International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Privacy Preserving for Sensitive Data using Data Masking Technique

Madhurya J A, Beena G Pillai

Abstract : ***In recent times, most the people are using internet where they are going to share sensitive information with other individual or with an organization like hospital, banking sector or business companies. So such huge amount of information will be stored on cloud. The attackers may try to hack the sensitive data and will try to misuse that data. So here the security for data comes first. There are numerous methods available to provide security for the data that is being shared among individuals or organizations. Most of the organizations take enough precautions to secure data that is shared with third party organizations. In recent times providing privacy for the sensitive data is high priority. The objective of this research is to discover the various data masking solutions for different applications for providing security to the data. Established data privacy method like AES or DES encryption technique proves to be proficient but time consuming. In order to avoid time consumption and to provide privacy for the data being shared, this paper proposes a information hiding method based on format-preserving encryption for sensitive data. This method will masquerade only sensitive data and make sure the encrypted data is still in the original format where it doesn't consume much memory space. Organization like hospitals or banking sector or any business companies can use this format-preserving method to enhance the security of the data being shared. Tested the information on Spark illustrate that information hiding method based on format-preserving encryption can provide data privacy for sensitive data and preserve data format.***

Keywords : big data, data mask, format-preserving, substitution.

I. INTRODUCTION

The meaning of responsive information is broad enough and differs from nation to nation, business to business and person to person. In some nation like United States – the person's employee id is most important data. Similarly in medical field health data of patients is considered as sensitive information which will be shared among different doctors for reviews. So for such sensitive data security is needed. Data masking means hiding original data with modified data[2]-[3].Meaning, important data is modified with realistic but original information is not used for testing thus achieving both the objectives – preserving important information and providing that tested information is true.

There are multiple methods to implement information hiding technique on sensitive data to provide privacy. It can be a replacement method for existing data with likely random data or shuffle of definite letters or digits, producing a new data. Alternately, it could be composite as use algorithms to shuffle or replace a original data with a random data generated by a algorithm , thus providing a security to the original information.

There are some information security methods which are used to encrypt huge amount of data and has been listed out. The established information encryption method can encode the information in irreversible format, like AES algorithm which is used to encode the id field. This can hide id and differentiate dissimilar persons, but the outcome is a binary bit string, that doesn't have the original information structure, so it can not be stored in the database nor it can be recognized as sensitive data. General industry applications need fixed patch and cycles and requires 6-8 copies of the function and information be used for testing mask algorithms so referential integrity is maintained.

In our proposed approach a information hiding method based on Format-Preserving Encryption (FPE)[1], which not only preserve data privacy and also preserve the data format. It is completely different form traditional encryption method where traditional encryption retains the original composition and arrangement of the normal text in illegible binary format. Our proposed FPE is completely different from traditional encryption method Format-Preserving Encryption is applied to resolve the confidentiality crisis in record and other application. Record and applications structure doesn't to be modified to store cipher text because the result of FPE will have same extent and structure as normal text. FPE can also be used to hide information[6] for performance testing and secure testing, which can avoid the original information from privacy disclosure.

II. RELATED WORK

Information veiling could be a prepare of de-identifying or darkening particular information inside a particular information component inside database table or column. In other words information veiling is the substitution of existing touchy data in test or improvement databases with data that looks genuine but is of no utilize to anybody who might wish to abuse it. In common, the clients of test, advancement or preparing database[8] don't have to be see the real data as long as what they are looking at looks genuine and is steady.

Manuscript received on March 15, 2020.
Revised Manuscript received on March 24, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Madhurya J A, Dept. of CSE, Gitam University, Bangalore, India.
Beena G Pillai, Dept. of CSE, Gitam University, Bangalore, India.

Successful information concealing requires information to be changed in a method that the genuine value cannot be decided or reengineered, utilitarian look is kept up, so compelling test is conceivable.
Information can be scrambled or decoded, social astuteness is kept up, security polices can be created and partition of obligation between privacy and business is set up.

