Indonesian Journal of Electrical Engineering and Computer Science
Vol. 11, No. 1, July 2018, pp. 268~274
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v11.i1.pp268-274

Big Data Security Architecture using Split and Merge Method

Archana R A[1], Ravindra S Hegadi[2], Manjunath T N[3]
[1]Research Scholar, R& D Centre, Bharathiar University, Coimbatore, Tamil Nadu, India
[2]Director, School of Compuational Sciences, Solapur University, Maharastra, India
[3]Professor, Dept of ISE, BMS Institute of Technology, Bangalore, Karnataka, India

Article Info

Article history:

Received Sep 5, 2017
Revised Nov 10, 2017
Accepted Jan 19, 2018

Keywords:

Big Data
Data Masking
Data Security

ABSTRACT

Due to rapid growth of unstructured data in contemporary information world, there is an essence of big data infrastructure for many applications spread across domains, due to the different source information type and huge volume, data ingestion and data retrieval is important activity during this process data security is a vital to protect user data, in connection with this, authors proposed a big data security architecture using split and merge security method in big data environment using hadoop.This work will help Data security professionals and organizations implementing big data projects.

Corresponding Author:

Archana R A,
Research Scholar, R& D Centre, Bharathiar University,
Coimbatore, Tamil Nadu, India.
Email: archana.tnm@gmail.com

1. INTRODUCTION

In this digital era, requirement of huge data is increasing rapidly, so are the security concerns. Everyone wants data to be safe with every aspect and dimension. Breach of privacy and security is one of the major concerns in the digital world we live in today. We have large amounts of data that is generated in the field of media, healthcare, technology, private sector, science, sports and security has to be maintained with every aspect. Big data refers to the management and analysis of huge amounts of data that exceeds the capability and efficiency of traditional data in every dimension. Big data collects and analyses large amounts of data from heterogeneous sources to discover unprecedented new knowledge and understanding of scientific and business scenarios[1][2]. Aiming to gain greater insight into patterns not generally discernible from smaller data sets, big data business intelligence enables visibility into associations and trends that otherwise go unnoticed. Designing a fool proof security measure for such a wide volume and variety is an extremely intricate and cumbersome procedure. In this paper, we propose a schema that encrypts data and also makes data access difficult to an intruder. This is achieved by dividing or splitting the concerned data following which, encryption is performed. This approach proves to be beneficial since it not only encrypts data but also focuses on data integrity. To an intruder it may seem like the data integrity is lost since the data is divided, but the information about the file locations and pairing is known only to an authorized user. Hence a two level security is established, one that scrambles or masks the data items and one that protects its integrity under attack [4][5]. Data Splitting, Encryption and Decryption. Encryption performed by using Secure Efficient data distribution algorithm and decryption is performed using Efficient Data Conflation algorithm. The data splitting is achieved through various steps for different kinds of data. The goal of this mechanism is to increase efficiency of performance and security. A massive data breach all over the news across the world has led to large multinational companies making information security their top most priority. This is not just restricted to securing the actual information but also extended to authorizing each individual

user of the company servers. Sensitive information is further declared confidential and only a set of executive employees are granted access. With such security standards and protocols in place, security enforcements are undergoing a revolution in order to ensure top level data protection. Since big data analytics has invariably become one of the most sought after domains in the IT industry, big data security also assumes top priority. Big Data analytics reveals hidden data patterns, unknown co-relations, customer preferences and other useful information that aids the economic growth of a company. Hence every leading company is leaning on this technology to make better informed decisions, improve its operational efficiency, predict trends and over all improve its performance. There are three properties to big data, volume, variety and velocity. The variety arises by the presence of structured, unstructured and semi structured data. Providing a security solution that encompasses all the three types of data has been a challenging task. The proposed solution attempts to decrease the difficulty in performing this task. Most encryption algorithms provide a single level security by encrypting data using a random key. In the proposed solution, a two level security is provided, by encrypting the contents of a piece of information and also creating an illusion to an attacker that data integrity is lost[5][6].

