A Study on Dynamic Data Masking with its Trends and Implications

Ravi Kumar G.K
Research Scholar
Dr.MGR University
Chennai, Tamil Nadu, India

Dr B Justus Rabi
Professor, Dept of E&E
Dr.MGR University
Chennai, Tamil Nadu, India

Manjunath TN
Research Scholar
Bharathiar University
Coimbatore, Tamil Nadu, India

ABSTRACT
In the present information age, Conventional data masking solutions perform Static Data Masking, where the obfuscated values are physically stored in the database. For obvious reasons, Data Masking has been limited to Development and QA environments. Now, for the first time, organizations/enterprises production environments can be dynamically masked without enhancing existing applications and building new security layers. By enhancing the existing application security with Dynamic Data Masking security layer, which enables the application of various security actions, includes masking, hiding or blocking incoming requests. Dynamic Data masking is an effective strategy in reducing the risk of data exposure to insiders and outsiders in organizations and is a best practice for securing production databases. Dynamic data masking is the process of masking specific data elements, Ascend without touching applications or physical production databases. This paper will help data security developers and analysts in securing production data effectively.

Keywords: Data masking, ETL, Mapping, Replacement technique.

1. INTRODUCTION
Data masking has become one of the important activity now a day, because of onsite offshore model of working, if we need to get data from onsite to offshore accessing via network, we need to ensure 100% security of sensitive data. The challenge was that it didn't want to display sensitive information in the existing key systems included in the database during testing, development and support phases. For example, customer's address, phone number, finance details, credit details etc. were sensitive information. But the rules for masking data were very complex to all levels. The customer required an innovative and centralized application which could handle complex requirements for different systems on masking the data in middle layer.Authors analysed and derived the solution framework for the data masking activity in an innovative way. It developed in two modules 1.GUI Based which performs as a Masking using random variable 2. Proxy application.

The XML based GUI developed in ASP maintained and executed the different masking rules for all the systems accessed through the proxy server. We used WebScrab web server for developing the server script. It is an open source but was customized in a larger level to suit client's requirements. The data masking at proxy level works at all levels including individual page, application level, and globally for all applications accessing through Proxy. Advantages of the dynamic data masking are:

i. Cost Savings (ii). Enhanced Productivity/Efficiency (iii).Improved Services and Operations

2. LITERATURE REVIEW
Authors has undergone literature review phase and evolved with the problem statement with the help of work, has published till today in the area of data masking

Muralidhar, K. and R. Sarathy, and R. Parsa (1996) – describes how to maintaining the Relationship between confidential and Non-Confidential Attributes in Statistical Databases for data masking.

Parsa, R.A., K. Muralidhar, and R. Sarathy (1997) - Discuss the general method for Data Perturbation.

Muralidhar, K. and R. Sarathy (2002) - Presented "The Two Step Data Shuffle: A New Masking Procedure," in Invited seminar presented to the Census Bureau and the Washington Statistical Society.

Muralidhar, K. and R. Sarathy (2009) - Proposed the idea of Privacy Violations in Accountability Data Released to the Public by State Educational Agencies.

Ravi kumar GK, Dr. Justus Rabi, Manjunath TN (2011) - Proposed a uniform architecture for data masking using random replacement.

Authors are exploring the importance of dynamic data masking with masking and subsetting results in effort and cost reduction with improved data security factors.

3. METHODOLOGY
Dynamic Data Masking solution transparently intercepts incoming SQL requests and applies security rules in real-time on them. One of the simplest illustrations:

a.) an application sends in 'select name,..'
b.)which dynamically gets re-written to 'select substr (name,1,2)
c.) which will take the name 'Obama' and display it as "Ob***"
d) The following expression is used to mask credit card numbers in productions: nvl2(CREDIT_CARD_NO,('****-****-****-'||substr(CREDIT_CARD_NO,length(trim(CREDIT_CARD_NO))-3)),")

Some of the frequently used masking, scrambling and blocking algorithms include, data substitution, replacing a value in the column with fictionalized data Truncating,

scrambling, hiding or nullifying, which replaces column values with NULL or '****"

Randomization, replacing the value with random data Skewing, which alters the numeric data by a random variance Using masking or scrambling functions with fictitious values Character substring masking, replacing a particular substring with a custom mask User extensibility is available with custom PL./SQL functions and/or a Java API Active Base.

