International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 2
ISSN: 2454-4248
366 – 370

A Robust Dynamic Data Masking Transformation approach To Safeguard Sensitive Data

Ruby Bhuvan Jain
Research Scholar, JSPM's Abacus Institute of Computer Applications, Pune, Maharashtra, India
ruby.jain81@gmail.com

Dr. Manimala Puri
Director, JSPM Group of Institutes, Survey No.80, Pune-Mumbai Bypass Highway, Tathawade, Pune, Maharashtra, India
manimalap@yahoo.com

Umesh Jain
Vice President, Deutsche Bank, Pune, Maharashtra, India
utjain@gmail.com

***Abstract*—** Large amount of digital data is generated rapidly all around the corners. Providing security to digital data is the crucial issue in almost all types of organizations. According to the Identity Theft Resource Center, there were 8,069 data breaches between January 2005 and November 2017, and in recent years the number of data breaches and compromised records has skyrocketed [1]. To provide protection to the digital sensitive data, from data breaches in the need of hour. Almost all domains like insurance, banking, health care, and educational and many more are concern about security of sensitive data. Data masking is one of the vital discussions everywhere as data breach leads to threats. 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 random deterministic masking algorithm with shift left approach. This paper describes methodology & experimental design and results.

Keywords- *Shift Left approach; Robustness; Data breach; Data de-identification; Sensitivity Diligence.*

*****

I. INTRODUCTION

In today's technology era, most of the organizations have digital data. Sensitive data is a part of every large organization's normal trade. Internal and external both attacks on the database, are on the increase. More than 70 percent of all attacks on databases are internal, making them very complicated to notice and control. Allowing copying sensitive data from production environment and used for non-production environments increases the likelihood for stealing, loss or exposure raises the organization's risk. Organizations have to share sensitive data with non-production environments like Development, User Acceptance Testing, Operational Acceptance Testing, Training Environment and Research as it is required for the purpose of product/tool upgrade. Instead of sharing production data AS IS with other environments it is always recommended to share masked data which is like preventive method instead of cure due to safety breaches.

Data masking is the process to defensibly protect sensitive information like PII (Personally Identifiable Information), PHI (Protected Health Information), SPI (Sensitive Personal Information) and PSI (Price Sensitive Information) by replacing the original sensitive values with new, legible, meaningful values, which retains all the properties of original data. Regulations and Compliances, enforces the organizations to protect sensitive data against data breaches. Data masking helps organizations meet compliance requirements for Payment Card Industry Data Security Standard (PCI DSS), Health Insurance Portability and Accountability Act (HIPAA), Gramm-Leach-Bliley Act (GLBA) and other data privacy regulations.

We have digital disrupters everywhere. The only way an industry can survive in today's global and highly digitized and virtually connected marketplace is by embracing technology and business change waves that act as tides that determine the direction of sail.

A lack of processes and technology to protect data in non-production environments can leave the company open to data theft or exposure and regulatory non-compliance. Data masking is one of the effective ways to reduce enterprise risk. Data masking is also known as de-identification, data securing, depersonalization, desensitization, obfuscation and data scrubbing.

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.

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

Authors in [12], proposed a PSO optimization technique which included clustering, encryption and distribution privacy techniques, to cluster medical data to check accuracy of resultant data.

Masking is not a specific process that designated masking team can follow. Masking is a philosophy and new way of thinking about securing sensitive data in such a way that accessible and usable data is still available and usable in non-production environment. To be truly "Masking", development team, ITAO and customers need to understand and actively participate in putting Masking framework and associated values and principles into practice. Shift Left emphasis that production data should be masked at the production itself in agreed frequently. The aim of this approach is that non production environment will always get masked data.

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.

 ISSN: 2454-4248
366 – 370

The remaining paper is organized as follows: Section II describes methodology & experimental design and section III describes results, followed by conclusion.

II. MASKING METHODOLOGY

Step1: Data Collection

This research work considers real open source data. UCI states that, Machine learning community donates data sets to UCI, for promoting research & development activities. The structure promoted by UCI to collect the data sets is suitable for machine learning applications. Till date, UCI has collected 425 data sets, belonging to various domains. Being data will be kept open for other to view, hence Donors have their own policy for sharing data and Donors decides the type of data to be shared and granularity of the data [2]. The dataset used in this research available with the name bank marketing under business category in UCI repository.

Step2: Data Description

The measurements of the data sets are shown in table I.

