IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 5, PP 01-06

Comparative Study on Approaches of Data Masking

Mridul Chavan[1], Ketki Joshi[2], Vidhya Chaudhary[3],
Ilu Mandaliya[4], Supriya Mandhare[5]

[1,2,3,4](Student, BE-Information Technology, Atharva College of Engineering, Mumbai University, India)
[5](Assistant professor, Information Technology, Atharva College of Engineering, Mumbai University, India)

Abstract: *In today's scenario, data is the most valuable commodity for every organization and securing organizational data is a top concern of the IT industry. The top data security issues are external attacks, lack of accountability, vulnerabilities in the system, data breaches. To tackle them, algorithms and mechanisms are set up for providing security from threats outside the organization, but data is left vulnerable to insider attacks. Data within an organization is accessed based on the privileges and access levels but this does not ensure complete data security from the breaches and passive attacks. Data is constantly moving from one environment to the other, that is, from production environment to non- production environment for testing, data interpretation and analysis, data warehousing, mining and other research purposes. This paper proposes the idea of masking the non-production data which contains personally identifiable information (PII) and at the same time maintaining compliance with various government policies and regulations. The aim is not just simple masking of the data but obfuscating the real data with a pattern that appears realistically similar to the real data. This paper illustrates the static and dynamic masking approaches to data.*

Keywords *-data masking, dynamic data masking, insider attacks, organizational data security, static data masking*

I. Introduction

As the volume of data grows across industries and the number of data attacks on enterprises continue to increase, organizations large and small are seeking best practices on how to protect their data. Security professionals and managers are increasingly concerned that the leading information security risk to organization comes from within. After evaluating all threats to an organization, surveys conclude that even though most attacks come from outside the organization, the most serious damage is done with help from inside[9]. Hence, there is a need to deal with exposure of sensitive organizational data at the hands of insider threats. The approach in this paper is to consider the aspect of data security which deals with securing non production databy, preventing the exposure of sensitive data to developers, testers, and via outsourcing by employing advanced data masking techniques to minimize the probability of such internal attacks affecting an organization/business etc[10]. Data maskingprimarily tackles the issue of data protection[5].

Data Masking is the technique of obfuscating sensitive data to prevent exposure of this data to users who don't have the authority to view the data. It is performed as per the access privileges of the user. In data masking, the aim is to mask sensitive information in non-production data with realistic looking but not real information. Data masking techniques ensure that data security is maintained by obscuring specific data within a database table thereby reducing the risk of data exposure and data breachesfrom both inside and outside an organization[8]. Effective data masking requires data to be altered in such a way that actual values are re-engineered, while retaining the functional and structural meaning of the data, so that it can be used in a meaningful way without compromising on security. The issues in data security have been highlighted by S. Selvakumar et al [5]. The intent of the published work was to integrate security by means of data masking in a multi-tenant cloud environment with the help of virtual machine masking and platform masking. The mechanism proposed increases the reliability in database service environments. Min Li et al have summarized the advancements in data masking and a generic model has been proposed from a theoretical perspective along with the shortcomings of the model[3]. G.Sarada et al have put forward four new approaches for masking the data using min-max normalization, fuzzy logic, and rail-fence and map range and also expounded on the limitations of the traditional methods[2]. The elementary idea of the causes and requirement for the implementation of this approach has been elaborated in the abstract and the introductory section I along with the work suggested and implemented by other authors and a brief on their approach. Section II expounds on the idea of data masking and the extensive process of masking sensitive data. The main objective is to mask the sensitive data in such a manner that the masked data appears more realistic along with facilitating analysis on the data. Section III describes the preeminent approaches in masking of data - Static Data Masking & Dynamic Data Masking. Section IV illustrates the various techniques which can be employed to mask the data followed by the conclusion in section V.

II. Data Masking

Organizations share the data from production environment for various business needs[11]. Som enterprises do not do much to protect their data in non-production environments. Hence, the data masking techniques are employed to safeguard sensitive data in non-production environments.

