Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

A Comprehensive Study of Security Issues and Challenges in Big Data

[1]M.Geethanjali and [2]Dr.P.Madhubala,
[1]Asst Professor Cum Research Scholar, Department of Computer Science, St Joseph's College Arts And Science For Women – Hosur., India
[2]Hod Cum Research Supervisor, Department Of Computer Science, Don Bosco College – Dharmapuri, India

Abstract **-** Big data is a collection of data sets which is very large in size as well as complex. The amount of data in world is growing day by day. Data is growing because of use of internet, smart phone and some social network. Traditional database systems are not able to capture, store and analyze this large amount of data. As the internet is growing, amount of big data is continues to grow rapidly. Big Data is going to play important role in future world. Big data changes the way that data is managed and used in some of the applications are in areas such as healthcare, traffic management, banking, retail, education and so on. In the digital and computing world, information is generated and collected at a rate that rapidly exceeds the limits. However, the fast growth rate of such large data generates numerous challenges, such as the rapid growth of data, transfer speed, diverse data and security. This paper shows overview of big data, its characteristics, and stages involved, big data in cloud and security issues and challenges with big data, strengthen of big data security and some proposals to big data security.

Keywords --- *Big data, volume, velocity, variety, veracity, value, cloud, security.*

I. INTRODUCTION

Big data refers to large volumes of data in our everyday lives. Extraordinary growth in data, although predictable, continues to strain corporate resources and government sectors [1]. Data is growing because of use of internet, smart phone and social networks like face book, you tube, twitter and so on. Many organizations demand efficient solutions to store and analyze these big amount data that are preliminary generated from various sources such as high throughput instruments, sensors or connected devices. For this purpose, big data technologies can utilize cloud computing to provide significant benefits, such as the availability of automated tools to assemble, connect, configure and reconfigure virtualized resources on demand [5]. These make it much easier to meet organizational goals as organizations can easily deploy cloud services.

Day by day, the security of confidential information is gaining more and more attention. According to 2016 Trust wave Global Security Report; nearly 97% of applications had at least one security vulnerability. By this, the security is extremely high ranked priority for any enterprises. However, with the ease of adoption of web-based, smart phones and cloud-based applications, the confidential information has become easy to access from various platforms. Such platforms are highly vulnerable to hacking, especially if they are not maintained properly. Unlike earlier, companies are now collecting and using lots of customer data. A lack of data security can bring some serious security issues and organization's reputation will be at stake [3]. The challenges include analysis, capture, search, sharing, storage, revelation, and privacy violations.

A. Big Data- Where is it?

Big data surrounds us, although we may not immediately realize it. Part of the problem is that, except in unusual circumstances, most of us don't deal with large amount of data in our everyday lives. Lacking of this immediate experience, we often fail to understand both opportunities as well as challenges presented by big data. There are two types of Big Data: structured and unstructured.

Structured data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smart phones and global positioning system (GPS) devices. Structured data also include things like account balances, sales figures and transaction data.

Unstructured data include more multifarious information, such as customer reviews from feasible websites, photos and other multimedia, and comments on social networking sites. These data cannot be separated into categorized or analyzed numerically.

II. BIG DATA CHARATERISTICS

A. Data Volume:

Data volume measures the amount of data available to an organization, companies, government, financial, medical institution, educational institution which are producing data in order of terabytes every day. Some applications are used to handle the data [2].

B. Data Velocity:

Data velocity measures the speed of data creation, streaming and aggregation. Data is generating at very fast rate [2]. Data velocity measures large data in short period of time. It deals with the pace at which data flows in from sources like business process, machines, networks and human interactions with things like social media sites, mobile devices, etc. The flow of data is massive and continuous.

C. Data Variety:

Data variety is a measure of the richness of the data representation – text, images, videos, audio, etc. At present data comes in different forms including data streams, text, picture, audio, video, structured, semi structured, unstructured. Unstructured data is difficult to handle with traditional tools and techniques.

D. Data Veracity:

Data veracity refers to the biases, noise and abnormality in data [4]. Is the data that is being stores and mined meaningful to the problem be analyzed. Data veracity is the degree to which the data is accurate, precise and trusted.

E. Data Value:

Data value measures the usefulness of data in making decisions. It has been noted that "the purpose of computing is insight, not numbers". Data science is exploratory and useful in getting to know the data, but "analytic science" encompasses the predictive power of big data.