3.1 Incremental Implementation Approach

The implementation of dynamic data masking solution is done using a unique methodology that accelerates implementation time from months into weeks. The methodology includes the following steps:

Step 1: Test Scenarios

The best way to get a full scope of the project is to start by defining test scenarios that will also be used for final acceptance testing. These scenarios should document:

1. The on-line application screens, packaged reports and batch processes impacted by the masking

2. The actual data to be masked, as presented in the application screens and packaged reports

3. The fields masking requirements

- Masking functions
- Business rules and constraints on masked data to be respected
- Expected behavior of queries on masked fields
- Expected behavior of update on masked fields

4. Map processes that should not be impacted by the masking (e.g. ETL…), by passing Active Base Security TM In many cases, such test scenarios already exist for application testing purposes.

Step 2: Mapping Data

Once data to be masked has been clearly identified at the application level, the first effort will be to map this data to actual technical objects such as tables, view, stored procedures This mapping can be achieved through multiple sources, which will most probably be used jointly:

- Application expertise, Database knowledge and naming standards.
- Active Base Security in logging mode can be used while running a specific scenario in order to see the complete interaction with the database.
- Database integrity constraints and object dependencies. Mapping data is the most time-consuming part of the project

Step 3: Developing Masking Rules

Once a masked field has been mapped to the database, rules are easily created in the Dynamic Data Masking solution in order to implement the masking. Most of the work required to handle masking of select statements with the Masking Rule is straightforward. However, special cases may have to be handled by specific rules:

- Behavior of queries on masked fields: The default behavior is to query with the entered value against the actual database value.
- Behavior of update of masked fields: This phase ends when the acceptance test scenarios can be run successfully [4].

Step 4: Acceptance Test

Using the masking rules defined in step 3, all test scenarios defined in step 1 are rerun, this time with masked results.

3.2 Data Masking Process

The process of data masking is designed to de-identify data, such that the data remains based on real information, but no longer has any practical usage or application. In other words, it is now data rather than information. This involves the obfuscation of any information subject to any internal procedure, national law or industry regulatory compliance, such that if this data is subsequently used in an inherently insecure context such as application testing, the data no longer contains meaningful private or confidential information in any way. The number of fields that will require masking will of course vary greatly by application, organization and legislative requirements, but in general it would be preferable only to mask columns that require masking to maintain a straightforward process that ensures the integrity of the data. Furthermore there is no single answer to the correct masking methodology or algorithm; indeed it could well be argued that a preferable approach is not to employ a single technique in masking data [8][7]. Masking in a large enterprise may well include elements unique to that organization and combine multiple techniques, but as building blocks common masking techniques include:

1. Simple masking In essence sensitive data is simply replaced with a static set of null values such as XXXX or 9999. This technique is sometimes used in simpler manual masking processes, and it is secure in that the original information is obviously effectively masked. However this approach departs from production-like data and increases the risk both of testing problems resulting from this data and, more importantly, of testing not detecting problems that will occur when the applications are executed against live data.

2. Numeric manipulation at its simplest this approach basically increments or decrements the data by a given range. For example an order value could be increased by 5% or the data could be aged by adding say 1000 days to a date of birth. The simpler versions of this approach should however be treated with caution, a simple algorithm could be deciphered and again the data may no longer represent the production data characteristics.

