International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-5, June 2020

Exploration of Big Data Security Framework using Machine Learning

K. Rajeshkumar, S. Dhanasekaran, V. Vasudevan

Abstract: ***As with prior technological advancements, big data technology is growing at present and we have to identify what are the possible threats to overhead the present security systems. Due to the development of recent technical environment like cloud, network connected smartphones and the omnipresent digital conversion of huge volume of all types of data poses more possible threats to sensitive data. Due to the improved vulnerability big data requires increased responsibility. During the last two years, the amount of data that has been created is about 90% of the whole data created. Strengthening the security of sensitive data from unauthorized discovery is the most challenging process in all kind of data processing. Data Leakage Detection offers a set of methods and techniques that can professionally solve the problem arising in particular critical data. The large amounts of existing data is mostly unstructured. To retrieve meaningful information, we have to develop superior analytical method in big data. At present we have more algorithms for security which are not easy to be implement for huge volume of data. We have to protect the sensitive information as well as details related users with the help of security protocols in big data. The sensitive data of the patient, different types of code patterns and set of attributes to be secured by using machine learning tool. Machine learning tools have a lot of library functions to protect the sensitive information about the clients. We recommend the Secure Pattern-Based Data Sensitivity Framework (PBDSF), to protect such sensitive information from big data using Machine Learning. In the proposed framework, HDFS is implemented to analysis the big data, to classify most important information and converting the sensitive data in a secure manner.***

Keywords – HDFS, EMR, Security, Big Data, Content Based Access Control, Sensitive Data Detection, Attribute-Based Access Control

I. INTRODUCTION

To create an Enhanced security framework for protecting confidential medical data from unauthorized users on big data, Machine learning approach is proposed in this study. The present security solutions are not able to provide pool proof security in big data. Due to complexity in time consumption, our present security mechanisms are not sufficient for providing the security against unauthorized users. In this approach, big data file is to be classified based upon the risk effect level into public and confidential. Due to the technological advancement and usage of the more internet based services, many users need to access the big data servers or services frequently. The user's confidential data like healthcare information, trading secrets and personal data can be kept in secured manner because of the technology advancement. The HDFS is the central information storage system that can store the large data through Hadoop application. With the help of Name Node and Data Node architecture, HDFS provides high performance access to highly scalable Hadoop collections. HDFS can manage the groups of large amount of data and relevant big data analytical applications. Our approach is to create a security protocol, techniques, tools, and security policy management framework to avoid unauthorized access in big data.

HDFS is a competence tool [2], [3] which maintains handles, stores large amount of data, provides quick programmed conclusions, and reduces the humanoid estimations. With the capacities of dependability, accountability, idleness, and distributed architecture HDFS is wildly recognized as commonly used dataset tool [4]. Due to this support, HDFS was designed to deal various big data types; structured, semi-structured and unstructured. In [5], Map Reduce Job-Scheduling algorithm guides grouping big data in an extented networking condition. We are to be able to secure the fixed data against vulnerabilities with normal security tools. The solution of the big data security would happen data accessibility, reliability, and privacy.

More encryption techniques are applied to protect the data from unauthorized access [6], [7]. The data supervision and classifications of data is the basic concern in big data. Kerberos management policies provides secure data at communications, transmission, authorization, and storage [8].

It is developed for transport layer secure communication, data encryption, and data authentication. The solutions of Kerberos is not easy to implement.

Revised Manuscript Received on May 15, 2020.
* Correspondence Author
K Rajeshkumar*, Assistant Professor, Department of Computer Science and Engineering, Theni Kammavar Sangam College of Technology, Theni, India, kumar85rajesh@gmail.com
Dr.S. Dhanasekaran, Associate Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankovil, srividhans@gmail.com
Dr. V. Vasudevan, Senior Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankovil, vasudevan_klu@yahoo.co.in

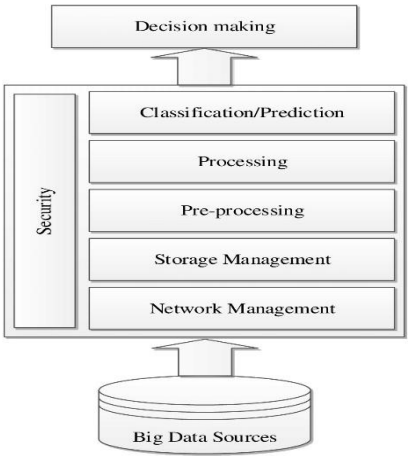

Fig. 1 Big Data security process

Based on the data criticality and sensitivity, we have proposed an integrated methodology for big data security and classification. Before initiating any data transmission process, we implement encryption and decryption tools on the confidential data. The purpose of the encryption methods is to secure the big data results in the healthcare organizations [19]. According to the sensitive information, we have adapted an exceptional algorithm that can assessment the sensitivity of the data.