Data Masking is an approach in which the sensitive production data which is obtained from live applications, is obfuscated into realistic looking fake data for non-production activities such as testing, quality assurance, development etc. The general process of how data masking takes place is given in Fig. 1 as follows:

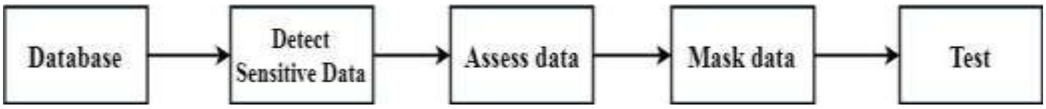


Fig. 1: The process of data masking

A comprehensive 4-step approachfor implementing data masking consists of the following steps[1,4]:

1. Detect sensitive data

To initiate the process of data masking, the data that needs to be masked must be identified. The decision on what constitutes as sensitive data is made by taking into consideration various government regulations and policies that dictate how sensitive data can be used or shared. This phase identifies sensitive or regulated data across the entire organization. The purpose is to come up with the list of sensitive data elements specific to the organization and discover the associated tables, columns and relationships across databases that contain the sensitive data. This is carried out usually by data, security and business analysts. Upon completion of this step, the next phase is the assessment of data.

2. Assess data

This step oversees the location of sensitive data in the organization schema/database. The DBA can then designate an attribute/column as sensitive for inclusion in the masking process or not sensitive for exclusion from future ad hoc pattern searches. Hence, identification of the masking algorithms to replace the original sensitive data. Developers or DBAs work with business or security analysts with their own masking routines.

3. Mask data

Once the detection and analysis of sensitive data in non-production environments is performed, the DBA can execute the masking algorithms decided in the previous phase to replace/mask all the sensitive data. This is the iterative phase.

4. Test

The final step of the masking process is to test whether the mask has been correctly defined and created. Once the masking process has completed and has been verified, the DBA then hands over the environment to the application testers. If the masking algorithms need further changes, then the DBA restores the database to the pre-masked state, makes the necessary adjustments to the masking algorithms and re-executes the masking process.

III. Approaches To Data Masking

Data Masking has two basic approaches namely static data masking and dynamic data masking[7]. The fundamental difference between the two is that, in static data masking the sensitive data is permanently masked by altering data at rest whereas in dynamic data masking, the sensitive data is masked in transit which leaves the original data intact and unaltered.

1.1 Static Data Masking

Static data masking mask the sensitive data in the production databaseby the use of pre-decided masking techniques[10]. It provides a basic level of data protection as it creates an offline version of the live production database. Generally organizations and enterprises employ static data masking when they want to contract out their data to third party or developers etc. Hence, realistic looking data can be used for testing, quality assurance and development without disclosing sensitive information.

Additional applications encompasses safeguarding sensitive data for benefit in analytics and training compliance with standards and regulations (GDPR, PCI, HIPAA) that impose limits on how organizations make use of data, especially PII[11].

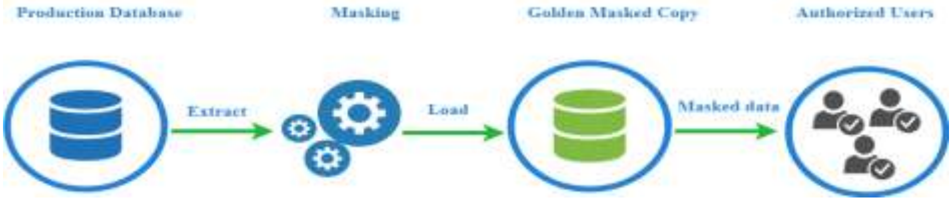


Fig. 2: Conceptual diagram of Static Data Masking

As shown in Fig. 2, in static data masking, the live production database is duplicated and an offline copy (Golden Masked Copy) constructed with all the sensitive fields masked. Hence, inherently there are two databases which are ordinarily not synchronized. The golden masked copy commonly lags behind the live production database, but it is updated on a periodic basis. The impediment of static data masking is that the actual live production database is left unprotected, so personnel who do have access to it can view the actualdata records and not masked records. Finally, the overhead incurred by having two copies is significant as it includes the cost of maintenance and hardware

.

1.2 Dynamic Data Masking

In dynamic data masking,the sensitive data is masked in transit which leaves the original data intact and unaltered[7]. Data is obfuscated as it is accessedin real time and the sensitive data never leaves the live production database. Dynamic data masking is also used as tool to enforce role-based security in applications.

