IJCTA, 9(26) 2016, pp. 181-185

A Review of Privacy and Security Issues in Big Data

Sumathy K.[1] and Swarnalatha P.[2]

ABSTRACT

Big data is similar to 'Small-data' having data bigger consequently which requires different approaches, techniques, tools & architectures. It is difficult to process using on-hand database management tools or traditional data processing applications. It is used in online social networks, smart phones, fine tuning of ubiquitous computing and many other technological advancements to the generation of multiple petabytes of structured, unstructured and semi-structured data. In this paper, the main focus is on the big data and various security issues and their related privacy issues.

Keywords: Big data, Privacy issues, Security issues

1. INTRODUCTION

The term big data is a catchword used to illustrate a huge volume of structured as well as unstructured data. As the data size is very large, it is difficult to use traditional database and software techniques to process it. In many organizations either the data is huge or it moves at extremely high-speed or it goes beyond existing processing capability. Big data is likely to easier business in improving their operations and help in making faster and more [1] intelligent decisions.

As developing new technologies and increasing the use of big data in several scopes, security and privacy has been considered as a challenge in big data. There are many security and privacy issues about big data. The top ten security and privacy challenges in big data is focused [2]. Some of these challenges are: secure computations, secure data storage, specific access control and data provenance.

2. MAJOR CHALLENGES OF BIG DATA ARE AS BELOW

A: How to deal with the size of big data.

B: How to handle multiplicity of types, sources, and formats.

C: How good is the data?

D: How broad is the coverage?

E: How fine is the sampling resolution? How timely are the readings?

F: How well understood are the sampling biases?

G: Is there data available, at all?

H: how to react to the flood of information in the time required by the application.

I: how to find high-quality data from the vast collections of data that are out there on the Web. Quality and purpose the challenge is determining the quality of data sets and relevance.

[1] Research Scholar, Department of Computer Science, VIT University, Vellore, TN, India, *E-mail: sumathymca90@gmail.com*

[2] Associate Professor, Department of Computer Science, VIT University, Vellore, TN, India, *E-mail: pswarnalatha@vit.ac.in*

All of these problems combined create a perfect storm of challenges and opportunities to create faster, inexpensive and better solutions for Big Data analytics than traditional approaches can solve.

3. FEATURES OF BIG DATA

Big data is a term refers to the collection of huge data sets which are described by what is often referred as multi 'V'.

In [3] 7 characteristics are used to describe big data:

Volume, variety, volume, value, veracity, volatility and complexity, however in [4], it doesn't point to volatility and complexity. Here we describe each property.

3.1 Volume: Volume is referred to the size of data. The size of data in big data is very large and is usually in terabytes and petabytes scale.

3.2 Velocity: Velocity offered to the speed of data producing and processing.In big data the rate of data producing and processing is very huge.

3.3 Variety: Variety refers to the different types of data in big data.It includes structured, unstructured and semi- structured data and the data can be in different forms.

3.4 Veracity: Veracity refers to the hope of data. Value: Value refers to the worth drives from big data.

3.5 Volatility: "Volatility refers to how long the data is to be valid and how long it should be stored" [3].

3.6 Complexity: "A complex dynamic relationship often exists in big data. The change of one data autotity result in the change of more than one set of data triggering a rippling effect" [3].

Some researchers explain the important characteristics of big data are volume, velocity and variety. In general, the characteristics of the big data are expressed as three Vs.

4. HOW BIG DATA CAUSES PRIVACY VIOLATION IN VARIOUS APPLICATIONS

(A) Health Care System

The Author [5] Because of the great advantages in protecting the health of patients, big data is highly supported by health care system. Big data information is used to recognize [6] people with a high risk of certain medical conditions at very early stage and providing improved quality care and lowering the increase cost of health care. Although there are great benefits, new studies are revealing that big data may be bringing than initially thought.

As per survey it is found that, the health care data is personal, it is very easily accessible and it is important to be conscious about security and privacy implication tapping into big data.

(B) Predictions can cause Discrimination

Big data allows the prediction of quite a bit of the other information about people. it can predict is increasingly develop the potential to be used as a way of discriminating against people in[7] a variety of demographics.

