ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 9, Issue, 05, pp.50320-50325, May, 2017

INTERNATIONAL JOURNAL
OF CURRENT RESEARCH

REVIEW ARTICLE

BIG DATA SECURITY

*Shraddha Singh

Galgotias University, India

ARTICLE INFO

Article History:
Received 17th February, 2017
Received in revised form
26th March, 2017
Accepted 04th April, 2017
Published online 23rd May, 2017

Key words:

Big Data, Privacy, Big data security, Security policy.

ABSTRACT

In today's era of IT world, Big Data is a new curve and a current buzz word now. Daily tremendous amount of digital data is being produced. It require an advance data management system to handle such a huge flood of data that are obtained due to advancement in tools and technologies being used. This has led human being in big dilemma. The security industry and research institute are paying more attention to the emerging security challenges in big data environment. The current security challenges in big data environment is related to privacy and volume of data. This paper discusses the security issues related to big data due to inadequate research and security solutions also the needs and challenges faced by the big data security, the security framework and proposed approaches.

Citation: Shraddha Singh, 2017. "Big data security", *International Journal of Current Research*, 9, (05), 50320-50325.

INTRODUCTION

Big data has become a buzzword in computer science, information system, business, marketing and other field in recent years. Due to technological advancement we are continuously generating an ever increasing amount of data. The data is constantly generated by people and devices and is being increased day by day over last 20 years and the web is overloaded with huge and exponential generation of data. According to a report from International Data Corporation (IDC), in 2011, the overall created and copied data volume in the world was 1.8ZB (≈ 1021*B*) which increased by nearly nine times within five years (Min Chen *et al.*, 2014). This figure will double at least every other two years in the near future. It becomes very important to store, manage and share huge and complex amount of data securely, in order to identify pattern and analyze complex data. With big amount of data comes the big security issues and challenges. Big data has become a part of life due to development of revolutionary tools & technology and also the popularization. The research and industries are both paying more attention to big data security in order to efficiently use big data technologies to solve big data analysis problems. The tremendous amount of structured, semi-structured and unstructured data are being developed every day. After gathering, sorting, analyzing and mining the user's sensitive data can be obtained. These sensitive data are not only limited to the use of enterprise but also to the other businesses when stored on big data platform. Due to high volume, velocity and variety of data privacy issue and security issue are magnified. The different source of data, different formats and data flow when combined with the high volume and streaming nature of data acquisition create a security risk. The enterprises are collecting and analyzing data for decades. There are many software infrastructure are used for data storage and analysis. The Hadoop infrastructure are most popular among them as it is being used by many enterprises for data storage and data analysis. The Hadoop framework supports large set of data in distributed environment and is part of apache software foundation. Hadoop framework is based on master and slave concept. The Hadoop framework helps to save file in distributed file system and do analytics on these files using Google map reduce algorithm. Due to distributed file system the failure of some node can be negotiated. The top companies like Google, Amazon, Yahoo and IBM are using Hadoop framework for their data support. The security challenges are arisen due to the heterogeneous data collected for storage and computation on these data. As many enterprises and businesses are adopting big data technologies the traditional existing method of storage will not be able to handle such a heterogeneous and huge amount of data. The traditional data management and data analysis system are based on RDBMS. The relational databases are used to store only the structured data in a tabular format and require expensive hardware. This paper deals with security challenges, need and solution for the problem and at the end concluded by the proposed security methods.

Corresponding author:* **Shraddha Singh,
Galgotias University, India.

Bigdata and big data security

Big Data

Big data refers to the extremely huge amount of data that need to be analyzed for computation in order to draw a meaningful pattern or format out of it. There are many definitions presented in different articles. Manyika *et al.* (2011) define Big Data as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze". Likewise, Davis and Patterson (2012) say "Big data is data too big to be handled and analyzed by traditional database protocols such as SQL". Both groups of authors previously mentioned go beyond the only size aspects of data when defining Big Data! EddDumbill in (O'Reilly Media, 2014) explicitly conveys the multi-dimensionality of Big Data when adding that "the data is too big, moves too fast, or doesn't fit the structures of your database architectures". The big data are characterized by its high volume, high variety, high velocity, high veracity and value. The big data have the 5 v's properties in them which are:

1. **Volume:** The volume in big data refers to the high amount of data being generated from different sources. The data in rest and high in masses are called volume.
2. **Velocity**: Velocity not only refers to the speed at which the big data are collected but also the analysis performed on the collected data to find the valuable information.
3. **Variety:** The variety refers to the type of data being collected from different sources like social media, government, healthcare etc. which are in different format like audios, videos, text, log files, images, PDF files and so on. These data are classified into three categories i.e. structured, unstructured and semi structured.
4. **Veracity:** The veracity in big data are the abnormalities, noises and biases in the data. These noise must be configured in advance to avoid any anamoly in user privacy.
5. **Validity:** The validity in big data means the data should be correct and accurate for intended use. A valid data helps to make a right decision.

