuses on using BD for security, but the other side of the coin is the security of BD.4) Humancomputer interaction, which mentions that the BD facilitates the analysis of diverse sources of data. Compared to the technical mechanisms developed for efficient computation and storage, the human-computer interaction with BD has received less attention and this is an area that needs to grow. The approach is to treat products and services as parts of complex systems that consist of both social and technological components.

Data Forms: Structured: When we talk about structured data, we often conclusively identify that, as soon as we placed our current data ware house in the relational database management system, the structure of the relational database management system was enforced on the current data ware house system, which is inclusive to understand the meaning associated with it. So we know, which columns are placed where, whom are they associated with and how the columns are associated in between tables and table spaces. The format of the data can be in text or numerical, but it is common understanding that for every person there is a unique identifier in terms of Age. The entire data is organized in terms of Entities (Semantic Chunks). Relations or Classes (Similar entities are grouped together).

- Attributes (Same descriptions for entities existing in the Same groups)
- Schema (All Entities in the group have a description associated with it.

• All are present & follow same order. o All of them have same format defined and length defined. Semi Structured: As we move on from structured data to semi structured data, there is little to demarcate and often the differentiating lines goes blurry. The data format that we are describing here does not conform to an explicit and fixed schema, however the tags associated with the data, if found associated with organizational structure, then the same data would be easier to analyze and organize. The same concept described here would predate the idea of XML but not HTML. Data is available in many formats, in the current scenario, electronically

• Database Systems o File Systems e.g., Bibliographic data, Web data o Data Exchange Formats, e.g. EDI, Scientific data Data that is not completely structured, but partially as spoken earlier

• Grouping of Similar Entities and semantically organized. o Entities may not have same attributes in the group

Unstructured: We have already discussed about the Structured and Semi Structured formats. Moving on to the unstructured format, this type would consists of formats that cannot be easily indexed. When we talk about indexing, it is with reference of relational tables and for the purpose of querying or analysis. This would include the file types that are associated with audio, video and image files

CONCLUSION

As the data is becoming bigger and bigger, there is a need to store this data in an efficient manner. I this paper, we have examined the innovative topic of big Data, which has recently gained lots of interest due to its applications. An analysis has been done on BDA, in order to provide an insight on the BDA concepts. Data Security is a challenging task to implement and calls for strong support in terms of security policy formulation and mechanisms. We plan to take up data collection, pre-treatment, integration, map reduce and prediction using Machine Learning techniques. In future, we are planning to develop security alerts, which will provide employees with the ability to view the activity. Events will be filtered down and summarized view will be available to each individual employee.

REFERENCES

[1] Nada Elgendy, Ahmed Elragal ,"Big Data Analytics: A Literature Review Paper", ICDM, LNAI 8557, pp. 214–227, 2014

[2] Bhawna Gupta , Dr. KiranJyoti ,"Big Data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3867-38702014

[3] Weiyi Shang , Zhen Ming Jiang , Hadi Hemmati , Bram Adams , Ahmed E. Hassan , Patrick Martin "Assisting Developers of Big Data Analytics Applications When Deploying on Hadoop Clouds",IEEE 978-1- 4673-3076-3/13 IEEE,2013

[4] Ulla Gain1 ,VirpiHotti ,"Big Data Analytics for Professionals, Data-milling for Laypeople ",International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 4, Number 1 (2014), pp. 33- 402013

[5] Alexander Ginsburg, Luciano JR Santos, KendallScoboria, Evan Scoboria, John Yeoh, "Big Data Analytics for Security Intelligence", 2014

[6] Jainendra Singh , "Big Data Analytic and Mining with Machine Learning Algorithm", World Journal of Computer Application and Technology 1(2): 51 -57, DOI: 10.13189/wjcat.2013.010205,2014

[7] www.forensicrisk.com/big-data-analytics-and-fraud-prevention

[8]www.itproportal.com/big-data-5-major-advantages-of-hadoop

[9] http://web.cs.ucla.edu/~miryung/teaching/EE379K-Spring2014/Papers/Paper%207.pdf

[10]http://www.techrepublic.com/resource-library/whitepapers/big-data-analytics-for-professionals-data-milling-for-laypeople

[11]http://researcher.watson.ibm.com/researcher/view_group.php?id=4

[12]http://www.skytree.net/machine-learning/why-do-machine-learning-big-data

[13]https://en.wikipedia.org/wiki/Big_data

[14] https://mobilesecuritywiki.com/

[15] https://en.wikipedia.org/wiki/Big_data

[16] Vignesh Prajapati , " Big Data Analytics with R and Hadoop ", Kindle Edition.