ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

# Elliptic Curve Based Data Scrambling with Encryption for Security of Big data

Soorya M, Swaraj K.P

M. Tech Student, Dept. of Computer Science and Engineering, Govt. Engineering College, Thrissur, Kerala, India

Assistant Professor, Dept. of Computer Science and Engineering, Govt. Engineering College, Thrissur, Kerala, India

**ABSTRACT:** Big data is the new hot topic in the software society. The huge amount of data that cannot be processed using traditional methods are called big data. Useful information obtained after analyzing big data are used for various purposes such as predictive analytics, business applications etc. But now security of big data is a huge concern since these huge amount of data contain a lot of personal information that are to be kept as a secret. Many techniques has been proposed for securing big data. Elliptic curve based data scrambling with encryption in a hybrid cloud architecture is proposed here for securing the big data in an organization. The data entering an organization is classified into sensitive and non-sensitive data. The sensitive data is split into several parts and encrypted. The encrypted parts of sensitive data are stored in one of the secured nodes in the private cloud that resides in the organization while the non-sensitive data are encrypted and stored in the public cloud for analytics purposes.

**KEYWORDS***:* Data Scrambling, Hybrid cloud, Sensitive and non-sensitive data.

## I. INTRODUCTION

### *A. BIG DATA AND BIG DATA ANALYTICS*

Big data actually refers to a large set of data for which the traditional data processing methods are in adequate. The main characteristics of big data are its huge volume, velocity in which the data is changing and the variety of data that is available. It is said that in one second almost 1Tb of data is generated in the world. 80 percentage of the data generated is actually unstructured data. There are a lot of challenges that are to be faced when it comes to big data. Some of the challenges are analysis capture, search, sharing, storage, transfer, visualization, and information privacy.

Big data analytics refers to the process of collecting, organizing and analyzing large sets of data (called big data) to discover patterns and other useful information. Big data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Analysts working with big data basically want the knowledge that comes from analyzing the data. Enterprises are increasingly looking to find actionable insights into their data. Many big data projects originate from the need to answer specific business questions. With the right big data analytics platforms in place, an enterprise can boost sales, increase efficiency, and improve operations, customer service and risk management. Organizations say that applying big data analytics helps to improve customer retention and also help with product development and gain a competitive advantage. Notably, the business area getting the most attention relates to increasing efficiencies and optimizing operations. Use of big data analytics improve speed and reduce complexity.

### *B. NEED FOR BIG DATA SECURITY*

There has been an increasing interest in big data and big data analytics with the development of network technology and cloud computing. Today each and every organization is using big data analytics to increase their revenue by predicting future trends in the industry. Government also uses big data analytics for policy making and

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

other activities. But if this useful information get into wrong hands it will cause great chaos. Some of the data breaches are discussed below.

*a) PFChangs*

In June 2014, P.F. Changs China Bistro reported a security breach that affected customers at 33 restaurants located in 16 states. The intruder may have stolen some data from certain credit and debit cards that were used during an eight-month period from October 19, 2013 to June 11, 2014. The potentially stolen credit and debit card data included the card number and, in some cases, the cardholders name and/or the cards expiration date.

*b) Kmart*

In a filing in early October with the Securities and Exchange Commission, Kmart said a month-long data breach took place in early September. Debit and credit card numbers appear to have been compromised. Kmart, which is owned by Sears Holding Corporation, did not disclose the scale of the breach. They did not believe hackers were able to obtain social security numbers, email addresses, PIN numbers or personal information from the system

*c) Sony*

Data continues to come out about this November 24 Sony breach. In early December, hackers leaked five unreleased movies online and some employees Social Security numbers. The security firm Identity Finder found the hack exposed over 47,000 Social Security numbers, including over 15,000 current or former employees. In addition, these numbers appeared more than 1.1 million times on 601 publicly-posted files stolen by hackers. A significant number of files containing the Social Security numbers were accompanied by other personal information, such as full names, dates of birth and home addresses, increasing the chances of identity fraud

*d) Ashley Madison*

The Impact Team hacker group accessed the Ashley Madison user database, revealing financial records and other useful information, including the personal data of 37 million users. The group's manifesto uncovered the "full delete feature" was false and personal information of its users was kept on file.

