International Journal of Engineering & Technology, 7 (4.39) (2018) 730-733

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Security and Privacy Mechanisms of Big Data

P Amarendra Reddy[1], Sheikh Gouse[2], P Bhaskara Reddy[3]

[1, 2, 3] Department of Information Technology, MLR Institute of Technology, Dundigal, Hyderabad - 500043, India.
**Corresponding author E-mail: amarpanyala88@gmail.com*

Abstract

Big Data advances in three key areas like storage, processing and analysis. Hadoop architecture is produced to store extensive measure of data through adaptable parallel handling with speed to get the outcomes. Organizations must guarantee that every single big data bases are resistant to security dangers and vulnerabilities. Amid data gathering, all the important security assurances, for example, continuous administration should to be satisfied. Remembering the big size of big data, Organizations should to recollect the way that overseeing such data could be troublesome and requires uncommon endeavors. Be that as it may, making every one of these strides would help keep up buyer protection. The Security and Privacy challenges for Big Data might be sorted out on big data community. Securing the framework of big data frameworks includes Securing appropriated calculations and data stores. Securing the data it is of vital significance, so we need to guarantee that data scattering is security saving and that touchy data is ensured using cryptography and granular access control. Overseeing tremendous volumes of data requires adaptable and conveyed answers for Securing data stores as well as empowering effective reviews and examinations of data provenance. At long last, the streaming data that is rolling in from various endpoints must be checked for respectability and can be utilized to perform continuous examination for security episodes to guarantee the strength of the framework. The explanation behind such ruptures may likewise be that security applications that are intended to store certain measures of data can't the huge volumes of data that the previously mentioned datasets have. Likewise, these security advances are wasteful to oversee dynamic data and can control static data as it were. In this way, only a customary security check can't distinguish security patches for nonstop streaming data. For this reason, we require full-time protection while data streaming and Big Data Analysis.

Keywords: *Big Data, Security, Hadoop, Map Reduce*

1. Introduction

The term Big Data means to expansive scale data administration and examination advances that surpass the ability of conventional data handling technologies. Big Data is separated from customary advances in three different ways: the measure of data i.e, volume, the speed of data generates and transmission i.e., speeds, and the kinds of organized and unstructured data i.e, variety.
Big Data is a term associated with data collection whose size or sort is over the limit of RDB's to get, administer, and process the data. Additionally, it has something like one of the going with characteristics tremendous volume, speed and variety. Big Data begins from different gadgets, wire or remote sensors, devices, video or sound sounds, log records, value based applications, web and online media a lot of it made continuously and in a huge scale.

2. Arrays Traditional Analytics and Big Data Analytics

The main characteristics between the both are:

a. Storage
b. Processing
c. Analysis [ ]

3. Need of Security for Big Data

Big Data is changing the analysis scene. Specifically, Big Data analysis can be utilized to enhance data security. Information driven data security goes back to fraud recognition and abnormality based interruption identification frameworks. Misrepresentation location is a standout amongst the most noticeable uses for Big Data examination. Most organizations have led misrepresentation identification for quite a long time. In any case, the custom-fabricated framework to dig Big Data for misrepresentation identification was not practical to adjust for other extortion recognition employments.

4. Privacy Challenges

It might be composed into different parts of the big data ecosystem system:

a. Framework Security
b. Data Privacy and Management
c. Integrity Security

Securing the framework of Big Data includes Securing appropriated calculations and data stores. Securing the data it is of fundamental significance, so we need to guarantee that data dispersal is security safeguarding and that touchy data is ensured using cryptography and granular access control. Overseeing tremendous volumes of data requires adaptable and disseminated answers for Securing data stores as well as empowering effective reviews and examinations of data provenance. At last, the streaming data that is rolling in from various endpoints must be checked for integrity and can be utilized to perform constant or real time analysis for security incidents to guarantee the soundness of the framework.

5. Security Challenges

a. Security and market patterns are making new security administration obstacles.
b. The existing security framework is not any more sufficient.
c. The time of Big Data Security Analytics.
d. Security analysis tools can't stay aware of the present data accumulation and handling needs.
e. Organizations need an enterprise wide security domain.
f. All existing security analysis tools depends human insight
g. Analytics aren't coordinated for mechanized incident reaction.