3.2.1 Request Based Dynamic Data Masking

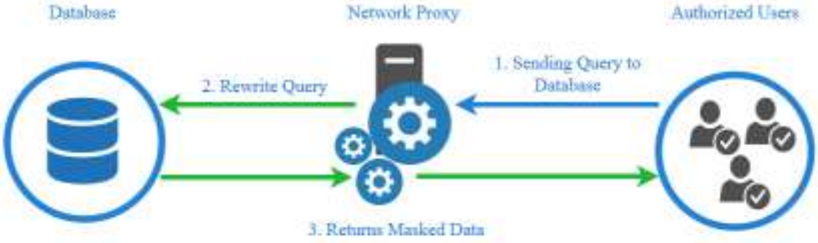


Fig. 3: Conceptual diagram of Request Based Dynamic Data Masking

As shown in Fig. 3, in request based dynamic data masking, the query sent by the user is reconstructed with the masking actions before it is sent to the database in real time. The database then receives the query with the masking applications to be performed.

3.2.2 Response Based Dynamic Data Masking

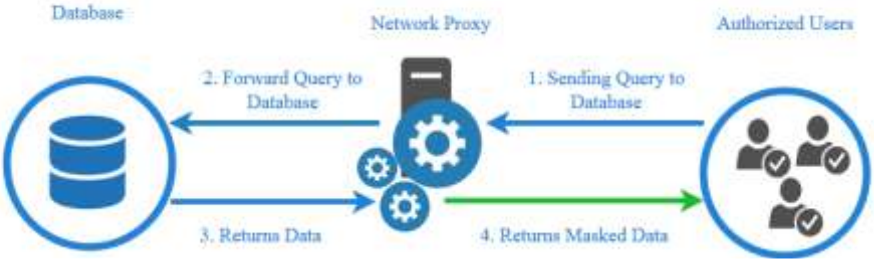


Fig. 4: Conceptual diagram of Response Based Dynamic Data Masking

As shown in Fig. 4, in response based dynamic data masking, the query is sent to the database in real time and data is masked in real time as it is received from the users.

A considerable advantage of dynamic data masking is that there are no copies of the production data and the sensitive data is masked from the live database itself. Since,the activities are performed on real data,time is substantially saved.

Most enterprises should employ both approaches of data masking, static data masking and dynamic data masking to ensure data protection. Even with the static data masking in place, almost any organization with sensitive data in live database should make use of dynamic data masking to protect live production systems. Static data masking is used for outsourcing and dynamic data masking to protect in premises live databases.

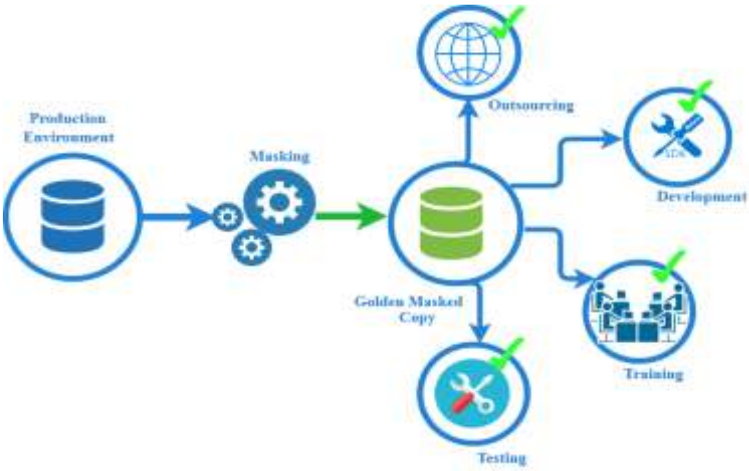


Fig. 4: Applications of Data Masking

IV. Data Masking Techniques

Depending upon how sensitive the data is and the requirements of masking, data masking techniques are implemented. Traditionally, several techniquesused for masking include substitution, shuffling, encryption, masking out, etc[1]. The limitation in using these techniques is that it fails in successful generation of random values that are unique for every original value[2]. For reducing these limitations, techniques as follows can be implemented and applied for unique masking.