3. Data substitution, this is a commonly used approach, and if used well can be extremely effective. Data is substituted with an alternative which can be determined randomly or through more sophisticated replacement mechanisms. The integrity of this approach is dependent on the data substituted; key of course is to preserve the usefulness of the data while obscuring its information value. In a form of Data Shuffling the data can be exchanged from different rows in the database, so one account number for example is now associated with the name of another account holder, and so on. Data substitution has the distinct advantage of preserving the variations and idiosyncrasies of the original production data. This approach

can be enhanced by using external data sources for substitution, for example a list of names from a phone book in place of customer names, other valid zip or postal codes for a given town, etc.Whatever techniques are employed, it is critical to define an appropriate action for each sensitive data element and to be able to repeatedly apply a consistent masking process which is propagated through all related data sources [6]

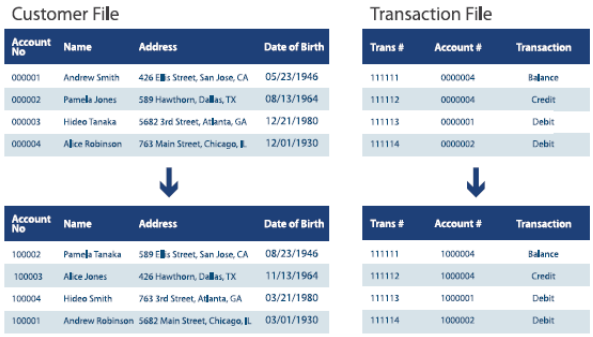

Fig-1: Simple data masking example, utilizing simple substitution and some basic numeric manipulation, preserving referential integrity

Five-step process can be defined:

1. Build the Knowledge Base from the production data through direct access or an unload process.
2. Analyze, inventory and classify the data in the Knowledge Base. This provides the key information for subsequent steps.
3. Define the extraction patterns and rules delivering repeated extraction schemes. These rules are reused in all subsequent extractions.
4. Perform the actual data extraction. This is a single process for both masked and subsetted data.
5. Load the reduced and secured test data into the test environment for testing against normal procedures [12].

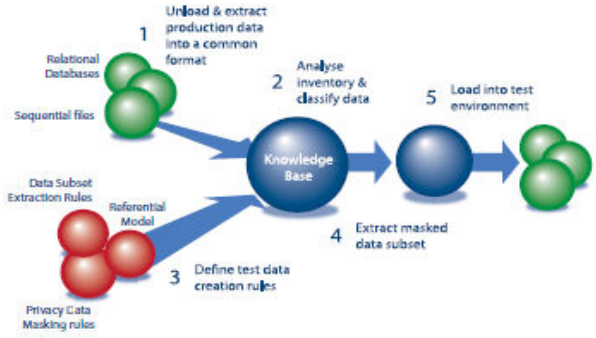

Fig-2: Five step process for masking and subsetting

Dynamic data masking Technique is used on large scale by the industries for data masking and with respect to the performance and security this method has been very much efficient. The way this method used is very much monotonous. For example, A column name with 10 characters data is replaced with X,#,$ or ? With the same number of characters. i.e the replaced symbol will also be repeated 10 times. Data and its graphical representation of the same is shown fig (2) and fig (3) :

Patterns	No Of Characters	Pattern	No Of Characters3
Replacement 1	4.3	2.4	2
Replacement 2	2.5	5	2
Replacement 3	3.5	5	4
Replacement 4	4.5	5	5

Fig-3: Table Data with Patterns and No of characters to be replaced

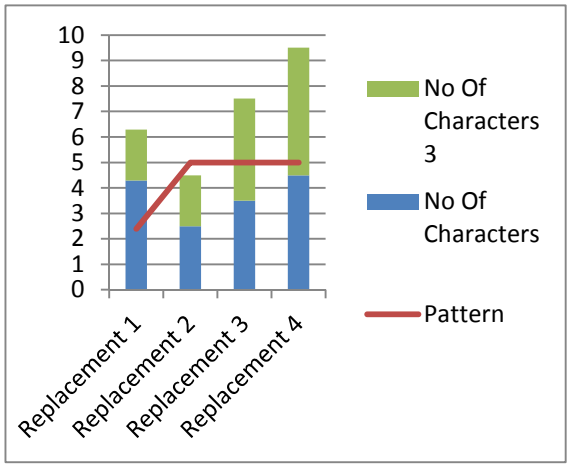

Fig-4: Representation of Replacement Technique.

The above graph depicts the stability of replacement rate across the industries. By far this method yields very good results. But as we see in the graph the replacement method is very much same for all sizes of data sets. And thus an enhancement to this method proposed would yield better results with more security.