Big Data Security

The most important asset for an enterprise is data. Most of the businesses and enterprises are using data for research and marketing hence must be sure from the perspective of security concern. Big data is not just a trending word for any business rather it's more than that to bring a fruitful benefits to the one who willingly use it. As big data are being used by many companies doesn't mean it has made its road to future. There are some issues, risk, challenges and security breaches present along with it which comes under the big data security. As many growing companies use big data technologies for storing and analyzing petabytes of data which they receive from social medias, web logs and click stream to get knowledge about their customer and business. The security breach would be too big with more serious and damaging property. The main concern in Information technology is the safety and privacy issue, the challenges in safety mechanism is due to data encryption to the diverse data obtained and in big data privacy there are two aspect i.e. one during data acquisition and second during storage, transmission and usage of the personnel privacy protection. Due to different software platforms, cloud infrastructure on different network the attack probabilities increase hence the traditional security measure applied to structured data is not sufficient.

Privacy and Security

Since there are variety of data which are being collected by the big data applications and which are transferred over the network they may also contain sensitive information like location based information, healthcare record etc. hence privacy and security concern naturally arises. There are several studies which are being performed to protect the data center. When the data are being transferred to data center the privacy and security must be addressed during data transmission to avoid the data being attacked.

Security a Big Question of Big Data

A big question to big data is which security and privacy technologies are best and adequate to be used to assure efficient access to data. The limitation of IT security allow

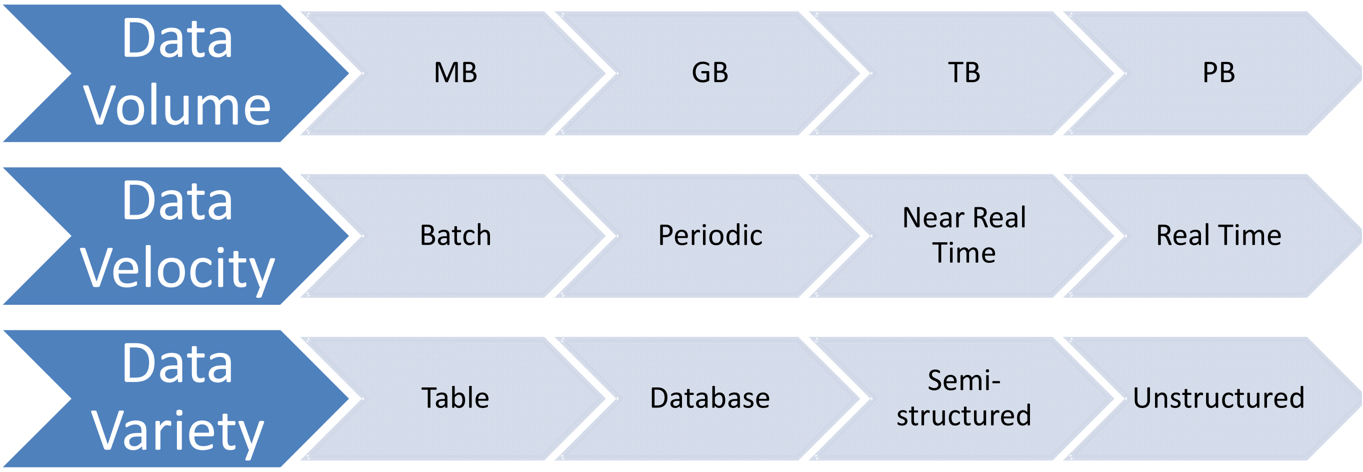

Fig: Big data expanding in three fronts at an increasing rate

attackers to use subversion of software to insert the malicious software into operating system and applications which could cause serious adverse impact to big data system. A verifiable protection must be made so that only the authorized access to data are made.

Is Your Structured, Semi-structured or Unstructured Data at Risk?

The best way to protect big data is not only focusing on volume, variety and velocity but to its security also. With great power of data comes great responsibility (KalyaniShirudkar and DilipMotwani, 2015). The right technology framework should be provided for deep visibility and multilayer security. The multilevel security means implementing security control at application, operating system and network level any keep track of any malicious activity performed.