A study shows that when observation of status like information from facebook was analyzed, it gave accurate information to discriminate men depending on the race alcohol consumption, gender etc. It is very much concerned by many people in organizations, employers, education system may use in such models and start discriminating people based on many human oriented parameters.

(C) Product Sales

One of the important applications of big data is marketing where the marketers try to place their products and services in the front of highly targeted customer.

But, when the customer is categorized into one category based on their behaviors, there is a possibility for harm. In spite of the possibility for harm, marketers still use big data to aim at people on social media platform like search engines and email. Forceful entry into personal area by providing advertisements based on the friends, likes and email content is causing anxiety among consumers.

5. BIG DATA SECURITY ISSUES AND CHALLENGES

In this paper, we are highlight the ten BIG DATA specific security and privacy challenges (8) (9)

5.1. Secure Computations in Distributed Programming Framework

Distributed programming framework used the parallelism concept in computational and storage to process massive amount of the data. The Map Reduce framework is a popular example which split an input file into the multiple chunks.

In the first phase of Map Reduce, it Mapped for each chunk reads the data, perform some computational, and outputs a list of key/values. In the next phase, a Reducer combines the values belonging to each distinct key and outputs the result. There are two important attack prevention measures: securing the mappers and securing the data in the presence of unrelay mapper [8] [9].

5.2. Security Best Practices for Non-Relational Data Stores

Non-relational data store have not reached security infrastructural maturity. These stores are designed mainly through the use of the NoSQL databases. NoSQL Databases were built to tackle different obstacles brought about by the analytics world and hence the security is never part of the model at any point of its design stage.

Developers using NoSQL databases usually embed security in middleware. NoSQL databases do not provide any support for enforcing it explicitly in the database. However, the clustering aspect of NoSQL databases poses additional challenges to the robustness of such security practices [8][9].

5.3. Secure Data Storage and Transactions Logs

Data and transaction logs are stored in the multi-tiered storage media. Manually moving data between tiers gives the IT manager direct authority over exactly what data is moved and when. However, as the size of data set has been, and continues to be, growing exponentially, expansible and availability have necessitated auto-tiering for big data storage management. Auto-tiering solution is not keep track, where the data is stored, which poses new challenges to secure data storage. New mechanism are imperative to thwart unauthorized access and maintain the 24/7 availability [8] [9].

5.4. End-Point Input Validation/Filtering

Many big data use cases in operation settings require data collection from many sources, such as end-point devices. A security information and event management system (SIEM) may collect event logs from millions of hardware device and the software applications in an enterprise network.

A key challenge in the data collection process is input validation: how can we hope the data? How can we validate that a source of input data is not malicious and how can we percolate spiteful input from our collection? Input validation and filtering is a daunting challenge posed by untrusted input sources, exclusively with the bring your own device (BYOD) model [8] [9].

5.5. Real-time Security/Compliance Monitoring

Real-time security monitoring has always been a challenges, and give the number of alerts generated by (security) devices. These alerts (correlated or not) lead to many false positive, which are mostly ignored or simply "clicked away," as humans cannot manage with the shear amount.

This problem might even increase with big data, given the volume and velocity of data streams. However, big data technologies ability also provide an opportunity, in the sense that these technologies do allow for the 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 expandable security analytics [8] [9]

5.6. Scalable and Composable Privacy-Preserving Data Mining and Analytics

Big data can be a troubling manifestation of Big Brother by potentially enabling invasions of privacy, invasive marketing, decreased in civil freedoms, and increase state and corporate control.

A recent analysis of how companies are leveraging data analytics for the marketing purposes identified an example of how a retailer was able to identify that a teenager was pregnant before her father knew and similarly, anonymizing data for analytics is not enough to maintain user privacy. For example, AOL released anonymized search logs for the 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 the IMDB scores. Therefore, it is valuable to establish guidelines and recommendations for preventing inadvertent privacy disclosures [8] [9].

5.7. Cryptographically Enforced Access Control and Secure Communication

To make secure that the most sensitive private data is end-to-end secure and only accessible to the authorized entities, data has to encrypted based on the access control policies. Specific research in this area such as attribute-based encryption (ABE) has to be made easy, more efficient, and scalable. To ensure authentication, agreement and fairness between the distributed entities, a cryptographically secure communication frameworks has to be implemented [8][9].