*e) Anthem*

In February, Anthem made history as the largest healthcare breach ever recorded. Initially, Anthem estimated approximately 78.8 million highly-sensitive patient records were breached, but that quickly increased to an additional 8.8
to 18.8 million non-patient records. Anthem's attack was just the first of many healthcare breaches of 2015; CareFirst BlueCross BlueShield and the UCLA Health Systems were also hacked.

From these attacks it is clear that if big data security is compromised it would cause a chaos

## II. LITERATURE SURVEY

Lei Xu et al. [1] proposed a user role based methodology for big data security. In the User role based methodology for providing security to big data mainly focuses on providing Privacy Preserving Data Mining. Big data analytics is an extended version of data mining. In this approach mainly four different types of users are identified, namely Data Provider: the user who owns some data that are desired by the data mining task, Data Collector: the user who collects data from data providers and then publish the data to the data miner, Data Miner: the user who performs data mining tasks on the data, and finally Decision Maker: the user who makes decisions based on the data mining results in order

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

to achieve certain goals. All users care about the security of sensitive information, but each user role views the security issue from its own perspective. Data provider hides all the data which he/she wants to hide using DNT (do not track) or other similar methods. Data collector modifies the collected data so that it does not contain any sensitive information. Data miner protects the sensitive mining results from the untrusted parties. Finally the decision maker protects the mining results and make decisions without disclosing sensitive information.

Sung-Hwan Kim et al. [2] Proposed an attribute relationship evaluation methodology for big data security. Here an attribute of dataset is considered as an essential source for creating value inside a big data. In the attribute relationship evaluation methodology first all the attributes of the data are extracted and then their properties are generalized. After that the correlation between attributes are compared and their relationship is evaluated. Finally selected attributes are protected using security measures based on correlation evaluation.

Md.Rafiqul Islam et al. [3] proposed a method for providing security to unstructured and structured big data. By analysis of Big Data structured data can be separated using SQL queries and security can be provided to it. Since unstructured data contains data types like text, XML, email, image, video etc., after analyzing the data we can build a node of databases, which holds different kinds of data. An algorithm interfaces the data node with a security suite, which contains several security services to provide security to data. The suite has services to maintain privacy, integrity, integrity with authentication and non repudiation. Thus the approach includes two processes, one is to do data analytics and another is to build a security suite.

Jakrarin Therdphapiyanak et al. [4] proposed a methodology that apply Hadoop for log analysis for distributed intrusion detection. Log files are used to enhancing system security not only for tracing but also for analyzing and determining their activities to obtain the new knowledge that cannot explicitly be seen with simple analysis. These log files are transferred to a hadoop cluster and are preprocessed. With files in a cluster, distributed K-Means algorithm is applied. With the analysis in hands, a visual report is generated.

Apart from these algorithms and methods big data analytics results itself is used for big data security. Big data analytics is used for intrusion detection systems, for network security, for enterprise event analytics and for advanced persistent threat detection. Several products such as HPArcSight CORR Engine uses big data analytics for APT detection. Similarly another product from eBay is the Eagle. Eagle is a User Profile-based Anomaly Detection system for Securing Hadoop Clusters.

The rest of the paper is organized as follows. In Section III, a detailed description of the proposed elliptic curve data scrambling method is given. Section IV concludes the paper.

## III. METHOD

The proposed data scrambling method is actually for a hybrid cloud architecture. Since a huge amount of data is generated in each organization day by day storage of the data is an issue. Organization cannot completely rely on public clouds since there will be very sensitive information in the generated data. These data are to be protected. So a hybrid cloud architecture is suitable for an organization that has a lot of data to deal with.