Figure 1. Big Data Security for Applications

2. RELATED WORK

While the term “big data” is relatively new, the act of gathering and storing large amounts of information for eventual analysis is ages old. The concept gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three Vs: Volume. Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would’ve been a problem – but new technologies (such as Hadoop) have eased the burden. Velocity. Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Variety. Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions. We can consider two additional dimensions when it comes to big data: Variability. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks [12]. Is something trending in social media daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data. Complexity. Today's data comes from multiple sources, which makes it difficult to link, match, cleanse and transform data across systems. However, it’s necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control. Keke Gai, et.al, IEEE 2nd International Conference on Big Data Security on Cloud, IEEE International Conference on High Performance and Smart Computing, IEEE International Conference on Intelligent Data and Security, 2016[1][2], Security-Aware Efficient Mass Distributed Storage Approach for Cloud Systems in Big Data. Yibin Li et.al, Intelligent cryptography approach for secure distributed big data storage in cloud computing, 2016 Elsevier Inc, This paper focused on the problem of the cloud data storage and aimed to provide an approach that could avoid the cloud operators reaching user’ sensitive data. Addressing this goal, we proposed a novel approach entitled as Security-Aware Efficient Distributed Storage (SA-EDS) model. In this model, we used our proposed algorithms, including Alternative Data Distribution (AD2), Secure Efficient Data Distributions (SED2) and Efficient Data Conflation (EDCon) algorithms. Our experimental evaluations had proved that our proposed scheme could effectively defend major threats from cloud-side[3][4]. The computation time was shorter than current active

approaches. Future work would address securing data duplications in order to increase the level of data availability since any of datacenter's down will cause the failure of data retrievals. Ahmed Alahmadi, Mai Abdelhakim, Jian Ren, and Tongtong Li. In IEEE Transactions on Information Forensics And Security, Vol. 9, No. 5, May 2014 emphasize on, a reliable AES-assisted DTV scheme was proposed for robust primary and secondary system operations under primary user emulation attacks. In the proposed scheme, an AES-encrypted reference signal is generated at the TV transmitter and used as the sync bits of the DTV data frames[5][6]. By allowing a shared secret between the transmitter and the receiver, the reference signal can be regenerated at the receiver and be used to achieve accurate identification of authorized primary users. Moreover, when combined with the analysis on the auto-correlation of the received signal, the presence of the malicious user can be detected accurately no matter the primary user is present or not. The proposed approach is practically feasible in the sense that it can effectively combat PUEA with no change in hardware or system structure except of a plug-in AES chip. Potentially, it can be applied directly to today's HDTV systems for more robust spectrum sharing. : B Eswara Reddy, Gandikota Ramu et.al : IEEE 2nd International Conference on Big Data Security on Cloud, IEEE International Conference on High Performance and Smart Computing, IEEE International Conference on Intelligent Data and Security, 2016, stress on The fulfilment of data privacy and identity privacy are the major challenging issues in the EHR's integrity auditing system. In this study, we introduced a framework for secure auditing of EHR's in cloud servers [7][8]. The proposed framework uses CP-ABE scheme with two-authority key computation method. Therefore, the public auditor verifies the data without accessing complete data, and he cannot reveal the patient identity in any case. This framework allows the auditors to perform auditing task simultaneously to enhance the efficiency of auditing task. In this framework, the cloud or KGA does not access the plaintext individually, so there is no chance of misusing of data in the cloud, it is a very important feature to support healthcare data auditing system [9][10]. This analysis explains that the framework is secure and efficient.

3. PROPOSED METHOD

The proposed system architecture is illustrated in Fig.-1, The system accepts the big data input file and identify the sensitive data using attribute relation methods and split the bug file into two and mask the split file then transmit the file to the destination and damask appropriately and calculate the performance parameters after the merging the split file to get the original file. Our approach is designed to divide the sensitive data into two encrypted parts for distributed storage in two different physical locations. The inputs include the initial data that consist of a string of data packets, which contain sensitive information. The outputs are two separate data packets that will be transmitted to different physical locations. The new generated data packets need to perfectly hide the sensitive information so that an attacker cannot read the information even though he or she has access to the data.