A. The main features are

1. Based on Data content and attributes create a novel data grouping techniques.
2. Classify the big data based on the predefined policies.
3. Suggest a new encoding method for private data.
4. Develop consumer support software Tool to execute security, help admin, and classification.
5. Incorporate big data security and classification.

II. RELATED WORK

SIEM tool is an encryption like tool to preventing potential risks [9] and Data Loss Preventing (DLP) is a tool to protect sensitive data [10]. Due to the large size in big data, the author in [11] demonstrates the data security strategies without affecting the performance. Implementation of data protection results in large time complexity. The various kind of policies are well-ordered in big data sources. Hence it is difficult to apply the protection technique. Data protection in cloud environment is main challenges for facility provider [12]. Normally the insider and outsider attacks are identified in most of the organizations. The insider attack implementation creates the restricted access of cloud environment security. A method cyber security model which is proposed for the solution of insider attack in cloud system. In this framework architecture, we are using the Hidden Markov model that is helpful to identify the upcoming activities. These are classified as four different types as per their outcomes in cloud service security: Hacked, Under Attack, Sensitivity, and Legitimate. An abnormal behavior of the user can be identified by an intelligent driven security model [13]. The security of the cloud environment is increased with the help of dynamic and self-adaptive suspicious user's record system and self-assuring (SA) tool. The SA tool comprises of archive to recognize and storing the keywords of the user's asymmetrical accomplishments and hazardous logs. There are no protection against data lose and data leakage. The main aspects of our goal is to protecting valuable information in big data. Securing the data is the utmost essential problems for Big Data investigation. Integrity of the data refers to protecting data from being altered by unauthorized users. Cryptography is a key role in certifying the data integrity. Hashing the data and comparing it with the hash value of original data is the technique to protect the data integrity. For that purpose the secured hash value of the original data must be delivered to users. The availability of data refers to ensuring that authorized users can access the data when required. Denying access to data has become a very public attack at present day. Nowadays all the administrations are using a single server or system to store important files like customer's personal information and patients health related data. It gives a high risks. Hence we implemented risk assessment taxonomy method to protect sensitive information in Big data. Hadoop Eco Systems is presented in [14]. The Distributed Data admission and storing on the cloud is authentic and protected arbitrary encoding methods are executed. Authentication is implemented to create the organization safe while retaining the performance principles. In this proposed system we discuss Hadoop in addition to effort in preserving the confidentiality and safety of big data. Our approach is precisely developed to reduce the threats of immigration and combination of a large volume of trusted data. CBAC is proposed in [24] that access control choices depends on the evaluations between the retrieved data and users IDs. Every user in our model fit to a distinctive role that is permissible to contact to positive types of data and it can be able to make sensible access control decision with a lesser overhead. The author in [25] applied the concepts of SSD framework that to be able to classify the sensitive data in automatic manner. Each and every organization have a skilled person to analyze the sensitive data. Their responsibilities are to identify the key (sensitive) information in given data value that can't be encountered in earlier process. After identified the sensitive data, that data set shall be used for the training the neural network. The similarities among the data is measured by Similarity analyzer. The semantic correlation among the set of data is determined by data usage pattern analyzer. Markov's algorithm is modified to recognize the semantic correlation. Data sensitivity estimator recognizes sensitive information substances from datasets.

The author in [20] defined the significance of confidentiality problems in Big Data arena, and deliberated how ABAC to be able to safeguard critical information from information abuse. Most of the big organization like medical field, Universities etc. maybe progressive from Attribute Based Access Control.

More number of algorithms established to excavation data from Electronic Health Record [21, 22]. Majorities of those algorithms are developed to classify the information depends on the particular features. The authors in [23] implemented Latent Semantic examination to determine pattern between remedial terms. With the help of our typical method, the communication among the patient and hospital management big data is improvised. Hospital Management data requires more security tools than other simple endorsements.

III. INTEGRATED CLASSIFICATION

The integrated methods comprise classification and protected (secure) file.