The need of security in big data

There are many companies which uses big data for research and marketing but may not be secure from security perspective. The security breach may cause more serious damage to the business if proper precaution is not taken care off. The data are being generated at exponential rate and solutions to many security issues that affect the data are not known still. Since data are being collected in petabytes from different source and in different format they are more likely to contain sensitive and personal information and hence the increase of risk and security breaches occur. There are many techniques such as logging, encryption, honeypot detection for making a big data secure and also deployment of big data for fraud detection in many different organizations are being preferred. Big Data implementations typically include open source code, with the potential for unrecognized back doors and default credentials (http://www.computerweekly.com/ feature/How-to-tackle-big-data-from-a-security-point-of-view (online)).

Challenges to big data security

As the traditional method of data management and data analysis are based on Relational database management which are used only to store structured data now faces many challenges. Since the data are both structured and unstructured it become difficult store it with traditional RDBMS and to enforce their security mechanism. There are so many challenges associated with the big data security that are faced by businesses and organization. Andrew Trait in Nov 16, 2016 has given an article on 5 Big Data Security Challenges which are described below:

The Data governance

Data governance refers to the effective management of data which consider issues like security, accuracy, availability and usability. A process must be defined and their effectiveness and adherence to that process must be managed and evaluated. A survey by Rand Worldwide, conducted in 2013, showed that, while 82% of companies know they face external regulation, 44% had no formal data governance policy and 22% had no plans to implement one. There is little evidence that the situation has improved three years on (Andrew Tait, 2016).

B. The Privacy-preserving analytics

The privacy preserving plays a very important role in big data analytics. Since the data are obtained from different online sources hence the data collected becomes doubtful that it has no missing piece of information to the original data and does not lead to the privacy violation. While privacy issue are considered, present datas*et al*one are not looked instead both present and future data are seen. Encryption and hashing are used to enhance data privacy. Data when not deal with care becomes toxic. Researchers are working on large scale analysis of data to provide data privacy in many ways. The researchers are also working on the area of homomorphic encryption in which analysis are performed at encrypted data. These analysis on encrypted data however leads to slower analysis of data which are encrypted to those which present in raw form. Microsoft are two IT giants who are currently conducting research into differential privacy. This approach adds mathematical noise to an individual's data, but enables the data set as a whole to be mined in search of overall patterns. Apple has said they are already deploying this technology in iOS 10 (Andrew Tait, 2016).

C. The Perimeter-based security

A perimeter-based security model is a common model for big data installation for data security. In this model the private and sensitive applications are kept under the secure network so that the outside attacker is unable to modify or steel the information. This model is based on the hope and faith that the outside attacker will not do any harm to the system private information.

D. The Non-relational data-stores

Any company small or large dealing with big data, NoSQL tool can play an important role for them. NoSQL is different from traditional RDBMS in a way that neither they require SQL as a query language nor fixed table schema. NoSQL are at evolutionary stage of its lifecycle and the attack vector are not well mapped out yet. The security feature for NoSQL are being compromised due to lack of knowledge on how to configure them since the relational database management system are being used by many and its security feature are much known due to availability of training courses and auditing tools. Soon the NoSQL technique will become better and data would be adequately protected.

E. The Configuration management:

Due to nature of data to be distributed among multiple cluster the configuration management becomes challenging. The configuration of cluster are spread among various XML,JSON and text file generally incompatible and when any new type of machine are added into the cluster they need to be set up, configured and patched so that no security hole is created.

Motivation and Related work

Motivation

Due to the increasing popularity in big data technology, the security issues are also introduced due to adaption of these technology are increasing. The attackers are continuously searching for the loop holes in the big data technology in order to get the access the user privacy. The traditional security will not be able to handle this security breaches hence it is require to visualize, control and inspect the security measure to ensure security. The main motive to understand the big data security is to find the component which are more prone to attacks, loop holes and challenges and to come up with the infrastructure and platform which are less vulnerable to attack.

B. Related Work:

The Hadoop is a java based distributed system framework that are used by many business that still requiring more feature to be developed to address security issues. The researchers have found some of the issues and working on that issues, some are presented below:

Kevin Hamlen *et al* has proposed that the data in the database can be stored in encrypted form as well instead of plaint text only. When the data are stored in the encrypted form in database the intruder does not get access to actual data even when they have access to the database. But due to encrypted data the processing overhead is increased and the processing has to be performed on encrypted data rather than the plain text and thus added extra security layer. The IBM researcher has proposed that there should be a secured environment for query processing. Kerberos can be an effective environment, in this an arbitrator which is a trusted third party perform the secure authentication in an open network. Kerberos was developed by MIT and is based on Needham-Schroeder protocol. Airavat (Kilzer *et al*., 2016) has proposed some security advancement in Map Reduce environment. They prevent information leak by providing the mathematical bound potential privacy algorithm.

