A Robust Approach to Secure Structured Sensitive Data using Non-Deterministic Random Replacement Algorithm

Ruby Bhuvan Jain
Research Scholar, JSPM's Abacus Institute of Computer Applications,
Pune, Maharashtra, India

Manimala Puri, PhD
Director, JSPM Group of Institutes, Survey No.80, Pune-Mumbai Bypass Highway,
Tathawade, Pune,
Maharashtra, India

Umesh Jain
Vice President, Deutsche Bank,
Pune, Maharashtra, India

ABSTRACT
The first list of Jan 2018 is one of the longest lists, with a count of 7,073,069cases, which include Cyber attacks & ransom ware, Data breaches, financial information, and others.Security and risk management leaders should use data masking to desensitize or protect sensitive data and address the changing threat and compliance landscape. Masking is a philosophy or new way of thinking about safeguarding sensitive data in such a way that accessible and usable data is still available for non- production environment. In this research paper authors proposed a dynamic data masking model to protect sensitive data using non-deterministic randomreplacement algorithm. This paper contains comparative analysis of proposed model with existing masking methods and result shows that proposed model is would be superior in terms of sensitive data discovery, dynamic data masking and data security.

Keywords
Data Discovery, Sensitivity Diligence File, Testing Integration, Security Integration.

1. INTRODUCTION
The data masking technology market is growing as industries ranging from Banking, Financial Services and Insurance (BFSI) to government are becoming extremely cautious to the internal hacking and data privacy concerns. Recent technologies like big data where massive databases are generated, exposing it to greater risk.

In order to address these concerns for preventing outside attacks, data masking technology is used which can be used to and analyze the data in proxy.

Data masking technology provide data security by cloning the original data into a non-sensitive proxy which look like similar data. This non-sensitive data can be used in business process for testing and analysis without the risk of breaking the business. Dynamic masking is the recent trend in database access.

The intention of data masking is to limit the exposure of sensitive data and at the same time data is also available at non-production environments. There are many environments where it is necessary to hide the data and to use that data for processing purposes. New Development or enhancement of existing applications or in security controls to existing environments are the most common examples where data security is most important concern. Developers are not allowed to see detailed personal information but at the same time they may need that data, or its equivalent, in order to test the application they are developing. The technology, or technique, used to enable this functionality is known as data masking.

Data masking originally emerged as a complement to test data management, to protect sensitive data from unauthorized eyes. Data masking (DM) is a technology aimed at preventing the abuse of sensitive data by giving users fictitious data instead of real sensitive data. It aims to deter the misuse of data at rest, typically in nonproduction databases and data in transit, typically in production databases. DM is not the same as encryption or tokenization, although masking vendors may also offer encryption or tokenization. Static masking is a non reversible process in which the data undergoes a one-way transformation. Tokenization and format-preserving encryption (FPE) are alternative methods that are designed to be reversible, but this reversibility may increase the risk of secrecy and privacy violations. Depending upon requirements, data masking can be implemented using multiple ways.

Data masking is primarily used by security and risk management team to protect sensitive data and address the regulatory standards and compliances. The Health Insurance Portability and Accountability Act (HIPAA), Payment Card Industry Data Security Standard (PCI DSS), Gramm-Leach-Bliley Act (GLBA), Federal Information Security Management Act (FISMA) and Statement on Auditing Standards (SAS 70) and others, enforces that organizations must establish compliance frameworks for data security. Data Masking enables regulatory compliance through rule-based application of masking formats across enterprise-wide databases.

Essentially there are various data, which are required masking. The first is data about individuals, which includes credit card numbers, social security numbers, health and medical information, address and postal details and so on. This is generally referred to as personally identifiable information (PII) or, in the case of healthcare, personal health information (PHI). The main reason for masking such data is to comply with data privacy regulations but also to avoid the sort of bad publicity and costs associated with data breaches caused by hacking and other cyber attacks.

With more structured and unstructured data in enterprise databases, companies need simple and consistent tools to comply with data privacy regulations and mask sensitive data during application development, testing or data analysis. Adopting data masking helps enterprises raise the level of security and privacy assurance against abuses. At the same time, data masking helps enterprises meet compliance

requirements with the security and privacy standards recommended by regulating/auditing authorities [1].