TABLE I. DATASET SUMMARY MEASUREMENTS

S.N.	Measurements	Value
1	Data Set Characteristics	Multivariate
2	Area	Business
3	Attribute Characteristics	Real, Numeric, Categorical
4	Associated tasks	Classification
5	Missing values	N/A
6	Date Donated	2012-02-14

The data set shown in table I do not have name of the client name and Account Number as it can reveal identity of the bank client. The researchers advice readers to see structure of the data sets on UCI repository site for further deep analysis.

Step3: Data Specification

Table II defines attribute name with the description of value, the category in which that can will fall and type of sensitivity in the data mentioned in Section I. The current work uses 12 attributes out of 20 attributes from original dataset. Table II contains attribute specification of the data set.

TABLE II: ATTRIBUTE SPECIFICATIONS

S.N.	Attribute Name	Attribute Description	Type of Attribute	Sensitivity Class
1	Age	Age	Numeric	PII
2	Job	Type of job	Categorical	PII
3	Marital	Marital status	Categorical	SPI
4	Education	Education	Categorical	SPI
5	Default	Has credit in default?	Categorical	PSI
6	Housing	Has housing loan?	Categorical	PSI
7	Loan	Has personal loan?	Categorical	PSI
8	Contact	Contact communication type	Categorical	PII
9	Month	Last contact month of year	Categorical	PII
10	day_of_week	Last contact day of the week	Categorical	PII
11	Duration	Last contact duration, in seconds	Numeric	-
12	Campaign	Number of contacts performed during this campaign and for this client	Numeric	PII

Step 4: *Experimental Design*

The following figure shows the technical architecture of proposed solution, followed by description.

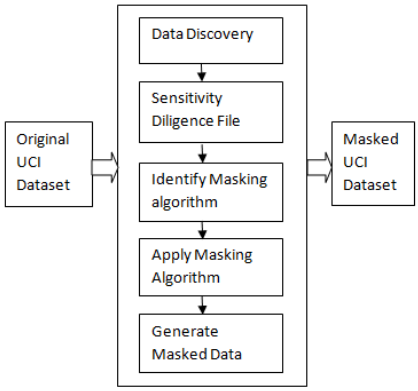

Figure1. Technical Architecture

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 layers of implementation. Left layer contains source dataset. The right layer is the masked dataset of left layer source dataset. The layer in the centre shows the process of data discovery, role of sensitivity diligence file and identify and apply masking algorithm.

A. Data Discovery: Data discovery need single input Source data files from the actor. Based on discovery algorithm, potential sensitive fields are identified, this file is reviewed by ITAO and outcome is final 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.

In this step, system will read all input files. To store masked file, output directory is required. If the directory is not available, directory will be created. Data Discovery is explained below:

Start

1. Read Master Masking file.
2. Read Input Data file.

 ISSN: 2454-4248
366 – 370

3. Read Sensitivity Diligence file.
4. Read Output directory
   4.1 If available do nothing
   4.2 Else create output directory

End

B. Sensitivity Diligence File: This input file have 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 *must be* 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. To store this input a separate file called Sensitivity Diligence file is required. This file will have all column names in first row and second row contains description of data whether it is sensitive or not. 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.

	A	B	C	D	E	F	G	H	I	J	K	L
1	age	job	marital	education	default	housing	loan	contact	month	day_of_w	duration	campaign
2	N	Y	Y	Y	N	Y	Y	Y	N	Y	Y	N

Figure 2: Sample Sensitivity diligence

This sample data given in figure 2 represents only two rows, this means only business as usual masking algorithm is applied.

Sensitivity diligence algorithm is described below:

Start

1. Read file F
2. File F has set of rows i.e. f = (r1, r2, r3, …….. , rn)
3. Each row r consists of multiple columns i.e. r = (c1, c2, c3 , ……. , cn)
4. Read r1 having name of columns of source file
5. ***Read r2***
   For each column in r2
   Identify the sensitive columns i.e. Scs
   5.1 If 'Y' Goto Step 6
   5.2 Else ***End***
6. ***Read r3***
   If condition
   Identify appropriate masking algorithm
   Else
   Business as usual

End

Approval of Sensitivity Diligence Report: This sensitivity diligence file generated 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 [13]. Once approved appropriate masking algorithm can be identified and applied.

C. Identify Masking algorithm: This step needs input from third row of sensitivity diligence file algorithm can be identified. Special masking conditions are mentioned here. 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.