A. Classification Technique

Essentially, big data is categorized as per their essential, significances, methods of fortification depending upon the data critical and compassion. As per the implementation of RIL, our big data services are classified as public and confidential.

B. Confidential data

Confidential data is secured data which are not accessible for all other people. Practically, that data is individually recognizable data that is considered isolated in natural manner, such as medical data, locations, previous labor understanding, and fiscal information.

C. Public data

In public data, info to be available to access of all other candidates, recycled and reorganized by means of someone with no present local, nationwide or global allowed limitations on right of entry. For example, in an organization all the information in public domain is available to all the users and all the individuals, like job description and marketing materials.

IV. EXPERIMENTAL RESULT AND PERFORMANCE CALCULATION

Basically big data is classified as structured, semi-structured, and unstructured. Some of the information in big data should be kept in public. In [15], [16] the author proposed Map-reduced Framework to scrutinize the result of the suggested methods and to validate its performance. Protecting the sensitive information in the large volume of data is a critical process in big data. We make use of Hadoop map-reduce framework to finish the challenging process.

A. Performance Evaluation

The author in [18], proposed a data transmission method that depends on Secure Socket Layer to connect among Name nodes of both user (sender and receiver) in cloud system. It generates temporary SK, a random hash function and more number of tickets encoded using SK. This imposes additional security overhead which decrease performance when large volume of data is transferred among dissimilar cloud systems. In [17], they ignored acknowledgement through communication process between sender and receiver. If there is no acknowledgement means that the unauthorized user can access the original data or the sender can delete the original data after transmitting the data. This ends in producing the retransmission of the packet that will rise the BW and the data handling overhead.

With the help of [17], [18] we can understand that it increases the both total transmission response time and total delay time. Hence it degrades the HDFS performance. Our protocol optimize the communication, the encryption and decryption operation among the users (sender and receiver).

V. CONCLUSION

From our research process, we clearly understand an objectives of the security in big data. With the implementation of the security framework the big data information should be kept in secured manner. Finally we conclude that we are using Elliptic Curve Cryptography to encrypt the data over Rivest-Shamir-Adleman mechanism. The reply time is more quickly as it needs less handling control and memory.

REFERENCES

1. Ismail Hababeh, Ammar Gharaibeh, Samer Nofal, Issa Khalil, “An Integrated Methodology for Big Data Classification and Security for Improving cloud Systems Data Mobility”, in IEEE Tran.,vol.7,pp.9153-9163, Jan 2019
2. Sonic. Accessed: Sep. 2018.[Online]. Available: http://mirrors.sonic.net/ apache/Hadoop/ common/ hadoop2.6.0/
3. K.Shvachko, H.Radia, S.Radia, and R.Chansler, ‘‘The Hadoop distributed file system,’’ in Proc. IEEE 26th Symp. Mass Storage Syst. Technol. (MSST), May 2010, pp. 1–10.
4. J. V. Gautam, H. B. Prajapati, V. K. Dabhi, and S. Chaudhary, ‘‘A survey on job scheduling algorithms in big data processing,’’ in Proc. IEEE Int. Conf.Electr.,Comput. Commun. Technol. (ICECCT), Mar.2015,pp.1–11
5. A. Holmes, Hadoop in Practice. Shelter Island, NY, USA: Manning Publications, 2012.
6. P. Adluru, S. S. Datla, and X. Zhang, ‘‘Hadoop eco system for big data security and privacy,’’ in Proc. Long Island Syst., Appl. Technol. Conf. (LISAT), 2015, pp. 1–6.
7. G. Raj, R. C. Kesireddi, and S. Gupta, ‘‘Enhancement of security mechanism for confidential data using AES-128, 192 and 256bit encryption in cloud,’’ in Proc. 1st Int. Conf. Next Gener. Comput. Technol. (NGCT), Sep. 2015, pp. 374–378.
8. N.Chaudhariand, S.Srivastava, ‘‘Big data security issues and challenges,’’ in Proc. Int. Conf. Comput., Commun. Autom. (ICCCA), 2016, pp. 60–64.
9. D. Miller, S. Harris, A. Harper, S. VanDyke, and C. Blask, Security Information and Event Management (SIEM) Implementation. New York, NY, USA: McGraw-Hill, 2011.
10. N. I. Readshaw, J. Ramanathan, and G. G. Bray, ‘‘Method and apparatus for associating data loss protection (DLP) policies with endpoints,’’ U.S. Patent 9311495, Apr. 12, 2016.
11. N.Chaudhariand S.Srivastava,‘‘Big data security issues and challenges,’’ in Proc. Int. Conf. Comput., Commun. Autom. (ICCCA), 2016, pp. 60–64
12. A. S. Sohal, R. Sandhu, S. K. Sood, and V. Chang, ‘‘A cybersecurity framework to identify malicious edge device in fog computing and cloud- of-things environments,’’ Comput.Secur., vol.74, pp.340–354,May2018.
13. A. Gupta, A. Verma, P. Kalra, and L. Kumar, ‘‘Big data: A security compliance model,’’ in Proc. Conf. IT Bus. Ind. Government (CSIBIG), 2014, pp. 1–5.
14. P. Adluru, S. S. Datla, and X. Zhang, ‘‘Hadoop eco system for big data security and privacy,’’ in Proc. Long Island Syst., Appl. Technol. Conf. (LISAT), 2015, pp. 1–6
15. Hadoop. Mapreduce Tutorial. Accessed: Sep. 2018. [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html/
16. I. Hababeh. (2015). ‘‘Data migration among different clouds.’’ [Online]. Available: https://arxiv.org/ abs/1512.08383
17. Q. Shen, L. Zhang, X. Yang, Y. Yang, Z. Wu, and Y. Zhang, ‘‘SecDM: Securing data migration between cloud storage systems,’’ in Proc. IEEE 9thInt.Conf. Dependable, Autonomic Secure Comput. (DASC),Dec.2011, pp. 636–641.
18. C. Zhonghan, Z. Diming, H. Hao, and Q. Zhenjiang, ‘‘Design and implementation of data encryption in cloud based on HDFS,’’ in Proc. Int. Workshop Cloud Comput. Inf. Secur. (CCIS), 2013, pp. 274–277.
19. Yin Chung Yau, Praveen Khethavath, J. A. Figueroa, “Secure Pattern-Based Data Sensitivity Framework for Big Data in Healthcare”, in IEEE, Comp. Society, May 2019pp.65-70,.