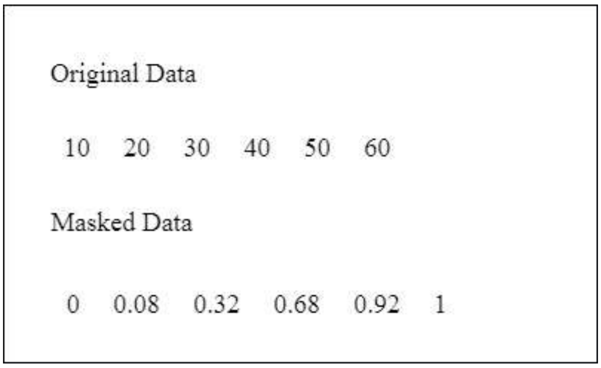
4.1. Fuzzy Based Approach:

In this approach, the concept of fuzzy set theory is used, that generates a fuzzy logic output that can be used as a masked result. This approach is more likely to maintain the interrelation in the data and protect privacy. A fuzzy membership function is used to map the data into a masking result, thus reducing the time to process. Using this approach, the data can only be masked within a range of 0-1. An example using S-shaped fuzzy functionis given below[6]:

$$f(x;a,b)=\begin{cases} 0\,, & x\le a \\ 2\left(\frac{\mathrm{x-a}}{\mathrm{b-a}}\right)^{\mathbf{2}}, & a\le x\le \frac{a+b}{2} \\ 1-2\left(\frac{x-b}{b-a}\right)^{2}, & \frac{a+b}{2}\le x\le b \\ 1, & x\ge b \end{cases} \qquad \text{……………………………..} \ (1)$$

Table 1: Specifications of the equation 1

No.	Variables	Specifications
1.	x	Value of sensitive attribute
2.	a	Minimum value in sensitive attributes
3.	b	Maximum value in sensitive attributes


Original Data

10 20 30 40 50 60

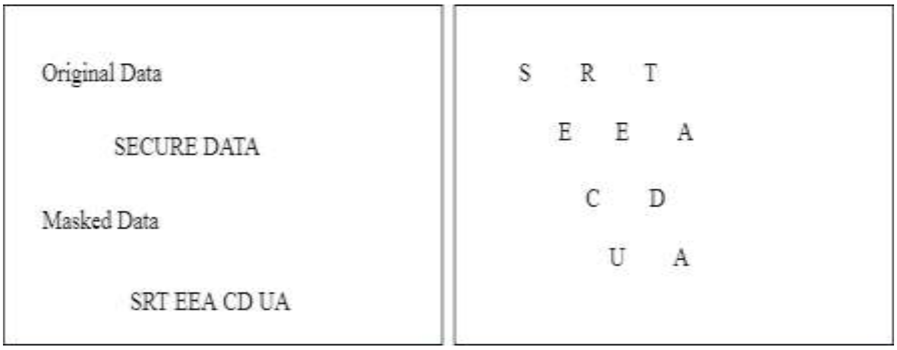
Masked Data

0 0.08 0.32 0.68 0.92 1

Fig. 5: Example of Fuzzy Based Approach

4.2. Rail-fence Method:

This technique is mostly applied to categorical data wherein the original data is written row/column-wise and the transformed data is fetched by traversing along columns/rowsrespectively[2]. An example of masking using this approach is given below:



Original Data

SECURE DATA

Masked Data

SRT EEA CD UA

S R T

E E A

C D

U A

Fig. 6: Example of Rail-Fence Method

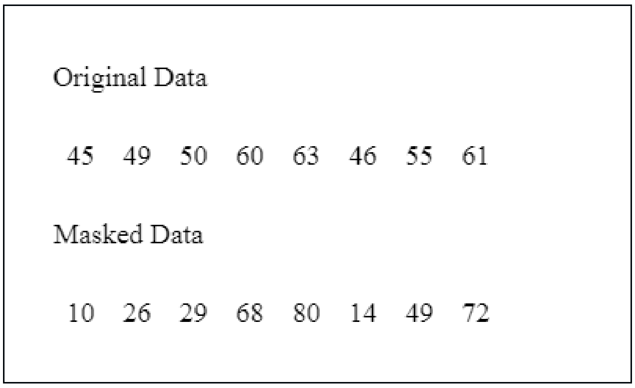
4.3. Map Range (Rosetta Code):