### *A. SYSTEM ARCHITECTURE*

The proposed system architecture is shown below in the figure 1. The general system architecture shown in the figure has mainly five modules which are private cloud, public cloud, pre-analyzing module, Data scrambler or data splitter and finally the encryption module

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

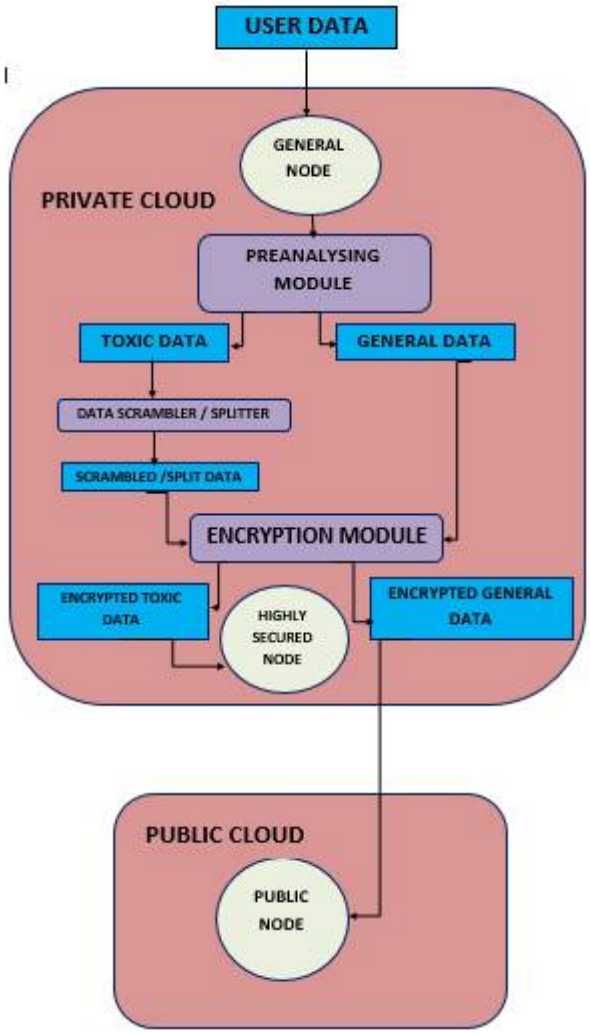


Fig 1. System Architecture.

The data generated from the users contain both toxic data and general data. Toxic data is the sensitive information that must be protected. It is termed as toxic because if it goes out of the organization it is harmful for both the user as well as the company. The toxic or sensitive data must be protected with in the organization and the general data can be stored in the public storage. So for an organization instead of maintaining a lot of data inside the organization using a lot of money and manpower, a simple NIST hybrid cloud architecture can be used. By using the hybrid cloud architecture the sensitive or toxic data can be stored in nodes inside the organization. These nodes collectively form a private cloud. Private cloud generally consists of 2 types of nodes: the general node where all the user data gets stored at first and the highly secured node where the sensitive information are stored after encryption. The latter node is highly secured because it has additional authentication measures so that it can only be used by high officials in the organization and any traffic to/from that node will be monitored. The general data can be stored in the public cloud which resides outside the organization. So this data can be accessed from anywhere for other purposes by authenticated people.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

### B. *PREANALYSING MODULE*

The data generated by the user contain both structured as well as unstructured data. The toxic data or sensitive information from the data are to be extracted at first. The pre analyzing module perform simple analytics and extracts or separate the toxic data such as credit card numbers, personal identification numbers(aadhar number), bank account numbers etc and general data from the user data which is stored in the general node.

### C. *DATA SCRAMBLER OR DATA SPLITTER*

Toxic data obtained after analyzing the general data are not encrypted as such. It is scrambled into several pieces and then encrypted and each encrypted piece is stored in a different location for high security. The sensitive information is divided and stored because if an attacker somehow find a way and compromise the security of one location it will not badly affect the organization since the information obtained by the attacker will not be useful. He cannot do anything with a partial information as it is tiresome and expensive to check every possibilities for completing the information. Elliptic curve based data scrambling method is used here as it is more efficient and secure.

*Elliptic curve data scrambling method*:

In elliptic curve data scrambling method at first an elliptic curve is fixed and all the points are generated using the equation $y^2 = x^3 + ax + b$. For example for an elliptic curve $E_{13}(1,1)$ a=b=1 and 13 represents that the addition is followed by modulo 13. The points generated for x values up to 13 will be (0,1) , (0,12) , (1,4) , (1,9) , (4,2) , (4,11) , (5,1) , (5,12) , (7,0) , (8,1) , (8,12) , (10,6) , (10,7) , (11,2) , (11,11) , (12,5) and (12,8). Here there are 17 points so sensitive data can be split into 17 parts and each part is placed in a corresponding point sequentially (first chunk on (0,1) next on (0,12) and so on). After that for shuffling either parts can be rejoined in the decreasing order of Y coordinate along with increasing order of X coordinates or using any new technique.

