Indonesian Journal of Electrical Engineering and Computer Science
Vol. 7, No. 2, August 2017, pp. 449 ~ 456
DOI: 10.11591/ijeecs.v7.i2.pp449-456

A Big Data Security using Data Masking Methods

Archana RA*[1], Ravindra S Hegadi[2], Manjunath TN[3]
[1]R& D Centre, Bharathiar University, Coimbatore, Tamil Nadu, India
[2]School of Compuational Sciences, Solapur University, Maharastra, India
[3]Dept of ISE, BMS Institute of Technology, Bangalore, Karnataka, India
*Corresponding author, e-mail: archana.tnm@gmail.com

Abstract

Due to Internet of things and social media platforms, raw data is getting generated from systems around us in three sixty degree with respect to time, volume and type. Social networking is increasing rapidly to exploit business advertisements as business demands. In this regard there are many challenges for data management service providers, security is one among them. Data management service providers need to ensure security for their privileged customers in providing accurate and valid data. Since underlying transactional data have varying data characteristics such huge volume, variety and complexity, there is an essence of deploying such data sets on to the big data platforms which can handle structured, semi-structured and un-structured data sets. In this regard we propose a data masking technique for big data security. Data masking ensures proxy of original dataset with a different dataset which is not real but looks realistic. The given data set is masked using modulus operator and the concept of keys. Our experiment advocates enhanced modulus based data masking is better with respect to execution time and space utilization for larger data sets when compared to modulus based data masking. This work will help big data developers, quality analysts in the business domains and provides confidence for end-users in providing data security.

Keywords*: Big Data, Data Security, Data Masking*

Copyright © 2017 Institute of Advanced Engineering and Science. All rights reserved.

1. Introduction

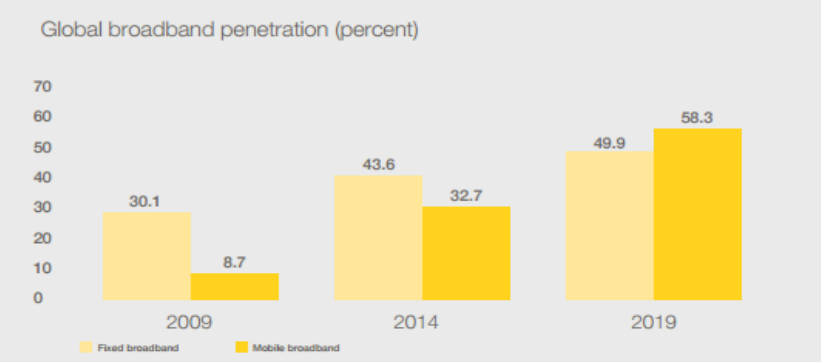
In Today's business world, Business organizations are more tend to use social media to exploit the business opportunities, social media applications are increasing day by day to make all their transactions and communications easy. Because of such transactions, huge volumes of heterogeneous data sets are generated. All these data sets need to be stored on big data platform-Hadoop for data analysis/predictive analysis. Business domains like banking, retail, insurance, finance and security are using intelligent applications to make their business effectively. Information management solutions are no longer just a business enabler but an integral part of providing enhanced customer services. Information management is crucial for management of data efficiently and effectively. However, an overlooked risk for various domain in security space for data analysis. The various factors on security aspects such as types of real data used in software development,testing, Information security. Use of cloud computing, distributed computing, outsourced services and which experiences data breaches involving real consumer data. In this connection data masking provides data protection to address privacy concerns. Data masking replaces sensitive data with a non-sensitive substitute, but does so in a way that preserves the integrity of the data. This means masked data can be used to facilitate business processes without changing the supporting applications, databases or data storage facilities which enables you to remove the risk without breaking your business. Securosis Research has developed five laws for data masking: (a) Masked data should not be reversible. (b) Masked data should be representative of the original data set. (c) Masked data should maintain application and database integrity. (d) Non-sensitive data should be masked only if it can be used to re-create or tie back to sensitive data and (e) Data masking routines must be repeatable. One-off masking is both ineffective and impossible to maintain. Today's Information Technology environments are highly dynamic, and masking routines need to keep pace. InfoSphere Optim Data Privacy provides a comprehensive set of data masking techniques to support data privacy and compliance requirements. For the first time, you can mask data across platforms, across data sources using a standard and repeatable process to ensure data privacy

Received April 25, 2017; Revised June 8, 2017; Accepted June 23, 2017

without impacting the stability of your applications with greater ease and unparalleled scalability and performance. With InfoSphere Optim Data Privacy, you mask and move. Masking and moving allows you to extract and mask data, and then insert or load the data into one or more destinations. Masking in place allows you to de-identify data and replace existing values. InfoSphere Optim Data Privacy provides the most comprehensive set of data masking techniques on the market. The method you use will depend on the type of data you are masking and the result you want to achieve. Out-of-the-box capabilities for specific data types are included, such as random or sequential number generation, string literal substitution, concatenating expressions, arithmetic expressions, lookup values and user-defined functions, to name a few. Some examples of situations in which masking techniques can be applied includes are data at rest or data in flight, relational data, flat files and data sets such as IBM IMS or VSAM, Data being transformed through an extract, transform and load (ETL) tool, Data accessed in SQL queries inside a database, Data in reports and documents, Data inside applications, Data moving to, in and from big data platforms such as Hadoop, Data used for testing big data environments, data used for analytics applications for example, PureData Analytics or Teradata and data used for testing data warehouses.Some of the Benefits of Data masking Focus on data security and privacy to deliver significant value such as prevent data breaches, ensure data integrity, reduce cost of compliance and protect privacy.