20. Cavoukian, Ann, Michelle Chibba, Graham Williamson, and Andrew Ferguson. "The importance of ABAC: attribute-based access control to big data: privacy and context." Privacy and Big Data Institute, Ryerson University, Toronto, Canada(2015).
21. Metsker, Oleg, Ekaterina Bolgova, Alexey Yakovlev, Anastasia Funkner, and Sergey Kovalchuk. "Pattern-based mining in electronic health records for complex clinical process analysis." Procedia computer science 119 (2017): 197-206.
22. Sikorskiy, Sergey, Oleg Metsker, Alexey Yakovlev, and Sergey Kovalchuk. "Machine Learning Based Text Mining in Electronic Health Records: Cardiovascular Patient Cases." In International Conference on Computational Science, pp. 818-824. Springer, Cham, 2018.
23. Gefen, David, Jake Miller, Johnathon Kyle Armstrong, Frances H. Cornelius, Noreen Robertson, Aaron Smith-McLallen, and Jennifer A. Taylor. "Identifying patterns in medical records through latent semantic analysis." Commun. ACM 61, no. 6 (2018): 72-77.
24. Zeng, Wenrong, Yohao Yang, and Bo Luo, “Access control for big data content”. In 2013 IEEE international Conference on Big Data,pp.17-22. IEEE, 2013.
25. TK, Ashwin Kumar, Hong Liu, Johnson P. Thomas, and Goutam Mylavarapu, “Identifying Sensitive Data Items within Hadoop” In 2015 IEEE 17th International Conference on High Performance Computing and communications.

AUTHORS PROFILE

Mr.K. Rajeshkumar, completed his M.Tech., Network Engineering in Kalaslingam Univerisity on 2010. He is working as Assistant Professor in Theni Kamavar Sangam college of Technoloty, Theni. His major research is Network Security, Big Data Analytics, and Cloud Computing.

Dr.S. Dhanasekaran, received PhD in 2017. He is working as an Associate Professor in Kalasalingam Academy of Research and Education, Krishnankovil. His research area is CloudComputing. He has published more than 20 research paper in SCOPUS and SCI. He is a life time member of ISTE and IEEE.

Dr.V. Vasudevan received Ph.D. degree in 1992. He is currently serving as Registrar and Senior Professor with the school of Computing and information sciences, Kalasalingam Academy of Research and Education, Krishnankovil. He has published 60 papers in International journals and International conferences. He has large volume of publications in refereed journals. He is a life time member of Indian Society of Technical Education (ISTE).