Nowadays data comes from different sources. And it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or data can quickly spiral out of control. These are certain characteristics of big data from normal data.

Figure 1: 5 V's of BIG DATA

III. STAGES INVOLVED IN BIG DATA

A. Data Acquisition:

The first step in Big Data is acquiring the data itself. With the growing medium the rate of data generation is rising exponentially. With the introduction of smart devices which are used with a wide array of sensors continuously generate data. Most of this data is not useful and can be discarded, however due to its unstructured form; selectively discarding the data presents a challenge. This data becomes more potent in nature when it's merged with other valuable data and superimposed. Due to the interconnectedness of devices over the World Wide Web, data is increasingly being collated and stored in the cloud.

B. Data Extraction:

All of the data generated and acquired is not of use. It contains a large amount of redundant or unimportant data. The challenges presented in data extraction are two folds: firstly, due to nature of data generated, deciding which data to keep and which to discard increasingly depends on the context in which the data was initially generated. Secondly, a lack of a common platform presents its own set of challenges. Due to wide variety of data that exists, bringing them under a common platform to standardize data extraction is a major challenge.

C. Data Collation:

Data from a singular source often is not enough for analysis or prediction. More than one data sources are often combined to give a bigger picture to analyze. For example a health monitor application often collects data from the health –rate sensor, pedometer, etc. to summarize the health information of the user. Likewise, weather prediction software take in data from many sources which reveal the daily humidity, temperature, precipitation, etc. In the scheme of Big Data convergence of data to form a bbigger picture is often considered a very important part of processing.

D. Data Structuring:

Once all the data is aggregated, it is very important to present and store data for further use in a structured format. The structuring is important so queries can be made on the data. Data structuring employs method of organizing the data in particular schema. Various new platforms, such as NoSQL, can query even on unstructured data and are being increasingly used for Big Data Analysis.

E. Data Visualization:

Once the data is structured, queries are made on the data and the data is presented in a visual format. Data Analysis involves targeting areas of interest and providing results based on the data that has been structured.

F. Data Interpretation:

The ultimate step in Big Data processing includes interpretation and gaining valuable information from the data that is processed. The information gained can be of two types: **Retrospective Analysis** includes gaining insights about events and actions that have already taken place. For instance, data about the television viewership for a show in different areas can help us to judge the popularity of the show in those areas. **Prospective Analysis** includes judging patterns and discerning trends for future from data that is already been generated. Weather Prediction using big data analysis is example of prospective analysis.

IV. BIG DATA IN CLOUD

Big Data in Clouds is a new generation data intensive platform for quickly building the analytics and deploying over an elastically scalable infrastructure. Cloud computing is widely used in association with Big Data due to the numerous advantages it provides namely as on-demand/real-time service availability, widespread access, and sharing of resources [16, 17, 18].

However, usage of cloud computing comes with a huge number of security challenges since this technology includes multiple areas and principals like networking, resource sharing, databases, virtualization, operating systems etc., therefore security issues of these systems and technologies are applicable to cloud computing [21].

One of the main issues with the cloud is securing storage data. Henceforth, cloud service providers have suggested secure ways for sharing Big Data on the cloud platform. These providers assure that their clients do not face issues like data loss or theft, caused by user impersonation [19].Based on the services rendered to the end users, these are broadly classified into four types as described below

A. Public Big Data clouds:

Large scale data organization and Processing over the elastically scalable infrastructure. The resources are served over internet as pay-as-go computing models. The examples include Amazan Big Data computing in clouds, Google cloud platform of Big Data computing and so on.

B. Private Big Data clouds:

Deployment of Big Data platform within the enterprise over a virtualized infrastructure, with a greater control and privacy to the single organization.

C. Hybrid Big Data clouds:

Federation of public and private Big Data clouds for scalability, disaster recovery and high availability. In this deployment, the private task can be migrated to the public infrastructure during peak workloads.

D. Big Data access network and computing platform:

Integrated platform of data, computing and analytics delivered as serviced by multiple distinct providers. Big Data computing in clouds "Big Data Clouds" is data intensive analytics platform of large scale, distributed compute and storage infrastructures.

Integrated cloud and Big Data access networks on cloud infrastructure for analytics development. The content from several sources like social media, web logs, scientific studies, sensor networks, business transactions etc. are growing rapidly. Deriving useful information for decision making from such large data, fusing the information from several sources would be a challenging task.