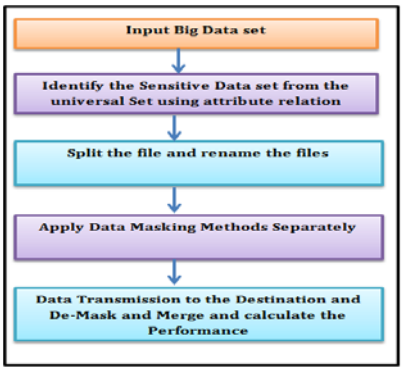

Figure 1: Architecture of the Proposed System

The proposed algorithm performs below steps:

1. Input the Big Data, α and β, that are acquired from different physical locations.
2. The key K from the registry that maintains keys.(Attribute Relation Method)
3. Initialize the datasets γ, γ 1, and D.
4. Split the file using XOR operation to both α and β by using the key, K. (Masking)
5. Merge γ, γ 1 to retrieve original data D. Output D. (Demasking)

In Masking Phase, the inputs of this algorithm include the two split components of the sensitive data that is to be encrypted and the randomly generated key.The two components are initialized as sets R and C for representation. Each of the two sets is subjected to XOR operation with the random key. This key is common for a given file input, but varies for different file inputs. It must be maintained in a special registry and communicated securely to a legitimate user of the information[1][2].The XOR operation results in two output sets namely, α and β. These are then saved in different physical locations within the target file system. This is depicted in Figure 2.

MASKING PHASE

Require: R, C
Ensure: α, β
1: Input R, C
2: Initialize α ← 0, β ← 0
3: Randomly generate a key K
4: FOR ∀ input data packets
5: IF R! = C && C! = 0
6: DO α ← C ⊕ K
7: β ← R ⊕ K
8: END IF
9: END FOR
10: Output α, β

Figure 2. Masking Phase

In demasking phase is designed to enable users to gain the information by converging two data components from distributed storage locations. Inputs of this algorithm include two data components, α, β, and the appropriate key, K. A log of all session keys must be maintained and tagged. According to the data to be decrypted the right key instance must be used. It is depicted in Figure 3.

DEMASKING PHASE

Require: α, β, K
Ensure: Original data, D
1: Input α, β, K
2: Initialize γ← 0, γ¹← 0, D ← 0
3: γ ← α (XOR) K
4: γ¹← β (XOR) K
5: D ← γ + γ¹
6: Output D

Figure 3. De-Masking Phase

The proposed model achieves data confidentiality with minimum overhead since an illegitimate user without the appropriate key cannot access the protected data. If in any case, an attacker obtains the key, he or she may not able to retrieve the complete information since the original data packet is split into components. The attacker is fooled into thinking the data has been altered. To retrieve the original data as it is, both the data components must be decrypted and merged in order to ensure that data integrity has not been lost.

4. RESULTS AND DISCUSSIONS

The proposed methods have been experimented on text file, image and audio files of size 1 MB data of each and executed the algorithms and extracted the time versus and size for individual data sets and provide flexibility around how the data will be masked and ensure that business rules of the enterprise application will not be impacted. A text file with a .txt extension is chosen at random and the text file is split into two components which is then encrypted using the above said algorithm. The Figure 4 represents the

Time vs Size graph that illustrates how the masking time behaves with increase in data size. As shown time is linearly proportional to the size.

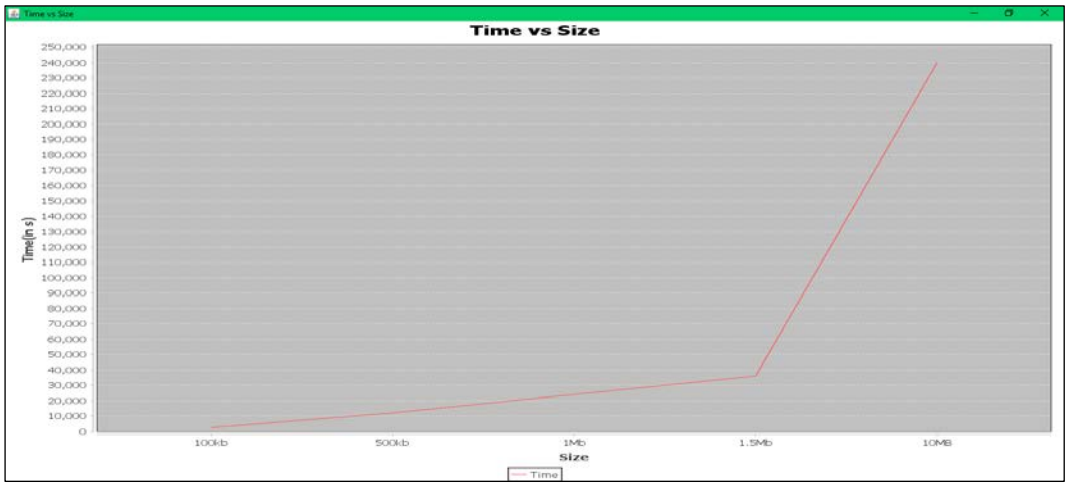

Figure 4. Time vs Size for Text File of 1MB Size

The image file with .jpg extension using the proposed algorithm. The Figure 5 repreents the Time Vs Size graph that illustrates how the masking time behaves with increase in data size. As shown time is linearly proportional to the size.