D. Apply Masking Algorithm: This step reads Master masking file which has list of input files- source data file and sensitivity diligence file as input. The algorithm creates output directory if it not exists. 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 detailed algorithm is described below:

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. Else if lowercase replace with any random lowercase character.
3. Else if number replace with any random number. If new masked number 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)

 ISSN: 2454-4248
366 – 370

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

E. Generate Masked Data: This step observes the masked data file for data usability, reversibility, robustness and lossless of data.

III. RESULT

The proposed algorithm need input data file along with master masking file. The master masking file template is design in such a way that it stitches input file and Sensitivity diligence file. Proposed algorithm is applied on the complete dataset along with all attributes of bank marketing. Given below in figure 3 contains only 15 records from the source data file. Figure 4 contains the masked data after first execution of the algorithm and figure 5 contains the masked sample data after second execution. This algorithm was applied 5 different times with same dataset and with multiple datasets but the result was consistently satisfying the following features:

- Robust: Results are showing that every run for the same source data with same master masking and sensitivity diligence file different masked data is generated which supports robustness. Figure 4 and Figure 5 have different masked values for same input files.
- Integrity: For same original values in specific columns same masked values are generated, which is required to maintain data integrity throughout the database.
- Accuracy: The algorithm replaces small case with small case, upper with upper, number with number and no change in special characters like '@' in case of mail id, '/' , space bar and '–' in address line. The algorithm also takes care of the length of the string, means unmasked data and masked data will have same number of characters.
- Reversibility: No character set is generated as the algorithm used here is non deterministic in nature. Due to this feature gaining the unmasked original data is next to impossible.

	A	B	C	D	E	F	G	H	I	J	K	L
1	age	job	marital	education	default	housing	loan	contact	month	day_of_w	duration	campaign
2	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1
3	57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1
4	37	services	married	high.school	no	yes	no	telephone	may	mon	226	1
5	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1
6	56	services	married	high.school	no	no	yes	telephone	may	mon	307	1
7	45	services	married	basic.9y	unknown	no	no	telephone	may	mon	198	1
8	59	admin.	married	professional.course	no	no	no	telephone	may	mon	139	1
9	41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	217	1
10	24	technician	single	professional.course	no	yes	no	telephone	may	mon	380	1
11	25	services	single	high.school	no	yes	no	telephone	may	mon	50	1
12	41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	55	1
13	25	services	single	high.school	no	yes	no	telephone	may	mon	222	1
14	29	blue-collar	single	high.school	no	no	yes	telephone	may	mon	137	1
15	57	housemaid	divorced	basic.4y	no	yes	no	telephone	may	mon	293	1

Figure 3: Source Data file

Figure 3 contains the first 15 data records of source data file.

	A	B	C	D	E	F	G	H	I	J	K	L
1	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign
2	56	tngpgajrc	xbzwdij	bybrr.5l	no	kf	kf	yxryoiexe	may	icw	119	1
3	57	popkrpue	xbzwdij	vquy.csbv	unknown	kf	kf	yxryoiexe	may	icw	777	1
4	37	popkrpue	xbzwdij	vquy.csbv	no	kor	kf	yxryoiexe	may	icw	610	1
5	40	slczg.	xbzwdij	wnjtj.4t	no	kf	kf	yxryoiexe	may	icw	881	1
6	56	popkrpue	xbzwdij	vquy.csbv	no	kf	kor	yxryoiexe	may	icw	334	1
7	45	popkrpue	xbzwdij	qhsev.9s	unknown	kf	kf	yxryoiexe	may	icw	280	1
8	59	slczg.	xbzwdij	sudnvvqtt	no	kf	kf	yxryoiexe	may	icw	852	1
9	41	cpon-jsesio	xbzwdij	amjvkcr	unknown	kf	kf	yxryoiexe	may	icw	888	1
10	24	vdgascgvtf	keqzbw	sudnvvqtt	no	kor	kf	yxryoiexe	may	icw	562	1
11	25	popkrpue	keqzbw	vquy.csbv	no	kor	kf	yxryoiexe	may	icw	43	1
12	41	cpon-jsesio	xbzwdij	amjvkcr	unknown	kf	kf	yxryoiexe	may	icw	54	1
13	25	popkrpue	keqzbw	vquy.csbv	no	kor	kf	yxryoiexe	may	icw	347	1
14	29	cpon-jsesio	keqzbw	vquy.csbv	no	kf	kor	yxryoiexe	may	icw	100	1
15	57	tngpgajrc	ymfztrqh	bybrr.5l	no	kor	kf	yxryoiexe	may	icw	267	1