Data Masking Techniques:

Right now numerous information concealing strategies are accessible within the industry and taking after are the vital information veiling procedures.

A. Substitution:

The replacement procedure replace the accessible information with arbitrary values from a pre-prepared dataset i.e. this strategy comprises of haphazardly supplanting the substance of a column information with data that looks comparative but is totally irrelevant to the genuine points of interest[5]. For case, the surnames in a client database may be sanitized by supplanting the genuine final names with surnames drawn from a largish irregular list. Substitution is exceptionally viable in terms of protecting the see and feel of the existing information. The drawback is that a largish store of substitutable data must be accessible for each column to be substituted.

B. Shuffling*:*

The Rearranging method employments the accessible information as its claim replacement dataset and move the value between lines in such a way that the no value are shown in respective unique columns i.e. rearranging is comparative to substitution but that the substitution information is determined from the column itself[6]. Mainly the information in a line is arbitrarily moved between lines until there's now not any reasonable relationship with the remaining data within the push .There's a certain peril within the rearranging method. In other words, the initial information is still display and now and then significant questions can still be inquired of it. Another thought is the calculation utilized to rearrange the information. In the event that the rearranging strategy can be decided and after that the information can be effortlessly "un-shuffled".

C. Number and Date Variance:

The Number and Date Change strategy changes the existing input in a indicated run to arrange it in jumble way. Example, birth date values might be changed inside a extend of +/- 60 days. The numeric change procedure is valuable on numbers or date information[9]. This calculation includes adjusting each number or date esteem in a column by a few arbitrary rate of its genuine esteem. This procedure has advantage of giving a sensible camouflage for the information whereas still observance the run and conveyance of value within the line to inside offered limits.

D. Encryption:

These Encrypt processes algorithmically scramble the data. more often than does not take off the data looking realistic and can now and then make the data larger. This strategy offers the choice of taking off the information in put and obvious to those with the fitting solution whereas remaining viably futile to no one without the solution[13]-[15]. This would emerge to be a really great alternative however, for mysterious database; it is one of the least valuable techniques.

The preference of having the genuine information accessible to anyone with the input is really a major impediment in a test or advancement database. The "optional" perceive ability gives no major advantage in a test framework and the encryption secret word as it were ought to elude once and all of the information is compromised.

Of course, you'll be able alter the key and recover the test instancegs – but outsourced, put away or spared duplicates of the information are all still accessible beneath the ancient password[7]. In today's data age, the information is an critical resource of the organization so the security of the data could be a imperative part within the industry. In arrange to realize the over aspect, the information veiling is used. The information masking is primarily the information is supplanted with practical but not the initial information. The most objective is to form delicate data isn't made accessible exterior of the environment. The information concealing is to fair give the duplicate of the generation information in back of the improvement environment and in this way it controls the spillage of the information.

Information concealing are planned to be repeatable so referential astuteness is kept up. We have investigated the information concealing engineering, strategies with reasonable information and arrange of veiling.

III. THE CHALLENGES OF MASKING DATA

Organizations has attempted to concentrate on and actualize all these issue with tradition hand-crafted arrangements or using existing information control instruments inside the venture to fathom this issue of sharing touchy data with the persons who are not clients. Example, the foremost frequent arrangement: catalog scripts. To begin with look, a benefit of the catalog program approach is to illustrate that particular information needs special protection. These sensitive data will be chosen by DBA to run at the earliest. Let's see at the issues with this approach base[16].

A. Reusability:

Since the rigid connection is established between a program and with the related record, these programs have to be re-written from beginning to the other connected record. There is no common features in a program that can be reused by other databases.

B. Transparency:

Thus script be likely to be solid set of instructions, reviewers are not straight forward to concealing strategies utilized within the program. The reviewers will discover amazingly troublesome to offer any proposal on whether the concealing prepare built into a script is secure and offers the venture the fitting degree of assurance.