Yuangang Yao and *et al*. has proposed a semantic security analysis framework for processing and analysing multi source heterogeneous security data. It provide advantages to current big data analysis techniques for information security technology by providing data collection, storage, processing, analysis and also meets up demand of current security incident analysis that supports big data analysis, breadth and depth. The propose framework for semantic security analysis presents novel method for analysing security data on different levels. The security analysis for searching, visualization and interaction are provided by HCI analysis for application and interfaces. The current focus is on data analysis and data processing and future work will optimize and implement the HCI analysis and semantic analysis. MdIleasPramanik *et al* (MdIleasPramanik *et al*., 2016) has proposed a big data analytics framework which explore four dominant factor of network criminals which are network extraction, subgroup detection, interaction pattern discovery and central member identification. In their proposed work they provide four important application that are related to two dimension of SNA directly by combining big data source, transformation, platform and tools. Deepak Puthal and *et al*. (2017) has proposed a shared key synchronization method to process an end to end security in big data stream processing system which consist of distributed sensor and cloud-hosted stream processing engines (DSM). There proposed method synchronize shared key without any communication between sensing devices and DSM where the sensing device obtain shared key re-initialization properties from their neighbours.
Xinhua and *et al*. (2015) has proposed a systematic framework for sharing, submitting and storing sensitive data securely on big data platform based on heterogeneous proxy re-encryption algorithm and guaranteeing secure use of clear text in cloud by private use of space by user process based on VMM. This framework protect security of user's important sensitive data and the data owner will have complete control to their own data hence a feasible solution which would balance the benefits of the involved parties under semi trusted condition.
Jiaqi, Lizhe and *et al*. (2014) has designed and proposed a security framework which runs Map Reduce tasks among different clusters in distributed environment. A single sign on process is provided in this framework to submit jobs to G-Hadoop framework and also a security mechanism to protect it from attackers or misusing. The security mechanism is based on SSL, cryptographic algorithm and GSI and has ability to prevent common attacks like replay attack, MITM attack and delay attack and ensure secure communication over network.

The security framework for big data

There are many security framework for big data has been developed as discussed in the related work above. In this section a security framework given by Forrester has been discussed. This framework allows to help risk and security professional to control big data. The Forrester has divided security framework into the three steps:

Defining the data

There are many industry which are focusing on big data and shifting there tools and infrastructure toward this initiatives. The companies may be unaware about the kind of data they are being depositing whether structured or unstructured. Based on level of toxicity the data classification must be defined to counterpart the legal and privacy. Once the data is classified it becomes easier to provide the security to it and to locate it in enterprise. The discovery and classification are critical in nature since data discovery index and locate the data in big data environment and the data classification catalogs the big data in order to make it easier to control it.

B. Dissecting and analyzing the data:

As data are valuable for the business the risk and security professional can also draw value from the data deposited. The big data environment store security information and allow access to more data. The security information management (SIM) and network analysis and visibility (NAV) solution are used to enhance the security measure.

C. Defending and protecting the data:

The risk and security professionals are defending the data by providing a solutions to the number of attacks increasing day by day. The Forrester framework provide the basic way to protect and defend the data:

- The access control helps to ensure that, the right data are accessed by right user and at the right time. Since the data are deposited in massive volume the intruder or cybercriminals can access the sensitive information. Hence it is important to limit the number of people who can access the data deposited and monitor the activities performed by this people throughout.
- The data inspection can help security team by providing an alert due to potential abuses according to their data

usage patterns. This can be achieved by deploying NAV tools like metadata analysis, flow analysis or packet capture analysis tools and integrate it with SIM solutions to protect toxic data proactively.

- The data disposing is a powerful defensive technique when the data are no longer needed by any enterprise. Any toxic data which are no longer needed for any real business interest, data investigation should be classified properly and disposed off.
- The data killing devalues the data so that no cybercriminals can use or sell it. Many techniques such as encryption, masking and tokenization be used by any enterprise to devalue the data. Once the data are encrypted it become useless for any cybercriminals.

Since the data are the most powerful fuel to any enterprises or organizations, their security becomes the prime importance. If sensitive data fall in wrong hand would cause serious disaster to the organization. Hence the security planning for the big data should be started at the early stage when the big data is initiated to reduce cost, failure, risk and deployment pain.