4. RESULTS

A Detailed Report on the web-based application implementing dynamic data masking techniques with comparison of various masking techniques.

Domains	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-5: Table Data with Patterns and No of characters to be replaced

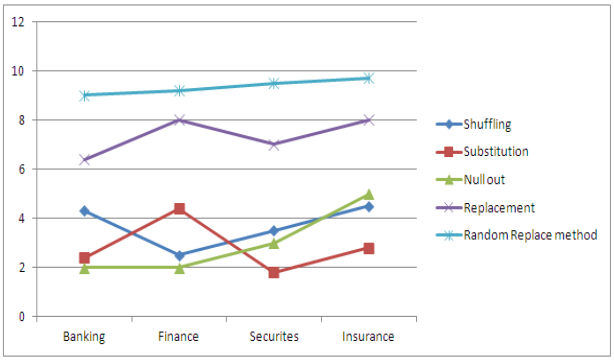

Fig-6: Comparison of various masking techniques with random replacement with respect to response time

4.1 Quick look at the UI after data is published:

Fig-7: Data Uploaded Page.

One can view the data that is uploaded into the database table on click of Show Data button. This Show Data button will display the table data that is uploaded recently

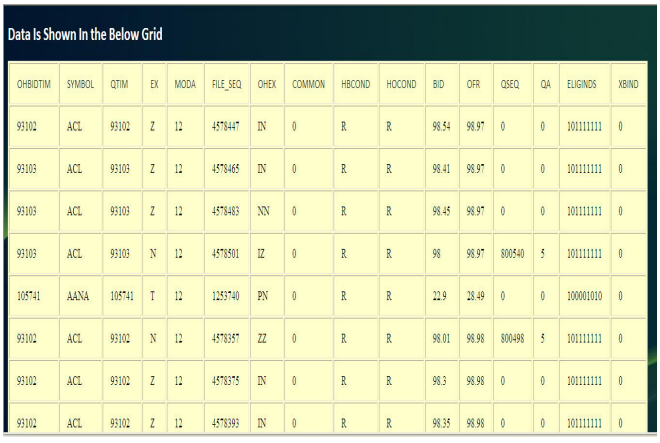

Fig-8: Data before masking.

On click BACK, user will be directed to the previous screen.

On click Mask, Data that is published recently can be masked but there is an authentication done here. Once the user clicks on Mask Button he/she is asked with the password to mask the data.

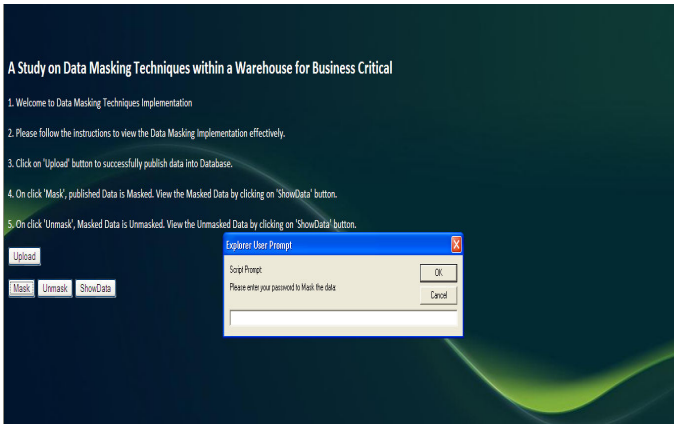

Fig-9: Authentication Page before masking.

On correct entry of the password user is let in for the Masking process else an alert message is displayed for the user saying it as an invalid password**.**

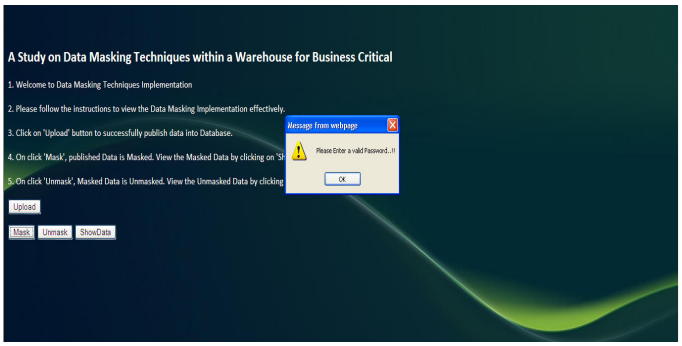

Fig-10: Authentication Page before Masking.