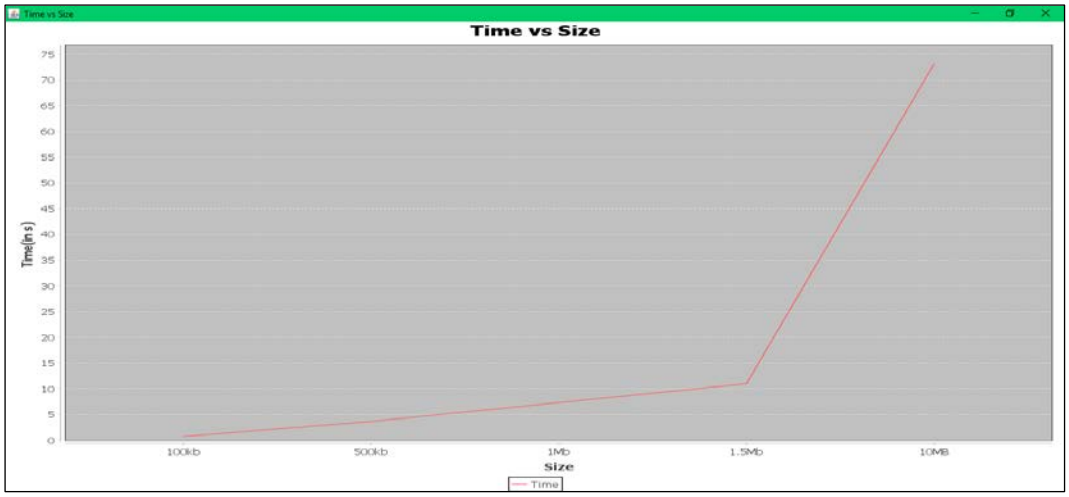

Figure 5. Time vs Size for Image File of 1MB Size

The Audio file with .avi extension using the proposed algorithm. The Figure 6 repreents the Time Vs Size graph that illustrates how the masking time behaves with increase in data size. As shown time is linearly proportional to the size.

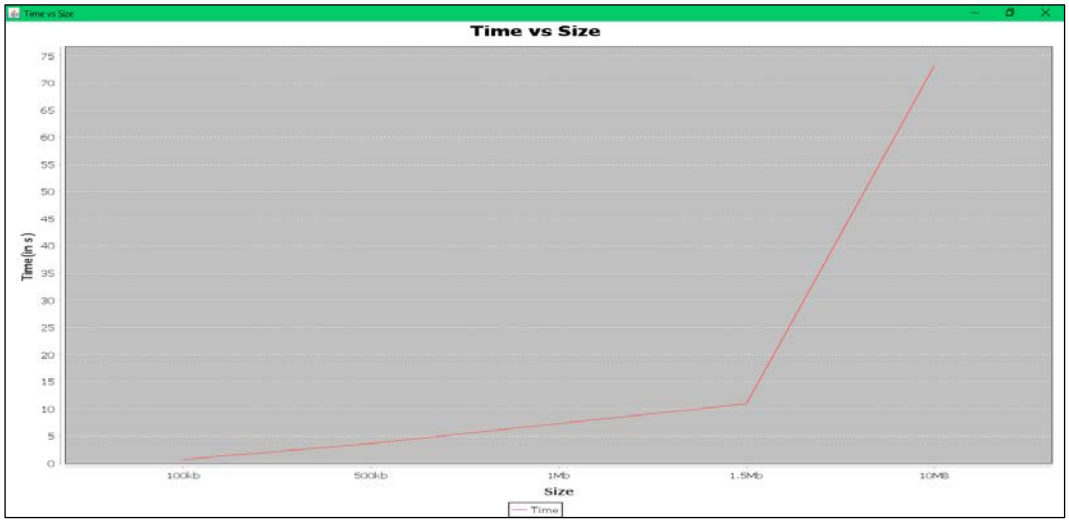

Figure 6. Time vs Size for Audio File of 1MB Size

The big data has heterogeneous data sets and securing such data sets is a challenging task in the current digital era. To achieve the said outcomes, organizations need to be able to handle it efficiently, quickly, and because often this data will include sensitive information needs security at different scales.