SECURITY CHALLENGES			
Infrastructure Security	**Data Privacy**	**Data Management and Integrity**	**Reactive Security**
1. Secure Distributed Processing of Data 2. Security Best Actions for Non-Relational Data-Bases	1. Data Analysis through Data Mining Preserving Data Privacy 2. Cryptographic Solutions for Data Security 3. Granular Access Control	1. Secure Data Storage and Transaction Logs 2. Granular Audits 3. Data Provenance	1. End-to-End Filtering & Validation 2. Supervising the Security Level in Real-Time

6. Big Data Security Issues

For this reason, you require full-time security while data streaming and Big Data examination.

Type of privacy Challenge needed	Description	Usage
Secure Computations in Distributed Programming	It uses parallelism concept in computation and storage of data	Map Reduce framework
Security Best Practices for Non-Relational Data Stores	Non-relational data stores have not yet reached security infrastructural maturity.	No SQL databases
Secure Data Storage and Transactions Logs	Data and transaction logs are stored in multi-tiered storage media	auto-tiering
End-Point Input Validation/Filtering	Many big data use cases in enterprise settings require data collection from many sources, such as end-point devices	security information and event management system (SIEM) bring your own device (BYOD) model weather sensors and feedback votes
Real-time Security/Compliance	Monitoring Real-time security monitoring has always been a challenge, given the number of alerts generated by (security) devices.	real-time anomaly detection. the fraud related to claims
Scalable and Composable Privacy-Preserving	Potentially enabling invasions of privacy, invasive marketing, decreased civil freedoms, and increase state and corporate control	AOL, Netflix, Sarbanes-Oxley
Cryptographically Enforced Access Control and Secure Communication	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.	attribute-based encryption (ABE)
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	protect corporate financial information , Health Insurance Portability and Accountability Act
Granular Audits	With real-time security monitoring , we try to be notified at the moment an attack takes place	HIPAA, PCI, Sarbanes-Oxley
Data Provenance	Metadata will grow in complexity due to large provenance graphs generated from provenance-enabled programming environments in big data applications	history of a digital record PCI or Sarbanes-Oxley.

7. Solution for Big Data of Privacy Challenge

Hadoop Security Then and Now

Component	Component Original Hadoop Release	Now Included/Available Encryption Not included
Encryption	Not included	DEK encryption automatically applied to data in HFDS and in motion; additional data protection features are specic to each commercial distribution; KMS manages encryption keys; Kerberos is commonly used; additional encryption methods are Hadoop compatible/available
Authentication	None	Kerberos is the foundation for Hadoop secure mode; Active Directory and LDAP extended to Hadoop; identity management solutions extended to Hadoop.
Access & Permissions	HDFS le permissions	Permissions can be set by individual, group, and role and set for specic data types and les; data masking can be applied to limit data that is accessed.

So here are some specific facets of clustered systems attackers will target.

Systemic Security	Operational Security	Embedded Security
Data access & ownership Data at rest protection Multi-tenancy Inter-node communication Client interaction Distributed nodes	Authentication and authorization Administrative data access Configuration and patch management Software bundles Authentication of applications and nodes Audit and logging Monitoring, filtering, and blocking API security	LDAP/AD integration Apache Ranger HDFS Encryption Apache Knox Apache Atlas Apache Ambari Monitoring
Architecting for Security	**Security Architectures**	**Our Recommendations**
Systemic Threat-Response Models Operational Threat-Response Models	Walled Garden Cluster Security Data Centric Security Enterprise Security Options	Use Kerberos for node authentication Use file layer encryption Use key management Use Apache Ranger

		Automate deployment Use logging and monitoring Use secure communication

Our base recommendations are as follows:

Recommendations to use	Purpose	Usage
Kerberos	To validate nodes and client applications before admission to the cluster, and to support other identity functions	Ambari
File Layer Encryption	To protect data at rest, ensure administrators and applications cannot directly access files, and prevent information leakage.	Data leakages
Key Management	To protect encryption keys and manage different keys for different files.	Public key cryptographic
Apache Ranger	To track module configuration and to set usage policies for fine grained control over data access.	Centralized security administration
Apache Knox	It is a pluggable framework, and a new REST API service can be added easily using a configurable services definition	Hadoop's REST and HTTP services
Automate Deployment	To virtualization technologies, cloud provider facilities, and scripted deployments based on products such as Chef and Puppet.	cloud
Logging and Monitoring	To logging tools that leverage the big data cluster itself — to validate usage and provide forensic system logs.	system logs
Secure Communication	SSL or TLSnetwork security to authenticate and ensure privacy of communications between nodes, name servers, and applications	client-toWeb application-layer
Aduit	As customers deploy Hadoop into corporate data and processing environments, metadata and data governance must be vital parts of any enterprise-ready data lake	Search and lineage for datasets • Metadata-driven data access control

Any security control used for Hadoop must meet the following requirements:

1. It must not compromise the basic functionality of the cluster.
2. It should scale in the same manner as the cluster.
3. It should address a security threat to the Hadoop cluster or data stored within it.

8. Big Data Security Generation

S.No	Type of Generation	Description	Use	Support
1	IDS(Intrusion Detection Systems)	Security engineers understood the requirement for layered security	Security and response	Defensive security is inconceivable.
2	SIEM(Security Information and EventManagement)	Managing cautions from various devices	Aggregate alerts from different areas	Data to security experts.
3	BDAS(Big Data Analytics in Security)	It gives a critical development in significant security intelligence	Decreases the ideal opportunity for corresponding, merging, and contextualizing various security occasion data	Stores historical data for feature usage

IDS has generally been a critical issue with innovations neglect to give the devices to help long haul, huge scale analysis for a few reasons:

1. Storing and holding an expansive amount of data was not financially doable
2. Performing analysis and complex inquiries on huge, organized data
3. Unstructured data has no devices to break down and oversee.
4. Systems utilize bunch registering frameworks.

Coming up next are just a portion of the inquiries that should be tended to:

1. Data provenance.
2. Privacy
3. Securing Big Data stores
4. Human-PC communication:

8.1. Big Data Security Analytics Technology Transformation

Eventually, the goal of Big Data security analysis is to give a thorough and up-to the second perspective of IT exercises with the goal that security experts and administrators can make convenient, data -driven choices. From an innovation point of view, this will require new security frameworks giving:

S.No	research project	Authors Name	Paradigm	Sample / test
1	BotCloud	Fraçois, J. et al. 2011	MapReduce	Netflow
2	Advanced Persistent Threat (APT)	Verizon, 2010	Malware and virus	Data breach was recorded
3	Worldwide Intelligence Network Environment (WINE)	Dumitras&Shoue, 2011	Data analytic	Validates real-time data.

9. Conclusion

The Big Data architecture will both become more critical to secure, and more frequently attacked. Thus growing the list of big data security issue. Although the information security practices, methodologies and tools to ensure the security and privacy of the Big Data ecosystem already exist, the particular characteristics of Big Data make them ineffective if they are not used in an integrated manner. It also presents some solutions for these challenges, but it does not provide a definitive solution for the problem. It rather points to some directions and technologies that might contribute to solve some of the most relevant and challenging Big Data security and privacy issues. Challenges represent only the tip of the iceberg about the problems that still need to be studied and solved on the development of secure and privacy-aware Big Data ecosystem. More researches required to overcome the security of big data instead of current security algorithms and methods.

Acknowledgment

The authors would like to acknowledge Science and Engineering Research Board, India for financial support.

References

[1] Tankard, Colin. "Big data security". Network Security, 2012, no.7: 5-8.

[2] Domenico Talia. "Clouds for Scalable Big Data Analytics". IEEE Computer, 2013, vol.46, no.5: 98-101.

[3] Karadsheh, Louay. "Applying security policies and service level agreement to IaaS service model to enhance security and transition". Computers & Security, 2012, vol.31, no.3: 315-326.

[4] ChunmingRong , Son T. Nguyen , Martin GiljeJaatun. "Beyond lightning: A survey on security challenges in cloud computing". Computers and Electrical Engineering, 2013, vol.39, no.1:47-54.