V. BIG DATA SECURITY ISSUES AND CHALLENGES

There is an increasing need of research in technologies that can handle the vast volume of data and make it secure efficiently [5]. Current Technologies for securing data are slow when applied to huge amount of data. Big data is used by many enterprises, organizations for marketing and research but they may not have fundamental assets particularly for security perspective. If a security threat occurs to big data, it would become even more serious issues.

A. Top 10 security & Privacy Challenges:

- Secure computations in distributed programming frameworks.
- Security best practices for non-relational data stores.
- Security data storage and transactions log.
- End-point input validation/filtering.
- Real-time security monitoring.
- Scalable privacy-preserving data mining and analytics.
- Cryptographically enforced data centric security.
- Granular access control.
- Granular audits.
- Data provenance.

The above challenges are grouped into four broad components by Cloud Security Alliance (CSA). They are,

- Infrastructure Security.
- Data Privacy.
- Data Management.
- Integrity and Reactive Security

In most cases, the distributed system's computations have a limited protection; say one or two levels [11]. At some or the other point, connections security and the encryption of access control will be ineffective and inaccessible. Automated data transforms needs extra security norms, which are frequently unavailable. Suggested detailed audits are not periodically done in Big Data because of the massive amount of data being involved. Because of the big data size, its stock is not always tracked or monitored.

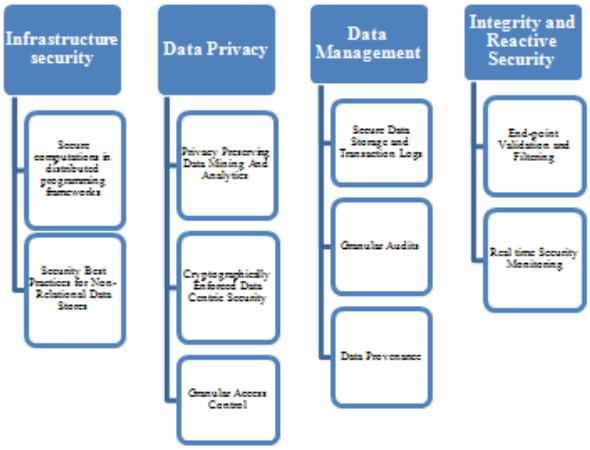

Figure 2: Classification of the top 10 challenges

B. Strengthen Big Data Security:

- By using cloud computing that the possible route of enhancing the security of Big Data is via the constant growth of the antivirus industry.
- Another way of protecting data is by using VPN.
- Concentrate on security of the application, rather than security of the device.
- Deploy real-time security information.
- Supply proactive and reactive security.

VI. PROPOSALS TO ADDRESS BIG DATA SECURITY

The basic and more common solution for this includes encrypting everything to make data secure regardless where the data resides (data centre, computer, mobile devices or any other). As Big Data grows and it's processing gets faster, then encryption, masking and tokenization are critical elements for protecting sensitive data [15]. Due to its characteristics, Big Data projects need to take into consider the identification of the different data sources, the origin and creators of data as well as who is allowed to access the data. As recommendation, different security mechanisms should be closer to the data sources and data itself. In order to provide security right at the origin of data, and mechanism of control and prevention on archiving, data leakage prevention and access should work together.

CONCLUSION

This paper described the big data concept, its characteristics and importance. Effectively managing and prioritizing the volume, velocity, variety, veracity and value of data requires human insight, a multipronged approach and multiple layers of defence. Using big data tools to analyze the massive amount of threat received daily, and correlating the different components of an attack, allows a security vendor to continuously update the global threat intelligence and equates to improved threat knowledge and insight. It initiates a collaborative research effort to begin examining big data issues and challenges. Big Data is also changing things in the business world. Companies are using big data analysis to target marketing at very specific demographics. To accept and adapt to the new technologies, many challenges and security issues exist which need to be brought up right in the beginning before it is too late. All those issues and challenges have been described in this paper. These challenges and issues will help

the business organizations which are moving the technology for increasing the value of the business to consider them right in the beginning and to find the ways to protest them. By reducing risk, they avoid potential recovery costs, adverse brand impacts, and legal implications. The future of big data is unlimited and evolution is unimaginable. Hence the hope is to develop better and better techniques and technologies towards finding solutions for big data security.