### D. *ENCRYPTION MODULE*

The sensitive information that is scrambled can be encrypted using any simple algorithms. Because already the data is scrambled so it is somewhat secure and again after encryption it will be stored in a highly secure node which has additional authentication measures and real time traffic monitoring mechanism. So the encryption method can be relatively simple based on the sensitivity level of the data.

The general data can be encrypted using DES or more secured algorithm since it has to be stored in a public cloud. The cloud service providers does not fully guarantee the safety of our data so as a precautionary measure we can provide a medium level encryption for the general data. Before applying encryption algorithms we can apply elliptic curve based data scrambling for the general data also if found necessary.

### E. *MAP REDUCE APPROACH FOR ELLIPTIC CURVE BASED DATA SCRAMBLING WITH ENCRYPTION FOR SECURITY OF BIG DATA*

Big data cannot be processed using traditional methods. Map reduce approach helps to break large problems or tasks into smaller individual tasks and helps to process it in parallel using distributed collection of computers. Thus the final result will be the aggregation of the individual results. Map reduce consists of two tasks: map and reduce. Both steps take a (key,value) pair as input. Map task actually does the processing. Reduce is an aggregation function which aggregates the output of the map tasks according to a specified criterion.

Here in this approach the sensitive data is split into separate chunks. The number of chunks depend upon the number of points in the elliptic curve. The map task is given the (point, chunk) pair as (key, value) input. The point corresponds to the elliptic curve number and the point on which the data chunk is placed. The map task performs the encryption and also gives the (scrambled order, encrypted chunk) output. The scrambled order highlights the Y coordinate and the corresponding X coordinate for the reduce tasks. The reduce task aggregates or rejoins the encrypted data chunks based on the scrambled order.

Thus the elliptic curve based data scrambling with encryption can be done in parallel using Map reduce framework.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

## IV. CONCLUSION

Big data is widely used for various business purposes. But in the view of hackers big data serves as an ocean of valuable credentials of private individuals. If the big data is not secured adequately or if its security is compromised then it will create a great chaos to both organizations and public. Although there are various products providing security for the data in an organization they do not guarantee complete safety. All the products aim at detecting and preventing intrusions. But if once the intrusion is not detected then the big data is compromised due to its weak security.

The proposed elliptic curve based data scrambling method along with suitable encryption method is keeps the data really safe. The scrambled pieces of data are stored in multiple locations within the private cloud. So even if the security of one location is compromised the attacker will not get full information and thus the partial information will be useless. Also the general data stored will be having dual security. The security provided by the cloud service provider will be there in addition to the encryption applied by the organization. Thus this method secures the big data of an organization efficiently.

## REFERENCES

[1] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan,Yong Ren, " Information Security in Big Data: Privacy and Data Mining",IEEE Access, 2014.

[2] Sung-Hwan Kim, Nam-Uk Kim,Tai-Myoung Chung, "Attribute Relationship Evaluation Methodology for Big Data Security", IEEE, 2013.

[3] Md.Rafiqul Islam, Md.Ezazul Islam, "An approach to provide security to unstructured Big Data" ,IEEE, 2014.

[4] Jakrarin Therdphapiyanak, Krerk Piromsopa,"Applying Hadoop for log analysis for intrusion detection",ICUIMC(IMCOM)ACM Conference,2013.

[5] N. Bakibayev, D. Olteanu, and J. Zavodny, " Big Data Analytics for Security Intelligence", Cloud Security Alliance Business white paper, 2013.

[6] Randy Franklin Smith , Brook Watson, "3 Big data Security Analytics Techniques You Can Apply Now To Catch Advanced Persistent Threats", HP business white paper, 2013.

[7] Chaitali Gupta, Ranjan Sinha, Yong Zhang,"Eagle: User Profile-based Anomaly Detection for Securing Hadoop Clusters",IEEE International Conference on Big Data (Big Data),2015.

[8] K. S.Vijayanad , T. Mala, "Ecids Elliptic curve inspired data scrambling method for securing data in hybrid cloud using N cloud architecture",Asian journal of Information technology, 2016.