Several methods, meeting different requirements have been proposed to improve data confidentiality. Among them, a handful of significant researches are presented in this segment; Ravikumar G K et al [6] proposed random replacement techniques and also represented a comparative study of various data masking techniques with proposed technique. The study is conducted in many domains like finance, banking and security. Results showed in the proposed method are better in terms of performance and data security. The research is applied only on statistical parameters; it can also be implemented on analytical parameters.

Waleed Ahmed, JaganAthreya[7], focused on challenges of masking. Authors also briefed on issues like reusability, transparency and maintainability. The model proposed in this paper is named as FAST. This is a 4 step comprehensive approach, FAST starts with discovering with sensitive data, accessing the masking algorithm, executing the high performance mask algorithm and integration testing of solution for quantity.

AshaKiranGrandhi et al [5], proposed a self-defined algorithm, named PSO optimization technique which included clustering, encryption and distribution privacy techniques, to cluster medical data to check accuracy of resultant data. The result could contain data set from different fields and of different sizes. At the same time, encryption time complexity could be considered.

XiaolingXia et al [11] suggested two methods m-invariance and NCm-invariance to convert the numerical data into categorical data to overcome from the defects.

Adrian Lane [9], focused on data masking, how the data masking technique work. Different types of Static and dynamic data masking techniques are discussed.

In this paper, the researcher's objective is to design and develop a model for protecting sensitivity of the given data. In this work, bank data sets are considered (UCI repository) for testing and evaluating the work (UC Irvine machine learning repository) [4]. In this present paper, the researcher adopts non-deterministic random replacement algorithm.

2. TECHNICAL ARCHITECTURE OF PROPOSED MODEL

The world is turning increasingly disruptive and volatile. Today's business and customers are smart, connected and tech savvy. This results a great amount of digital data, which requires security in many aspects. Dynamic data masking solutions safeguard sensitive and personal data from breach. Furthermore, increasing need for protected big data by dynamically masking sensitive information is further expected to observe a robust growth during the forecast period. In this work authors proposed three-tier technical architecture in order to secure sensitive data.

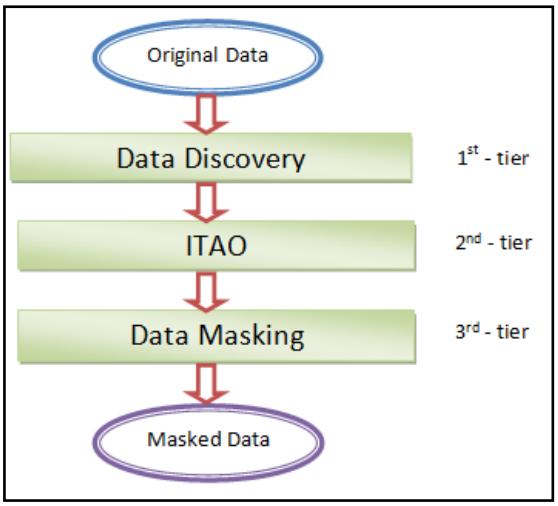

Figure 1: Technical architecture

Figure 1 contains the technical architecture of this research work. The main objective of this work is to design an effective data masking tool to secure sensitive data from unauthorized access.

The above design has three-tier of implementation. Three-tier architecture is a software architecture pattern in which the user interface (presentation), functional process logic ("business rules"), computer data storage and data access are developed and maintained as independent modules, most often on separate platforms. In the figure 1, the 1st tier focuses on discovery sensitive data from the given data set. The 2nd tier is the Information Technology Application Owner (ITAO). ITAO is the subject matter expert of the given dataset. The 3rd tier focuses on masking algorithm used to secure sensitive data.