The proposed approach

There are several security measure which can be used to improve the security of big data environment. The cloud environment has many different technology hence we require many solution to make the environment secure. The following measure of security must be taken to ensure security in big data environment.

The File Encryption

The data are present in different machine on different cluster, the attacker can steal all sensitive information. The data should be in encrypted form for this different machine should use different encryption key and the sensitive information should be kept secure firewalls. In this way the attacker will not get the important information and misuse it even when they are able to get the data.

The Network Encryption

As per industries standard the communication over the network should be in encrypted form and all the RPC call should be over SSL. This way even if attacker access network communication packets, they can’t extract or manipulate packet information.

The Logging

The data should be logged every time when a map reduce job modifies the data or when the user’s information responsible for that job is used. Log audit should be done regularly so that if any malicious operation or activities are done can be found easily.

The Software Format and Node Maintenance

All the software should be updated so that the system become secure and the nodes should be formatted regularly to eliminate the virus present.

The Nodes Authentication

A node before joining a cluster should be authenticated and if any node is found to be a malicious node then they should not be allowed. The Kerberos can be used to authenticate nodes.

The Honeypot Nodes

Honeypot nodes appears similar to a regular node and present at the clusters but is generally a trap which trap the attackers or hackers.

Conclusion

In this paper we have represented big data security. Why we need big data security and the challenges faced by it. There are many challenges related to big data security such as security of data storage, distributed programming, scalable data analytics and mining, input filtering from client, secure communication and access control. The various big data security framework has been studied and presented in brief. A security framework given by Forrester has been presented and the proposed approaches for big data security has been provided.

REFERENCES

Andrew Tait“5 Big Data Security Challenges” Nov 16, 2016 http://blog.learningtree.com/five-big-data-security-challenges(online).

Davis, K., D. Patterson, “Ethics of Big Data: Balancing Risk and Innovation”, O’Reilly Media, 2012page 4.

Deepak Puthal, Surya Nepal, Rajiv Ranjan, Jinjun Chen “A Synchronized Shared Key Generation Method for Maintaining End-to-End Security of Big Data Streams”Proceedings of the 50th Hawaii International Conference on System Sciences | 2017.

JiaqiZhaoa, LizheWangb, JieTaoc, JinjunChend, WeiyeSunc, Rajiv Ranjane, Joanna Kołodziejf, AchimStreitc, DimitriosGeorgakopoulose “A security framework in G-Hadoop for big data computing across distributed Cloud data centres”Journal of Computer and System Sciences 80 (2014) 994–1007.

KalyaniShirudkar, DilipMotwani “Big-Data Security” Volume 5, Issue 3, March 2015 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering Research Paper.

Kilzer, Ann, Emmett Witchel, Indrajit Roy, VitalyShmatikov, and Srinath T.V. Setty. "Airavat:Security and Privacy for MapReduce."Nov 16, 2016

Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C.Roxburgh, A.H. Byers, “Big Data: The Next Frontier for Innovation, Competition, and Productivity”, McKinsey Global Institute, 2011page 1-J.

MdIleasPramanikWenping Zhang Y.K.LauChunping Li “*A Framework for Criminal Network Analysis Using Big Data”* 978-1-5090-6119-8/16 $31.00 © 2016 IEEE DOI 10.1109/ICEBE.2016.12

Min Chen, Shiwen Mao, Yunhao Liu “Big Data: A Survey” Mobile NetwAppl (2014) 19:171–209DOI 10.1007/s11036-013-0489-0

O’Reilly Media, “Big Data Now”: 2014 Edition, O’Reilly Media 2014page 3-I.

Peter Wood“How to tackle big data from a security point of view”http://www.computerweekly.com/feature/How-to-tackle-big-data-from-a-security-point-of-view(online).

Xinhua Dong, Ruixuan Li, Heng He, Wanwan Zhou, Zhengyuan Xue, and Hao Wu “Secure Sensitive Data Sharing on a Big Data Platform” TSINGHUA SCIENCE AND TECHNOLOGY ISSNl l1007-0214l l08/11l lpp72-80 Volume 20, Number 1, February 2015.

Yuangang Yao, Lei Zhang, Jin Yi, Yong Peng, Weihua Hu, Lei Shi “A Framework for Big Data Security Analysis and the Semantic Technology”, China Information Technology Security Evaluation Center Beijing, Chinayaoyg@itsec.gov.cn

****