Figure 4: First Run

	A	B	C	D	E	F	G	H	I	J	K	L
1	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign
2	56	ndztphgyy	sskcyxz	xsnjq.0s	no	dd	dd	xfangmjki	may	kge	876	1
3	57	omszpjko	sskcyxz	wcxr.vengap	unknown	dd	dd	xfangmjki	may	kge	129	1
4	37	omszpjko	sskcyxz	wcxr.vengap	no	vop	dd	xfangmjki	may	kge	186	1
5	40	pchsq.	sskcyxz	rsiky.7p	no	dd	dd	xfangmjki	may	kge	626	1
6	56	omszpjko	sskcyxz	wcxr.vengap	no	dd	vop	xfangmjki	may	kge	179	1
7	45	omszpjko	sskcyxz	xwwee.4l	unknown	dd	dd	xfangmjki	may	kge	140	1
8	59	pchsq.	sskcyxz	hthmrclvjrwz.vxuwdb	no	dd	dd	xfangmjki	may	kge	879	1
9	41	olqh-jzalcv	sskcyxz	kkhuqjr	unknown	dd	dd	xfangmjki	may	kge	334	1
10	24	tnlomhioxw	ciasns	hthmrclvjrwz.vxuwdb	no	vop	dd	xfangmjki	may	kge	179	1
11	25	omszpjko	ciasns	wcxr.vengap	no	vop	dd	xfangmjki	may	kge	92	1
12	41	olqh-jzalcv	sskcyxz	kkhuqjr	unknown	dd	dd	xfangmjki	may	kge	17	1
13	25	omszpjko	ciasns	wcxr.vengap	no	vop	dd	xfangmjki	may	kge	493	1
14	29	olqh-jzalcv	ciasns	wcxr.vengap	no	dd	vop	xfangmjki	may	kge	903	1
15	57	ndztphgyy	upidjalm	xsnjq.0s	no	vop	dd	xfangmjki	may	kge	298	1

Figure 5: Second Run

Table III is the representation of comparative study of proposed solution with some existing solutions. First column of the table contains list of the masking features. As the solutions are not open source, the comparison is based on the only features which are available free of cost.

TABLE III: COMPARISION OF MASKING FEATURES

Masking Capabilities	**Proposed Solution**	Informatica ILM	IBM Optim	Oracle OEM
Shift-left	Implicit	Explicit	Explicit	Explicit
Continuous feedback	Implicit	Explicit	Explicit	Explicit
Dynamic data masking	Implicit	Implicit	Implicit	Explicit
Scalability	Very High	High	High	High
Robustness	Very High	NA	NA	NA
Integrity	Very High	High	High	High
Application Usability	High	High	High	High
Application Cost	Free	High	High	High

 ISSN: 2454-4248
366 – 370

IV. CONCLUSION

A job well begun is half done. Apply 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 random deterministic in nature, the system is not maintaining any kind of character map. Due to this reversibility is highly impossible. 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 extended with additional data sets with varying size and initialization values in optimization algorithm. The proposed algorithm preserves proper balance between the protection and knowledge discovery.

ACKNOWLEDGMENT

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.

REFERENCES

[1] https://www.opswat.com/blog/11-largest-data-breaches-all-time-updated
[2] S. Moro, R. Laureano and P. Cortez.Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS.
[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, http://archive.ics.uci.edu/ml/machine-learning-databases/00222/
[5] 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
[6] Muralidhar, K. and R.Sarathy,(1999). Security of Random Data Perturbation Methods, ACM Transactions on Database Systems, 24(4), 487-493.
[7] Ravikumar G K, Manjunath T N, Ravindra S Hegadi, Umesh I M (2011). A Survey on Recent Trends, Process and Development in Data Masking for Testing-fIJCSI- Vol. 8, Issue 2, March 2011-p-535-544.
[8] 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
[9] Adrian Lane (2012). Understanding and Selecting Data Masking Solutions: Creating Secure and Useful Data, Securosis Version 1.0, August 10, 2012
[10] Waleed Ahmed, JaganAthreya (2013). Data Masking Best Practices- white paper
[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] AshaKiran, ManimalaPuri, Srinivasa Suresh, PSO Enabled Privacy preservation, Indian Journal of Science and Technology, Vol 10(11), DOI: 10.17485/ijst/2017/v10i11/89318, March 2017, ISSN:0974-5645(online)
[13] 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