5.8. Granular Access Control

The security property that substance from the perspective of access control is secrecy—preventing access to data by people should not access. The problem with the course-grained access mechanisms is that data that could otherwise be shared is frequently swept into a more restrictive category to guarantee sound security. Granular access control gives data managers a scalpel alternatively of a sword to share data as much as possible without compromising secrecy [8] [9].

5.9. Granular Audits

With real-time security control, we try to be notified at the moment an attack takes place. In reality, this will not always be the case (examples, new attacks, missed true positives). In order to get to the bottom of a missed attack, we need audit information. This is not only appropriate because we want to understand what happened and what went wrong, but also because compliance, regulation and forensics reasons. In that regard, auditing is not something new, but the scope and granularity ability be different. For example, we have to deal with more data objects, which probably are distributed [8] [9].

5.10. Data Provenance

Provenance metadata will be grow in complexity due to the large provenance graph generated from provenance-enabled programming environments in the big data applications. Analysis of such large

provenance graphs to detect metadata dependencies for security/confidentiality applications is computation intensive [8] [9].

Table 1
Related work

S.no	*Authors Name*	*Title of the paper*	*Inference*
1	CSA Big Data Working Group Co-Chairs Lead: Sreeranga Rajan, Fujitsu Co-Chair: Wilco van Ginkel,	Cloud Security Allience, Top Ten Big Data Security and Privacy Challenges,	Authors discusses various Security issues and privacy issues, attacks in big data.
2	Hervais Simo Fraunhofer-Institut für Sichere Informationstechnologie, Darmstadt, germany	"Privacy, technology and the law Big Data for everyone through good design	Authors discusses privacy issues and challenges in big data
3	P. Kamakshi1,Professor, Dept. of Information Technology, Kakatiya Institute of Technology and Science, Warangal, India	Survey on big data and related privacy issues	Author focuses the strength and applications of big data as well as various privacy issues are discussed
4	Zhang Hongjun1, Hao Wenning1, He Dengchao1, Mao Yuxing1 1PLA university of Industry and Technology Nan Jing, China	Survey of Research on Information Security in Big Data	Author discusses various security challenge caused by big data has attracted the attention of information security and industrial community domain.
5	Hervais Simo Fraunhofer-Institut für Sichere Informationstechnologie, Darmstadt,germany	Big Data: Opportunities and Privacy Challenges	Author focuses presented an overview of some key privacy and ethical issues that accompany the rise of big data. Our work is aimed at developing a better understanding of these issues.

6. CONCLUSION

Throughout the paper it was possible to present some of the most important security and privacy challenges that affect Big Data projects and their specificities. The particular characteristics of Big Data make them ineffective if they are not used in an integrated manner. This paper also presents some solutions for these challenges, but it does not provide a final solution for the problem. It rather points to some directions and technologies that might contribute to solve some of the most applicable and challenging Big Data security and privacy issues.

REFERENCES

[1] Agrawal R.,Srikant R., "Privacy Preserving Data Mining " *In the Proceedings of the ACM SIGMOD Conference*, 2000.

[2] Cloude Security Allience, "Top Ten Big Data Security and Privacy Challenges", *www.cloudsecurityalliance.org*, 2012.

[3] K. Zvarevashe, M. Mutandavari, T. Gotora, "A Survey of the Security Use Cases in Big Data", *International Journal of ACSIJ Advances in Computer Science: an International Journal*, **4(4)**, 4259-4266, 2014.

[4] M.D.Assuncau, R.N.Calheiros, S.Bianchi, A.S.Netto, R.Buyya, "Big Data computing and clouds" *Trends and future directions, Journal of Parallel and Distributed Computing*, 2014.

[5] "Big Data is the Future of Healthcare", *Cognizant 20-20 insights*, September 2012.

[6] "Big Data Analytics" ericsson White paper, 284 23-3211 Uen, August 2013.

[7] VINT research report on "Privacy, technology and the law Big Data for everyone through good design".

[8] Top Ten Big Data Security and Privacy Challenges CLOUD SECURITY ALLIANCE https://cloudsecurityalliance.org/

[9] Reference by urn, https://cloudsecurityalliance.org/media/news/csa-releases-the-expanded-top-ten-big-data-security-privacy-challenges/