ISSN 2349-6193
Impact Factor (PIF): 2.243

IJESMR

International Journal OF Engineering Sciences & Management Research

THE SECURITY ISSUES IN BIG DATA: A SURVEY

Prof. Nilesh Sable*, Prof. Harshawardhan Bhosale
*Assistant Professor, Computer Engineering Department JSPM's ICOER, Wagholi, Pune.
Professor, Computer Engineering Department JSPM's ICOER, Wagholi, Pune

Keywords: big data, apache Hadoop, MongoDB, Kerberos, NoSQL and social networks.

ABSTRACT

Big data is the collection of large and complex data sets that are difficult to process using on-hand database management tools or traditional data processing applications. The invention of online social networks, smart phones, fine tuning of ubiquitous computing and many other technological advancements have led to the generation of multiple petabytes of both structured, unstructured and semi-structured data. These massive data sets have lead to the birth of some distributed data processing and storage technologies like Apache Hadoop and MongoDB. To tackle the security issues in Hadoop, the Kerberos protocol has been introduced in its second edition. However, this technological movement has created some security loopholes in the processing and storage of the data sets. This paper tries to list some of the directions research on Big Data challenges has taken for the past five years together with their corresponding Use cases.

INTRODUCTION

According to SINTF, 90% of the world's data was has been generated over the past two years. The emergence of new and advanced technologies over the past decade has boosted the data consumers' appetite to create, store and consume data [1] [2]. CISCO VNI Mobile Focast has highlighted that Asia alone is expected to have a 76% compound annual growth rate for mobile only data [3]. This has since stimulated the desire to address the problems of processing and storage of these vast amounts of data sets. Apache Hadoop and many other technologies have been the knights in shining armour for this problem. This development has had a positive impact on the way large data sets are being processed together with the storage issues. However, less activities have been done to strengthen the security of the Big data infrastructure together with the data. Some researchers have come up with the Kerberos protocol to handle the security issues in Hadoop but apparently there is a plethora of security issues which range from computations in distributed programming frameworks to data provenance. The phenomenon of large data already exists in the fields of physics, biology, environmental ecology, automatic control and other scientific area. It is also crucial in military, communication finance and many other areas. This clearly qualifies Big data as an information security problem which has a lot of challenges which have to be curbed.

RELATED WORK

The growth of big data has raised a number of eyebrows as far as the challenges are concerned. Several authors have discovered a plethora of challenges which include data storage and privacy. Xiaoxue Zhang et al described the storage challenges of Big Data and they analysed them using Social Networks as examples. They further classified the related research issues into the following classifications: small files problem, load balancing, replica consistency and deduplication. Meiko Johnson also did some work on the privacy issues involved with Big Data. He classified these challenges into the following taxonomy: interaction with individuals, re-identification attacks, and probable vs. provable results, targeted identification attacks and economics effects. Visualize and understand their algorithm results. Kapil Bakshi et al [9] discussed the architectural considerations for Big data are concluded that despite the different architectures and design decisions, the analytics systems aim for Scale-out, Elasticity and High availability. Sachchidanand Singh et al in [10] described all the concepts of Big data along with the available market solutions used to handle and explore the unstructured large data are discussed. The observations and the results showed that analytics has become an important part for adding value for the social business.

CHARACTERISTICS OF BIG DATA

Big data is a term used to describe the collection of large and complex data sets that are difficult to process using on hand database management tools or traditional data processing applications. Big data spans across seven dimensions which include volume, variety, volume, value, veracity, volatility and complexity [4].

Volume: The volume of data here is very huge and is generated from a lot of different devices. The size of the data is usually in terabytes and petabytes. All this data also needs to be encrypted for privacy protection.

Velocity: This describes the real time attribute found in some of the data sets for example streaming data. The result that misses the appropriate time is usually of little value.

Variety: Big data consists of a variety of different types of data i.e. structured, unstructured and semi structured data. The data maybe in the form of blogs, videos, pictures, audio files, location information etc.

Value: This refers to the complex, advanced, predictive, business analysis and insights associated with the large data sets.

ISSN 2349-6193
Impact Factor (PIF): 2.243

IJESMR
International Journal OF Engineering Sciences & Management Research

Veracity: This deals with uncertain or imprecise data. It refers to the noise, biases and abnormality in data. This is where we find out if the data that is being stored and mined is meaningful to the problem being analysed.
Volatility: Big Data volatility refers to how long the data is going to be valid and how long it should be stored.
Complexity: A complex dynamic relationship often exists in Big data. The change of one data might result in the change of more than one set of data triggering a rippling effect.

THE SECURITY ISSUES IN BIG DATA

***A) Secure Computations in Distributed Programming Frameworks* :** Distributed programming frameworks use the parallelism concept in computation and storage to process massive amounts of data. The MapReduce framework is a popular example which splits an input file into multiple chunks. In the first phase of MapReduce, a Mapper for each chunk reads the data, performs some computation, and outputs a list of key/value pairs. In the next phase, a Reducer combines the values belonging to each distinct key and outputs the result. There are two major attack prevention measures: securing the mappers and securing the data in the presence of an untrusted mapper [6][5].