ACKNOWLEDGEMENT

I would like to express my sense of gratitude to ST. JOSEPH'S COLLEGE OF ARTS AND SCIENCE FOR WOMEN, Hosur for their support and encouragement. And I also like to thank PERIYAR UNIVERSITY, Salem for providing me the opportunity to carry out the research work in Big Data. Finally I would like to thank my Research Supervisor Dr.P.MADHUBALA for her guidance and valuable suggestions.

References

[1] Bermen, Jules J. "Principle of Big Data", Morgan Kaufmann, Waltham,2013.

[2] Cooper,Mell, "Tackling Big Data", NIST Information Technology Laboratory Computer Security Division.

[3] Gaddam, "Securing your Big Data Environment", Black Hat USA 2015.

[4] GeethaKumari, Srivatsava, "Big Data Analysis for Implementation of Enterprise Data Security", International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN:2249-9555

[5] Inukollu, Aris and Ravuri, "Security Issues Associated with Big Data in Cloud Computing", International Journal of Network Security & Its Applications (IJNSA), Vol. 6, No.3, May 2014.

[6] Bilge, L. & T. Dumitras, (2012, October) Before We Knew It: An empirical study of zero-day attacks in the real world. Paper presented at the ACM Conference on Computer and Communications Security (CCS), Raleigh,

[7] Bryant, R., R. Katz & E. Lazowaska.(2008). Big Data Computing: Creating revolutionary breakthroughs in commerce, science and society. Washington, DC: Computing Community Consortium.

[8] N. Mirmura Gonzalez, M. Torrez Rojas, Maciel da Silva, F. Redigolo, T. Melo de Brito Carvalho, C. Miers, M. Naslund, and A. Amhed, "A framework for authentication and authorization credentials in cloud computing", in Trust, Security and Privacy in Computing and communications(TrustCom), 2013 12th IEEE International Conference on, pp. 509-516, July 2013.

[9] R. Banyal, P. Jain, and V. Jain, "Multi-factor authentication framework for cloud computing, " in computational Intelligence, Modelling and simulation(CIMSim), 2013 Fifth International conference on pp. 105-110, sept 2013.

[10] Kapil Bakshi, "Considerations for Big Data: Architecture and Approach", IEEE, Aerospace Conference, 2012.

[11] Sachchidanand Singh, Nirmala Singh, "Big Data Analytics", IEEE, International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, 2012.

[12] Domenico Talia. "Clouds for Scalable Big Data Analytics". IEEE Computer, vol 46, no:98-101. 2013.

[13] J.Singh, "Big Data : Tools and Technologies in Big Data, "vol. 112, no.15,pp. 6-10, 2015

[14] Dona Sarkar, Asoke Nath, "Big Data – A Pilot Study on Scope and Challenges", International Journal of Advance Research in Computer Science and Management Studies (IJARCSMS, ISSN: 2371-7782), Vol 2, Issue 12, 31st Dec 2014.

[15] Hashem, I., Yaqoob, I. & Anuar, N. et. al., 2015. "The rise of "big data" on cloud computing: Review and open research issues", Information Systems, Vol. 47, pp. 98-115, Available from: Science Direct, [Accessed on 3rd August 2016].

[16] Mysore, D. & Khupat, S. & Jain, S., 2013. "Introduction to Big Data classification and architecture". Available from:https://www.ibm.com/developerworks/library/bd-archpatterns1/, [Accessed on 1st August 2016].

[17] Perreault, L., 2015. "Big Data and Privacy: Emerging Issues", Conf-IRM 2015 Proceedings, Available from: IEEE Computer Society Digital Library, [Accessed on 6th August 2016

[18] Jain, P, 2012. "Security issues and their solution in cloud computing", International Journal of Computing and Business Research, Available from: http://www.researchmanuscripts.com/isociety2012/1.pdf, [Accessed on 3rd August 2016].

[19] Kumar, A. & Lee, H., n.d. "Efficient and Secure Cloud Storage for Handling Big Data", No. 3, pp. 162-166, Available from: IEEE Computer Society Digital Library, [Accessed on 1st August 2016].

[20] Saranya, R. & MuthuKumar, V.P., 2015. "Security issues associated with big data in cloud computing", International Journal of Multidisciplinary Research and Development, Vol. 2, No. 4, pp. 580-585, Available from:www.allsubjectjournal.com/archives/download?id=716&refnum=253, [Accessed on 5th August 2016].