C. Maintainability:

While these enterprise applications are overhauled, modern table and column containing touchy information can be included as a part of the update handle. With a script-based approach, the whole program will be returned and overhauled to suit vacant table and column included as a portion of an application fix or an update.

IV. IMPLEMENTATION

Below Fig1 shows the complete process which will be used to justify linear based Models and Data Shuffling methods are used to pull the data from Production environment to QA Environment with examples and their comparison study between masked and unmasked data selected from data table , those data was sensitive information. Like, in an employee file the attributes are staff id and statistical performance of the original data and modified data. The algorithm repository and metadata repository as two entities from these data create as association rules used for masking using info-mask Rule Engine then store it in Mask rule Database and other side production data and archived mask data is compared and replaced and replicated as masked data with high security method, here we have used replacement data masking method for masking.

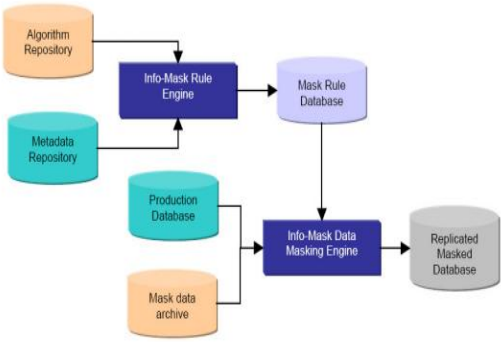


Fig.1 Complete process of data model

V. RESULT

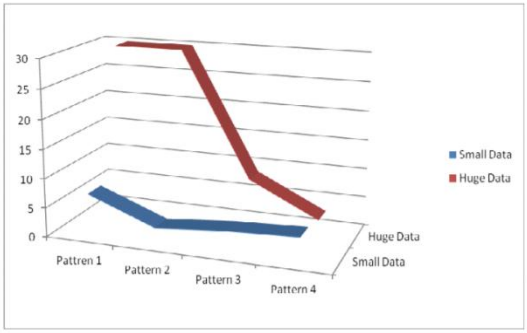


Fig. 2 Random replacement

A detail evaluation of methods used and proposed by us: The Fig.2 show about the improvement and projected method called Random replacement. Based on detail study and analysis of the usage of different method used across the industries, Random Replacement would be far better in terms of performance and data security [12]-[15] is shown in Fig.3

	Shuffling	Substitution	Null out	Replacement	Random Replace method
Banking	4.3	2.4	2	6.4	9
Finance	2.5	4.4	2	8	9.2
Insurance	3.5	1.8	3	7	9.5
Securities	4.5	2.8	5	8	9.7

Fig.3 Random replacement in terms of security

VI CONCLUSION

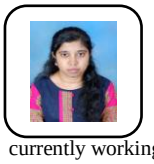
Now a days sharing of information is common, so need to give security for the sensitive information like employee information or customer information which will be shared by person's or authorities for business or with organization. In recent times, most of the sensitive data is hacked by hackers because malicious individuals easily access the servers where data is stored and misusing the hacked data. This causes bad impression on the organizations and effect business where no individual will be ready to share their information as there is no security of the shared data. Hence providing security to the information is crucial and it can be achieved by information hiding technique for the original data.