B) Security Best Practices for Non-Relational Data Stores: Non-relational data stores have not yet reached security infrastructural maturity. These stores are designed mainly through the use of NoSQL databases. NoSQL Databases were built to tackle different obstacles brought about by the analytics world and hence security was never part of the model at any point of its design stage. Developers using NoSQL databases usually embed security in the middleware. NoSQL databases do not provide any support for enforcing it explicitly in the database. However, clustering aspect of NoSQL databases poses additional challenges to the robustness of such security practices [6][5].

C) Secure Data Storage and Transactions Logs: Data and transaction logs are stored in multi-tiered storage media. Manually moving data between tiers gives the IT manager direct control over exactly what data is moved and when. However, as the size of data set has been, and continues to be, growing exponentially, scalability and availability have necessitated auto-tiering for big data storage management. Auto-tiering solutions do not keep track of where the data is stored, which poses new challenges to secure data storage. New mechanisms are imperative to thwart unauthorized access and maintain the 24/7 availability [6][5].

D) End-Point Input Validation/Filtering: Many big data use cases in enterprise settings require data collection from many sources, such as end-point devices. For example, a security information and event management system (SIEM) may collect event logs from millions of hardware devices and software applications in an enterprise network. A key challenge in the data collection process is input validation: how can we trust the data? How can we validate that a source of input data is not malicious and how can we filter malicious input from our collection? Input validation and filtering is a daunting challenge posed by untrusted input sources, especially with the bring your own device (BYOD) model [6][5].

E) Real-time Security/Compliance Monitoring: Real-time security monitoring has always been a challenge, given the number of alerts generated by (security) devices. These alerts (correlated or not) lead to many false positives, which are mostly ignored or simply "clicked away," as humans cannot cope with the shear amount. This problem might even increase with big data, given the volume and velocity of data streams. However, big data technologies might also provide an opportunity, in the sense that these technologies do allow for fast processing and analytics of different types of data. Which in its turn can be used to provide, for instance, real-time anomaly detection based on scalable security analytics [6][5].

F) Scalable and Composable Privacy-Preserving: Data Mining and Analytics Big data can be seen as a troubling manifestation of Big Brother by potentially enabling invasions of privacy, invasive marketing, decreased civil freedoms, and increase state and corporate control. A recent analysis of how companies are leveraging data analytics for marketing purposes identified an example of how a retailer was able to identify that a teenager was pregnant before her father knew. Similarly, anonymizing data for analytics is not enough to maintain user privacy. For example, AOL released anonymized search logs for academic purposes, but users were easily identified by their searchers. Netflix faced a similar problem when users of their anonymized data set were identified by correlating their Netflix movie scores with IMDB scores. Therefore, it is important to establish guidelines and recommendations for preventing inadvertent privacy disclosures [6][5].

G) Cryptographically Enforced Access Control and Secure Communication: To ensure that the most sensitive private data is end-to-end secure and only accessible to the authorized entities, data has to be encrypted based on access control policies. Specific research in this area such as attribute-based encryption (ABE) has to be made richer, more efficient, and scalable. To ensure authentication, agreement and fairness among the distributed entities, a cryptographically secure communication framework has to be implemented [6][5].

H) Granular Access Control: The security property that matters from the perspective of access control is secrecy—preventing access to data by people that should not have access. The problem with course-grained access mechanisms is that data that could otherwise be shared is often swept into a more restrictive category to guarantee sound security. Granular access control gives data managers a scalpel instead of a sword to share data as much as possible without compromising secrecy [6][5].

I) Granular Audits: With real-time security monitoring, we try to be notified at the moment an attack takes place. In reality, this will not always be the case (e.g., new attacks, missed true positives). In order to get to the bottom of a missed attack, we need audit information. This is not only relevant because we want to understand what happened and what went wrong, but also because

ISSN 2349-6193
Impact Factor (PIF): 2.243

IJESMR
International Journal OF Engineering Sciences & Management Research

compliance, regulation and forensics reasons. In that regard, auditing is not something new, but the scope and granularity might be different. For example, we have to deal with more data objects, which probably are (but not necessarily) distributed [6][5].

J) Data Provenance: Provenance metadata will grow in complexity due to large provenance graphs generated from provenance-enabled programming environments in big data applications. Analysis of such large provenance graphs to detect metadata dependencies for security/confidentiality applications is computationally intensive [6][5].

CONCLUSION

This paper has exposed the major security problems that need to be addressed in Big Data processing and storage. Some researchers have brought about the use of encryption together with Kerberos protocol in order to make the data more secure. However, these security and privacy issues come in different forms such that Kerberos might not be enough to fully secure the data.

REFERENCES

1. http://www.sintef.no
2. http://www.sciencedaily.com/releases/2013/05/130522085217.html
3. http://www.cisco.com/web/solutions/sp/vni/vni_mobile_forecast_highlight/index.html
4. Xiaoxue Zhang, Feng Xu, "Survey of Research on Big Data Storage", 2013 12th International Symposium on Distributed Computing and Applications to Business, Engineering & Science
5. Top Ten Big Data Security And Privacy Challenges CLOUD SECURITY ALLIANCE https://cloudsecurityalliance.org/
6. https://cloudsecurityalliance.org/media/news/csa-releases-the-expanded-top-ten-big-data-security-privacy-challenges/
7. http://www.darkreading.com/views/dont-get-ha-duped-by-big-data-security-p/240144305
8. Sachchidanand Singh, Nirmala Singh, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India
9. Kapil Bakshi, “Considerations for Big Data: Architecture and Approach”, IEEE Aerospace conference 2012