3. DIFFERENT COMPONENTS OF TECHNICAL ARCHITECTURE

The technology solution proposed in this work has following components:

3.1 Data Discovery:

In the process of data discovery, source data files from the actor are required. Based on discovery algorithm, potential sensitive fields are identified and outcome is Sensitivity Diligence File. This file contains complete dataset which is used for data masking. To provide ease to actor, research here proposed a master masking file. This file contains name of input data file in first row along with extension, file delimiter data into second row, extension/ type of expected output file in third row and name of sensitivity diligence file in fourth row.

3.2 Sensitivity Diligence File:

This input file has information about sensitive and regular data columns. Intensity is defined in three parameters- Mandatory, Optional and Not Applicable. Mandatory specifies that the particular column contains highly sensitive information and mustbe masked. Optional specifies that the particular column contains moderate sensitive information, and should be masked. Not Applicable specifies that the particular column contains non sensitive information, and there is no need to mask that column. Sensitivity Diligence File will have all column names in first row and intensity parameter in second row. To specify sensitivity 'Y' or 'y' is the symbol and for non sensitive data 'N' or 'n' is used.

Respective attribute in first row will have 'Y' or 'N' in second row to check that the column is to be masked or not. The third row have option for conditional masking like Male/ Female replace with 0/1 and likewise.

3.3 Information Technology Application Owner (ITAO):

This sensitivity diligence file generated by data discovery tier, needs approval of Information Technology Application Owner (ITAO) before being masked because ITAO knows end to end flow of data in application and corresponding data integrity in application [8]. ITAO primary job is to re-confirm the number of columns identified by Discovery Engine required for masking. Once approved appropriate masking algorithm can be identified and applied.

3.4 Data Masking:

This step reads third row of sensitivity diligence file. If special masking conditions are requested, then associated algorithm is identified and applied. For example, in case of credit card number first four digits should be kept AS IS and rest digits should be masked, in case of Male/Female they can be replaced with 0/1. One more input file required for the system is Master masking file. This file has list of input files-source data files in the first row, delimiter in the second row, source data file extension in third row and name of output file in the last row. This step reads Master masking file and sensitivity diligence file as input. The algorithm generates data in separate directory called as output to distinguish the masked data from original data. Header and Trailer information part of source data file is added to targeted output masked file and as per information mentioned in the sensitive file about sensitive columns, default non deterministic random masking algorithm will be applied.

The masked data file is observer for data usability, reversibility, robustness and lossless of data.

4. NON-DETERMINISTIC RANDOM REPLACEMENT ALGORITHM

Algorithm: Non-deterministic Random replacement.

Purpose: To raise the level of security of sensitive data.

Input: Bank Data

Output: Masked data.

Begin:

Step1: Create output file with the same name as input file.

Step 2: Add headers of the input file into output file.

Step 3: For each row:

For each column:

a. For 'N' in second row
b. If masking not required copy the data from input file and paste into output file.
c. Else

For each character:

1. If uppercase replace with any random uppercase character other than character in memory
2. Elseif lowercase replace with any random lowercase character.
3. Else if number replace with any random number. If new masked first digit is 0, then replace with some other random number.
4. Else special symbol no replacement.
5. Check the length of input string with output string.
6. First Char in numeric field will have non-zero digit
7. To change Year field in Date, use Year@Range@format example Year@1920-2016@mm/dd/yyyy
8. To change Month field in Date, use Month@Range@format example Month@03-12@mm/dd/yyyy (02 is avoided due to leap year)
9. To change Day field in Date, use Day@Range@format example Day@1920-2016@mm/dd/yyyy (30,31 is avoided due to leap year)
10. To change Date format, @@format example @@mm/dd/yyyy
11. To replace char, use Translate@FromString@ToString example Translate@01@UD (0 will be replaced with U and 1 will be replaced with D)
12. To support and cross change the file extension, enter data in 3rd row of Master Masking file

Add the result into output file.

Step 4: Add footer of input file into output file (if any).

Step 5: Save the output file in output directory with same name along with timestamp.

End

5. RESULTS

The report presented by Bloor research, evaluated the performance of various data masking models on the bases of following parameters [2].