REFERENCES

1. Bellare, M., Ristenpart, T., Rogaway, P., Stegers, T.: Format-Preserving Encryption. In: Jacobson, M.J., Rijmen, V., Safavi-Naini, R. (eds.) SAC 2009. LNCS, vol. 5867, pp. 295–312. Springer, Heidelberg (2009).
2. Radhakrishnan, R., Kharrazi, M., &Memon, N. (2005). Data masking: A new approach for steganography?. The Journal of VLSI Signal Processing, 41(3), 293-303.
3. Lee, J. K., Koo, B., Roh, D., Kim, W. H., & Kwon, D. (2014, December). Format-Preserving Encryption Algorithms Using Families of Tweakable Blockciphers. In International Conference on Information Security and Cryptology (pp. 132-159). Springer International Publishing
4. HemaShekhawat, Samiksha Sharma, ReetikaKoli. Privacy-Preserving Techniques for Big Data Analysis in Cloud. 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP) 978-1-5386-7989-0/19 ©2019 IEEE.
5. Suriyapriya, M. and A. Joicy, "Attribute Based Encryption with Privacy Preserving In Cloud," International Journal on Recent and Innovation Trends in Computing and Communication, 2014, ISSN: p. 2321-8169.
6. Jaiman, V. and G. Somani, "An Order Preserving Encryption Scheme for Cloud Computing," in Proceedings of the 7th International Conference on Security of Information and Networks, 2014, ACM.
7. Li, M., Liu, Z., Li, J., &Jia, C. (2012). Format-preserving encryption for character data. JNW, 7(8), 1239-1244.
8. Osama Ali (Ozkan) and AbdelkaderOuda. A Content-Based Data Masking Technique for A Built-In Framework in Business Intelligence Platform2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE) 978-1-5090-5538-8/17/$31.00 ©2017 IEEE
9. G. K. Ravikumar, T. N. Manjunath, S. HegadiRavindra, I. M Umesh, "A Survey on Recent Trends, Process and Development in Data Masking for Testing", (IJCSI) International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.
10. S.Vijayarani, Dr.A.Tamilarasi . An Efficient Masking Technique for Sensitive Data Protection. IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011 978-1-4577-0590-8/11/$26.00 ©2011 IEEE MIT, Anna University, Chennai. June 3-5, 2011
11. Jeremy Kepner, Vijay Gadepally, Pete Michaleas, Nabil Schear, MayankVaria, ArkadyYerukhimovich, and Robert K Cunningham, "Computing on masked data: a high performance method for improving big data veracity," in the Proceedings of High Performance Extreme Computing Conference (HPEC), 2014.

12. DiaaSalama Abdul. Elminaam1, Hatem Mohamed Abdul Kader2 and Mohie Mohamed Hadhoud3," Performance Evaluation of Symmetric Encryption Algorithms", IJCSNS International Journal of Computer Sci 280 ence and Network Security, VOL.8 No.12, December 2008.
13. Md Imran Alam, Mohammad RafeekKhan,"Performance and Efficiency Analysis of Different Block Cipher Algorithms of Symmetric Key Cryptography", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.
14. ShivlalMewada, Pradeep Sharma, S. S. Gautam"Exploration of Efficient Symmetric AES Algorithm", IEEE Conference Publications, March 2016.
15. A.Nadeem, "A performance comparison of data encryption algorithms", IEEE information and communication technologies, pp.84-89, 2006.Bn.
16. Chen, Deyan, and Hong Zhao. "Data security and privacy protection issues in cloud computing." In Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on, vol. 1, pp. 647-651. IEEE, 2012.

AUTHORS PROFILE



Madhurya J A received the B.E degree in Computer Science & Engineering from Visvesvaraya Technological University, Belgaum in 2011, the M.Tech degree in computer science and Engineering from Visvesvaraya Technological University, Belgaum, in 2017. She is currently working as an Assistant Professor in the Department of Computer Science and Engineering, Gitam University, Bangalore. Her current research focuses on the security in Cloud Computing, BigData, IoT.

Beena G Pillai received the B.Tech degree in Computer Science & Engineering from Acharya Nagrajuna University, Guntur, in 2012, the M.Tech degree in computer science and Engineering from Jawaharlal Nehru Technological University, Anantapur, in 2015. She is currently working as an Assistant Professor in the Department of Computer Science and Engineering, Gitam University, Bangalore. Her current research focuses on the security in Block chain technology, Cloud Computing and Cyber Security