This method is of use when original data is needed to be mapped to a specified range. This method is mostly used for mapping large numbers to small numbers. The formulais given as follows[2]:

$$t = b1 + \frac{(s-a1)(b2-b1)}{(a2-a1)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \ (2)$$

Table 2: Specifications of the equation (2)

	Variables	Specifications
1.	s	value of sensitive attribute
2.	a1	minimum value in sensitive attributes
3.	a2	maximum value in sensitive attributes
4.	t	mapped value of sensitive attribute
5.	b1	minimum value in mapped range
6.	b2	maximum value in mapped range



Original Data

45 49 50 60 63 46 55 61

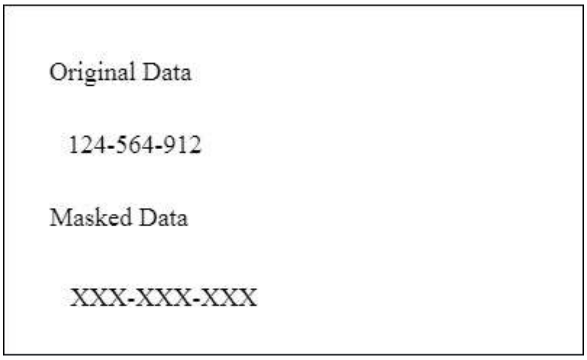
Masked Data

10 26 29 68 80 14 49 72

Fig. 6: Example of Map Range Method

4.4. Masking Outs:

This is a techniquethat simply masks some part of the original data with a specific character[2]. This can be used when masking the data would not affect the regular processing, otherwise it is of no use when the original data holds required information. An example is given below:



Original Data

124-564-912

Masked Data

XXX-XXX-XXX

Fig. 7: Example of Masking Out

V. Conclusion

A comparative study between the two fundamental approaches of data masking, namely static data masking and dynamic data masking was performed and reviewed in this paper. The primary difference between the two is that in static data masking, there is an offline copy of the live production database, called the test database which has all of the sensitive data masked whereas in dynamic data masking, the live database is intact and unaltered and the masking is performed as and when the sensitive data is accessed i.e. Data is masked in transit. Enterprises need to employ both of these approaches to safeguard their data. Static data masking is preferable in cases where data is outsourced – testing, development, data analysis etc. Dynamic data masking is employed to protect the organizational data on premises and is used to fulfill the various government policies (GDPR, PCI, HIPAA) pertaining to the usage of sensitive data. The new masking techniques proposed- fuzzy based approach, rail fence method, etc. will prove advantageous over traditional methods of substitution, shuffling as they create data which is realistic looking but fake as opposed to just obfuscating the data which makes it pretty evident that the data is masked. These new techniques would also allow for robust non production activities to be carried out onmasked data thereby safeguarding sensitive information from insider attacks.

References

[1] Data Masking Best Practice, Oracle White Paper, June 2013.

[2] G Sarada, G Manikandan, Dr.N. Sairam,"A Few New Approaches to Data Masking", *International Conference on Circuit , Power and Computing Technology* 2015.

[3] Min Li, Zheli Liu, ChunfuJia ,Zongqing Dong, "Data Masking Generic Model",Fourth International Conference on Emerging Intelligent Data and Web Technologies 2013.

[4] Osama Ali and AbdelkaderOuda, "A Classification Module in Data Masking Framework for Business Intelligence Platform in Healthcare", IEEE, 2016.

[5] S.Selvakumar and M. Mohanapriya, "Securing Cloud Data in Transit using Data Masking Technique in Cloud Enabled Multi-Tenant Software Service", *Indian Journal of Science and Technology*, vol. 9, no. 20, 2016.

[6] Timothy J. Ross "Fuzzy Logic with Engineering Applications", McGraw Hill International Editions, 1997.

[7] https://www.gartner.com/doc/3153926/static-dynamic-data-masking-explained

[8] http://searchsecurity.techtarget.com/definition/data-breach

[9] https://www.isdecisions.com/insider-threat/statistics.htm

[10] https://www.red-gate.com/hub/product-learning/sql-clone/protecting-production-data-non-production-environments

[11] https://compliancy-group.com/gdpr-compliance-hipaa-software