[5] Mark Dermot Ryan. "Cloud computing security: the scientific challenge, and a survey of solutions". Journal of Systems and Software, 2013, vol.86, no.9: 2263-2268.

[6] Min Chen, Shiwen Mao, Yunhao Liu. "Big data: a survey". Mobile Networks and Applications, 2014, vol.19, no.2: 171-209.

[7] Yuri Demchenko, Paola Grosso, Cees de Laat, Peter Membrey, "Addressing Big Data Issues in Scientific Data Infrastructure",IEEE 2014.

[8] Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu, and Jun Shao "Toward Efficient and Privacy-Preserving Computing in Big Data Era" IEEE Network , July/August 2014.

[9] Alvaro A. Cárdenas , Pratyusa K. Manadhata , Sreeranga P. Rajan "Data Analytics for Security", Co published by the IEEE Computer and Reliability Societies, Nov/Dec 2013.

[10] Ren, Yulong, and Wen Tang. "A Service Integrity Assurance Framework for Cloud Computing Based on Mapreduce."Proceedings of IEEE CCIS2012. Hangzhou: 2012, pp 240 – 244, Oct. 30 2012-Nov. 1 2012.

[11] N, Gonzalez, Miers C, Redigolo F, Carvalho T, Simplicio M, de Sousa G.T, and Pourzandi M. "A Quantitative Analysis of Current Security Concerns and Solutions for Cloud Computing.". Athens: 2011., pp 231 – 238, Nov. 29 2011- Dec. 1 2011

[12] Hao, Chen, and Ying Qiao. "Research of Cloud Computing based on the Hadoop platform.". Chengdu, China: 2011, pp. 181 – 184, 21-23 Oct 2011.

[13] Y, Amanatullah, Ipung H.P., Juliandri A, and Lim C. "Toward cloud computing reference architecture: Cloud service management perspective.". Jakarta: 2013, pp. 1-4, 1314 Jun. 2013.

[14] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.

[15] Foster I, Zhao Y, Raicu I, Lu S (2008) Cloud Computing and Grid Computing 360-Degree Compared. In: Grid Computing Environments Workshop (GCE'08). oi:10.1109/GCE.2008.4738445

[16] Fellowes, W. (2008). Partly Cloudy, Blue-Sky Thinking about Cloud Computing. White paper. 451 Groups.

[17] M. Casassa-Mont, S. Pearson and P. Bramhall, "Towards Accountable Management of Identity and Privacy: Sticky Policies and Enforceable Tracing Services", Proc. DEXA 2003, IEEE Computer Society, 2003, pp. 377-382

[18] J. Salmon, "Clouded in uncertainty – the legal pitfalls of cloud computing", Computing, 24 Sept 2008, http://www.computing.co.uk/computing/features/2226701/clo ud-ed-uncertainty-4229153

[19] Khajeh-Hosseini, A., Greenwood, D., Sommerville, I., (2010). Cloud Migration: A Case Study of Migrating an Enterprise IT System to IaaS. Submitted to IEEE CLOUD 2010

[20] S. Overby, How to Negotiate a Better Cloud Computing Contract, CIO, April 21, 2010, http://www.cio.com/article/591629/How_to_Negotiate_a_Better_Cloud_Computing_Contract

[21] Krautheim FJ (2009) Private virtual infrastructure for cloud computing. In: Proc of HotCloud

[22] Santos N, Gummadi K, Rodrigues R (2009) Towards trusted cloud computing. In: Proc of Ho

[23] Hortonworks, Inc. 2014. "Setting Up Kerberos for Hadoop 2.x." Hortonworks Data Platform: Installing Hadoop Using Apache Ambari. Palo Alto, CA: Hortonworks, Inc.Advantech. (2013).

[24] Enhancing Big Data Security. Retrieved from http://www.advantech.com.tw/nc/newsletter/whitepaper/big_data/big_data.pdf

[25] Agrawal, D., Das, S., & El Abbadi, A. (2011). Big data and cloud computing. In Proceedings of the 14th International Conference on Extending Database Technology - EDBT/ICDT '11 (p. 530). New York, New York, USA: ACM Press. doi:10.1145/1951365.1951432