5.1 Sensitive data discovery:

Data discovery involves identifying and locating sensitive or regulated data in order to adequately protect it or securely remove it. Data discovery is a priority for the proposed model because it is a crucial component of compliance readiness. Data discovery involves auditing sensitive or regulated information, including confidential or proprietary data as well as protected data such as personally identifiable information (PII) or electronic protected health information (ePHI). The true goals of data discovery, therefore, are to identify and classify data in order to determining the threat, the affected resources requiring protection, and the fallout of potential data leaks more manageable. ITAO plays a very important role in the proposed model, which minimizes the data discovery process in every aspect. As data discovery algorithm checks the pattern of data into dataset from given set of rules, this process increases the complexity of the system. The proposed system has ITAO to approve the sensitivity diligence report.

5.2 Testing integration:

Testing integration check the integrity of data in test and training environment. The masked data should be meaningful for non-production environments. For same original values in specific columns same masked values are generated, which is required to maintain data integrity throughout the database.

5.3 Dynamic data masking:

For every run of proposed model on same source data with same master masking and sensitivity diligence file different masked data is generated which supports robustness.

5.4 Structured data masking:

Structured data is highly organized information that uploads neatly into a relational database, lives in fixed fields, and is easily detectable via search operations or algorithms. The proposed dynamic data masking model has a complete focus on structured data.

5.5 Unstructured data masking:

Unstructured data may have its own internal structure, but does not conform neatly into a spreadsheet or database. While disorderly in nature, it is also incredibly valuable and increasingly available in the form of complex data sources, such as web logs, multimedia content, email, customer service interactions, sales automation, and social media data. Many models involved in this analysis have this feature.

5.6 Security Integration:

The model in this work is robust in nature, it don't create any character set in memory. This leads the masked data towards security.

Given below figure 2 contains the comparative analysis study of existing model with the proposed model. Based on detail study and survey of the usage of different method used across the industries, proposed algorithm would be far superior in terms of sensitive data discovery, dynamic data masking and data security.

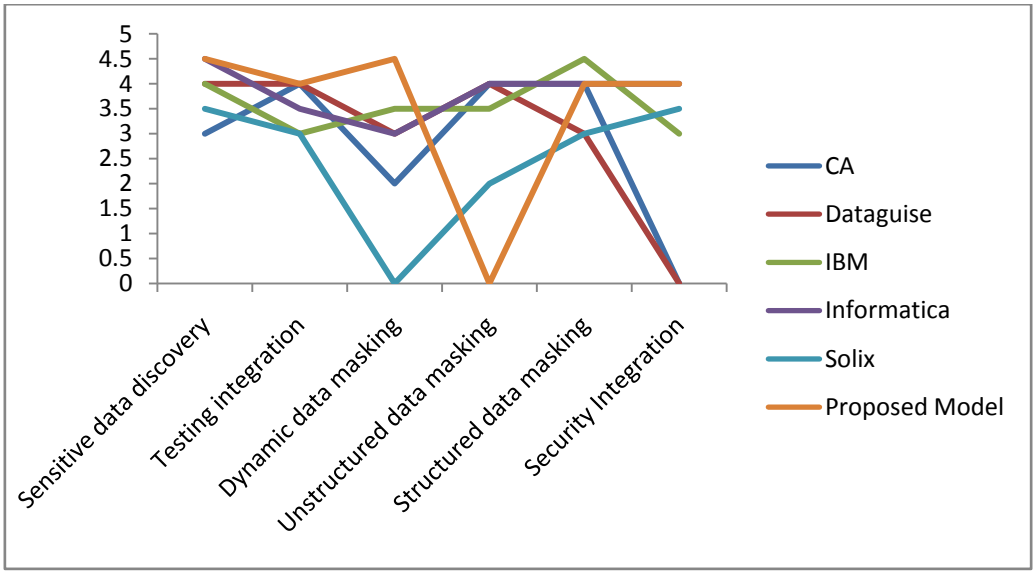

Figure2. Performance evaluation of various Data masking techniques alongside Proposed Model