On correct entry of the password the data is masked and the required message is displayed for the user after the masking is done.

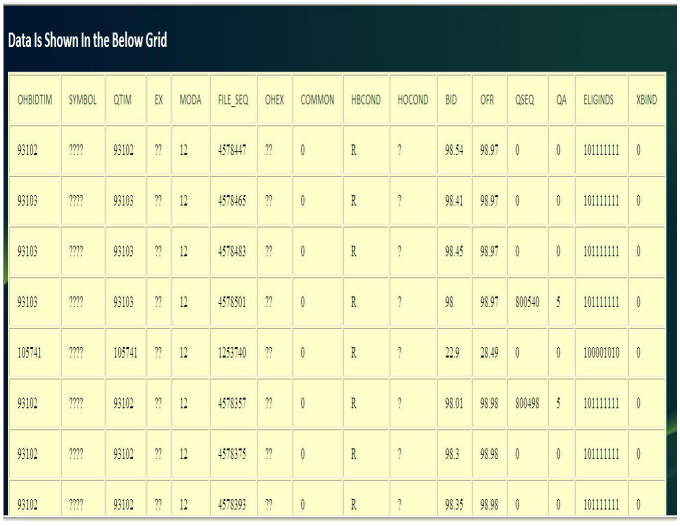

Fig-11: Data after masking.

5. CONCLUSIONS

This research addresses the necessity of dynamic data masking in present information age, no researcher has consolidated all the data masking techniques and importance of dynamic data masking for realistic situations.Comparision study of various techniques with the replacement method with respect to response time and our results shown random replacement is strongest method of data masking used across all domain which gives maximum confidence for all the customers and data masking will enable to accomplish the following:(a). Increases protection against data theft. (b). Enforces 'need to know access'. (c). Researchers in 2009 found that almost 80 to 90 percent of Fortune 500 companies and government agencies have experienced data theft. (d). Reduces restrictions on data use. (e). Provides realistic data for testing, development, training, outsourcing, data mining/research, etc. (f). Enables off-site and cross border software development and resource sharing (g).Supports compliance with privacy legislation & policies. (h). Data masking demonstrates corporate due diligence regarding compliance with data privacy legislation. (i).Improves client confidence. (j).Provides a heightened sense of security to clients, employees, and suppliers.This paper will help in analyzing the level of security needed for real time applications when publishing it in QA environment.

6. FUTUTE SCOPE

Security, privacy, and identity management will remain at the top of information security spending priorities the incidence of data breaches will continue to rise unless organizations enforce additional measures to protect sensitive data, both in production and non-production environments. We can enhance comparison study with respect other statistical features.

7. REFERENCES

[1] Muralidhar, K. and R. Sarathy "Generating Sufficiency-based Non-Synthetic Perturbed Data," Transactions on Data Privacy,1 (1).2008,p-17-33.

[2] Li, H., K. Muralidhar, and R. Sarathy, "Assessment of Disclosure Risk when using Confidentiality via Camouflage," Operations Research, 55(6), 2007,p-1178-1182.

[3] Muralidhar, K. and R. Sarathy, (2006) "A Comparison of Multiple Imputation and Data Perturbation for Masking Numerical Variables," Journal of Official Statistics, 22(3), 507-524.

[4] Muralidhar, K. and R. Sarathy, (2006) "Data Shuffling- A New Masking Approach for Numerical Data," Management Science, 52(5), 58-670.

[5]. Muralidhar, K. and R. Sarathy,(2005) "An Enhanced Data Perturbation Approach for Small Data Sets," Decision Sciences, 36(3), 513-529.

[6] . Muralidhar, K. and R. Sarathy,(2003) "A Rejoinder to the Comments by Polettini and Stander on 'A Theoretical Basis for Perturbation Methods'," Statistics and Computing, 13(4), 339-342.