5. CONCLUSION

In current business scenario, onsite offshore model and factory model are familiar and many organizations hesitate looking at security factors, big data boom is increasing day by day which has different varieties of data sources which needs to be secured based on the customer expectations and application specific, the proposed solutions are concerned about deploying the proposed solution in the first place. In this work we focused on the problem of security of the variety of data that is present on a big data platform and aimed to provide an approach that could protect sensitive data. Addressing this goal, we proposed a novel approach entitled as split and merge method to evaluate the expected performance dimensions, we evaluated the proposed model by assessing its execution time and size while different input data sizes were operated. This model has proven to be successful in performing the masking and demasking of different file formats

REFERENCES

[1] Meenakshi RK et.al, "RTL Modelling for the Cipher Block Chaining Mode (CBC) for Data Security", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 8, No. 3, December 2017, pp. 709 ~ 711 DOI: 10.11591/ijeecs.v8.i3.pp709-711

[2] S Kumaraswamy et.al, "Secure Cloud based Privacy Preserving Data Minning Platform", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 7, No. 3, September 2017, pp. 830 ~ 838 DOI: 10.11591/ijeecs.v7.i3.pp830-838

[3] Keke Gai, Meikang Qiu, Hui Zhao, "Security-Aware Efficient Mass Distributed Storage", *2nd IEEE International Conference on High Performance and Smart Computing (HPSC)*, Newyork, 2016.

[4] Approach for Cloud Systems in Big Data, "IEEE International Conference on High Performanceand Smart Computing", *IEEE International Conference on Intelligent Data and Security*, 2016

[5] K. Gai, M. Qiu, B. Thuraisingham, and L. Tao, "Proactive attribute based secure data schema for mobile cloud in financial industry", In *The IEEE International Symposium on Big Data Security on Cloud; 17th IEEE International Conference on High Performance Computing and Communications*, pages 1332–1337, New York, USA, 2015.

[6] K. Gai, M. Qiu, L. Chen, and M. Liu, "Electronic health record error prevention approach using ontology in big data", In *17th IEEE International Conference on High Performance Computing and Communications*, pages 752–757, New York, USA, 2015.

[7] Y. Li, M. Chen, W. Dai, and M. Qiu, "Energy optimization with dynamic task scheduling mobile cloud computing", *IEEE Systems J.*, pages 1–10, Jun 2015.

[8] Y. Li, W. Dai, Z. Ming, and M. Qiu, "Privacy protection for preventing data over-collection in smart city", *IEEE Transactions on Computers*, PP: 1, 2015.

[9] X. He, C. Wang, T. Liu, K. Gai, D. Chen, and L. Bai, "Research on campus mobile model based on periodic purpose for opportunistic network", In *2015 IEEE 17th International Conference on High Performance Computing and Communications*, pages 782–785, New York, USA, 2015. IEEE.

[10] K. Gai and A. Steenkamp, "A feasibility study of Platform-as-a-Service using cloud computing for a global service organization", *Journal of Information System Applied Research*, 7:28–42, 2014.
[11] M. Qiu, Z. Ming, J. Wang, L. Yang, and Y. Xiang, "Enabling cloud computing in emergency management systems", *Cloud Computing, IEEE*, 1(4):60–67, 2014.
[12] C. Wang, Q. Wang, K. Ren, N. Cao, and W. Lou, "Toward secure and dependable storage services in cloud computing", *IEEE Trans. On Services Computing*, 5(2):220–232, 2012.
[13] M. Mozaffari-Kermani and A. Reyhani-Masoleh, "A lightweight high-performance fault detection scheme for the advanced encryption standard using composite fields", *IEEE Transactions on Very Large Scale Integration Systems*, 19(1):85–91, 2011
[14] M. Qiu, H. Li, and E. Sha, "Heterogeneous real-time embedded software optimization considering hardware platform", In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1637–1641. ACM, 2009.