6. CONCLUSION

Data masking does not exist in isolation, it is an integrated process with data discovery, data profiling, environment sizing, and other production and non-production environments. Apply non-deterministic masking to obfuscate the real data is a good beginning to secure data so that it cannot be recovered by anyone -- insider or outsider -- who gains access to the masked data. The proposed algorithm generates masked data which is completely usable in all aspects. The integrity of the data is also preserved by calculating the length of unmasked strings with masked strings. As the algorithm is non deterministic in nature, the system is not maintaining any kind of character map. Due to this reversibility is highly impossible. Embracing changes and Nurturing High Performance requires Data masking/Security analyst Leadership and Data masking analyst/Security Mindset.There are big decisions to be made including the participation in many thoughts to measure the different aspects of proposed work. The present work does not focus on time and space complexity of the proposed algorithm. The work can be scalable with additional data sets with varying size and initialization values in optimization algorithm. The proposed model can be extended for masking of unstructured data.

7. ACKNOWLEDGMENTS

Thanks to the Research Center, Abacus Institute of Computer Applications, Savitriphule Pune University for supporting me for pursuing research in the privacy preserving topic. Thanks to, Moshe Lichman-UCI center for Machine Learning Repository, for his timely response to queries regarding data sets.

8. REFERENCES

[1] Oracle leads gartner's data masking technology available on http://www.firstpost.com/business/biztech/oracle-leads-gartners-data-masking-technology-quadrant-1892231.html

[2] Report on data masking Tools available on https://www.bloorresearch.com/research/data-masking-2017/

[3] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[4] S. Moro, Machine learning repository, bank marketing dataset, available on http://archive.ics.uci.edu/ml/machine-learning-databases/00222/

[5] Asha Kiran Grandhi, Manimala Puri, S. Srinivasa Suresh, Application of PSO Optimization Technique on Medical Data to uphold data privacy, proceedings of 67th IASTEM international conference, Dubai, UAE, 1st-2nd august 2017

[6] Ravikumar G K, Dr. B. Justus Rabi, Manjunath T. N, Dr.Ravindra S. Hegadi, Archana.R.A,Design of Data Masking Architecture and Analysis of Data Masking Techniques for Testing, International Journal of

Engineering Science and Technology (IJEST) Vol. 3 No. 6 June 2011, ISSN : 0975-5462

[7] Oracle White Paper—Data Masking Best Practices JULY 2013

[8] Ruby Bhuvan Jain, Dr. Manimala Puri, Umesh Jain, An Approach To Safeguard Sensitive Data Using Shift Left Masking Model, 978-1-5386-4304-4/18/$31.00 ©2018

[9] Adrian Lane (2012). Understanding and Selecting Data Masking Solutions: Creating Secure and Useful Data, Securosis Version 1.0, August 10, 2012

[10] Ruby Bhuvan Jain, Dr. Manimala Puri, Umesh Jain, A Robust Dynamic Data Masking Transformation approach To Safeguard Sensitive Data, International Journal on Future Revolution in Computer Science & Communication Engineering, Volume: 4 Issue: 2 ISSN: 2454-4248.

[11] XiaolingXia ,Qiang Xiao and Wei Ji (2012). An Efficient Method to Implement Data Private Protection for dynamic Numerical Sensitive Attributes, The 7th International Conference on Computer Science & Education (ICCSE 2012) July 14-17, 2012. Melbourne, Australia.

[12] Waleed Ahmed, Jagan Athreya (2013). Data Masking Best Practices- white paper

[13] Kamlesh Kumar Hingwe, S. Mary SairaBhanu (2014). Sensitive Data Protection of DBaaS using OPE and FPE, 2014 Fourth International Conference of Emerging Applications of Information Technology, 978-1-4799-4272-5/14 $31.00 © 2014 IEEE DOI 10.1109/EAIT.2014.22 pg no. 320-327

[14] Muralidhar, K. and R.Sarathy,(1999). Security of Random Data Perturbation Methods, ACM Transactions on Database Systems, 24(4), 487-493.

[15] S. Vijayarani, Dr. A. Tamilarasi (2011). 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

[16] Data Masking: What You Need to Know What You Really Need To Know Before You Begin A Net 2000 Ltd. White Paper.