[7] Muralidhar, K. and R. Sarathy, (2003)"A Theoretical Basis for Perturbation Methods," Statistics and Computing, 13(4), 329-335,

[8] Sarathy, R., K. Muralidhar, and R. Parsa,(2002) "Perturbing Non-Normal Confidential Attributes: The Copula Approach," Management Science, 48(12), 1613-1627.

[9] Muralidhar, K. and R. Sarathy,(2002) "What Could They Find? An Assessment of Security Risk," Proceedings of the 2002 National Meeting of the Decision Sciences Institute, San Diego, November.

[10] Sarathy, R. and K. Muralidhar, (2002)"The Security of Confidential Numerical Data in Databases," Information Systems Research, 13(4), 389-403.

[11] Muralidhar, K., R. Sarathy, and R. Parsa,(2001) "An Improved Security Requirement for Data Perturbation with Implications for E-Commerce," Decision Sciences, 32(4), 683-698.

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

[13] Muralidhar, K., R. Parsa, and R. Sarathy, (1999)"A General Additive Data Perturbation Method for Database Security," Management Science, 45(10), 1399-1415.

[14] Muralidhar, K., D. Batra, and P. Kirs (1995) "Accessibility, Security, and Accuracy in Statistical Databases: The Case for the Multiplicative Fixed Data Perturbation Approach," Management Science, 41(9), 1549-1564.

[15] Muralidhar, K. and R. Sarathy,(2008) "A Theoretical Comparison of Data Masking Techniques for Numerical Microdata," 3rd IAB Workshop on Confidentiality and Disclosure - SDC for Microdata, Nuremberg, Germany, November 20-21.

[16] Ravikumar G K, Dr.Justus rabi, Manjunath T.N, Ravindra S Hegadi,"Design of Data Masking Architecture and Analysis of Data Masking Techniques for Testing -IJEST11-03-06-217- Vol. 3 No. 6 June 2011 p.5150-5159

[17] Ravikumar G K, Manjunath T N, Ravindra S Hegadi, UMESH I M "A Survey on Recent Trends, Process and Development in Data Masking for Testing"-IJCSI- Vol. 8, Issue 2, March 2011-p-535-544.

[18] Manjunath T.N, Ravindra S Hegadi, Ravikumar G K."A Survey on Multimedia Data Mining and Its Relevance Today" IJCSNS. Vol. 10 No. 11-Nov 2010, pp. 165-170.

[19] Manjunath T.N, Ravindra S Hegadi, Ravikumar G K."Analysis of Data Quality Aspects in Data warehouse Systems", (IJCSIT)-Jan-2011.

8. AUTHORS PROFILE

Ravikumar GK. received his Bachelor's degree from Siddaganga Institute of Technology, Tumkur (Bangalore University) during the year 1996 and M. Tech in Systems Analysis and Computer Application from Karnataka Regional Engineering College Surthakal (NITK) during the year 2000. He is currently working towards his PhD degree in the Area of Data mining. He has published several papers in International and national level conferences. He is having around 14 years of Professional experienced which includes Software Industry and teaching experience. His area of interests are Data Warehouse & Business Intelligence, multimedia and Databases.

Dr. B. Justus Rabi: was born in Kanyakumari district of Tamilnadu.He obtained his Bachelor of Engineering Degree from Manonmonium Sundranar University. He has M.E. and Ph.D. Degrees from Madras University and CEG, Anna University, Chennai respectively His research interests include intelligent controllers, electrical drives, signal processing, wavelet transforms embedded systems. and data mining/making He has over 35 international publications to his credit. Currently he is working as an professor in the department of Electrical and Electronics Engineering at M.G.R University, Chennai.

Manjunath T N. received his Bachelor's Degree in computer Science and Engineering from Bangalore University, Bangalore, Karnataka, India during the year 2001 and M. Tech in computer Science and Engineering from VTU, Belgaum, Karnataka, India during the year 2004. Currently pursing Ph.D degree in Bharathiar University, Coimbatore. He is having total 10 years of Industry and teaching experience. His areas of interests are Data Warehouse & Business Intelligence, multimedia and Databases. He has published and presented papers in journals, international and national level conferences.