2. Related Work

In contemporary information technology trend, social media is one among the top communication media to share the opinions of likes and dislikes. According to Featured Insights, Global, Media and Entertainment report, internet users spend more time with social media sites than any other type of site. At the same time, over the next five years, we project mobile broadband penetration to overtake fixed broadband, rising to 58.3 percent of the total in 2019, from 32.7 percent in 2014 (Figure 1). Fixed broadband penetration growth will slow over the same period, rising by just 6.3 percentage points, from 43.6 percent in 2014 to 49.9 percent in 2019.



Source: McKinsey & Company, Wilkofsky Gruan Associates

Figure 1. Global Broadband Penetration

In this regard data security is very important for any business marketing across social media sites, intruder or insider may steal the data which imbalances the business in the business market. The process of obscuring specific data elements within data stores is called as data masking. It ensures that sensitive data is replaced with realistic but not real data. The goal is that sensitive customer information is not available outside of the authorized environment. Data masking is an effective strategy in reducing the risk of data exposure from inside and outside of an organization and should be considered a best practice for curing non-production databases [14] [15]. The following authors have exertion on data masking algorithms for various business domains in their perspectives. Muralidhar, K., R. et.al, proposed his work on “Random Data Perturbation for Non-normal Data,” in 2000, Proceedings of Annual Meeting of the

Decision Sciences Institute. Sarathy, etl.al presented his work on "The Two Step Data Shuffle: A New Masking Procedure," in Invited seminar presented to the Census Bureau and the Washington Statistical Society, in 2002. Later in 2003 they gave the idea of "The Data Shuffle: A New Masking Procedure for Numerical Data," in 8th INFORMS Computing Society Conference Sarathy, R and K Muralidhar gave the idea of "Data Masking - Problems, Solutions, and Opportunities," in TRDDC - TCS, Pune, India in 2006 [16] [17]. Muralidhar K and R Sarathy gave the idea of "Privacy Violations in Accountability Data Released to the Public by State Educational Agencies," in Federal Committee on Statistical Methodology Research Conference, in 2009. Ravikumar G K,et.al in (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 8, August 2011, Data Masking Techniques with Random Replacement using data volume, which is difficult for hackers to steal the data. Ricardo Jorge Santos et.al, International Joint Conference of IEEE TrustCom, 2011, showed how balancing Security and Performance for Enhancing Data Privacy in Data Warehouses using modulus based method. Muralidhar K and R Sarathy, states “Interval Responses for Queries on Confidential Attributes: A Security Evaluation.” Journal of Information Privacy and Security, 9(1), 3-16, 2013. A white paper by camouflage data masking specialist, titled “A Proactive Approach to Data Security for Cloud-Based Testing and Development”, May 2014, emphasis any cloud-based application development offers organizations many tangible benefits, yet organizations struggle with how to work with data in the cloud-big data while complying with key regulations and meeting data security requirements. Data masking offer organizations a way to guard data for big data application development/testing using confirmed expertise while extenuating the risk of a security breach. Vishnu B et.al, International Journal of Computer Applications, recent Advances in Information Technology, 2014, proposed An Effective Data Warehouse Security Framework which highlights on the usage of modulus operator in data security for data warehouse system [2] [3]. No literatures found on creating uniform data security framework using enhance modulus based for big data. Hence, hereby a model is proposed which can be uniformly used across the industry for data security on big data environment which are business critical. According to recent Gartner research, unstructured data accounts for at least 80% of an organization’s data. If left unmanaged, the sheer volume of unstructured data that’s generated each year within a company can be costly in a number of ways, ranging from security vulnerabilities to compliance risks. The new era of computing has arrived: organizations are now able to process, analyze and derive maximum value from structured, unstructured and streaming data in real time. However, in the rush to achieve new insights, are privacy concerns being neglected, how can you support business goals while also ensuring the privacy of sensitive data, with the average cost of security-related incidents in the era of big data estimated to be over USD40 million, according to this Aberdeen Group Research Brief, you can’t afford to ignore data privacy as a top requirement. With 2.5 quintillion bytes of data created every day, now is the time to understand sensitive data and establish business-driven privacy policies to keep customer, business, personally identifiable information (PII) and other types of sensitive data safe. Remember, however, that different types of data will require different protection policies. For example, text, audio, log files and click streams have unique characteristics and challenges around privacy.

3. Big Data Environment

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of 2016 ranging from a few dozen terabytes to many petabytes of data. Big data is not just about volume or rate of acquisition, but also heterogeneity/diversity, Multiple levels of granularity, Multiple media and modalities, Scientific disciplines, Complexity, Uncertainty, Incompleteness and Representation Opportunities. Big Data presents unprecedented opportunities to accelerate scientific discovery and innovation, Lead to new fields of inquiry that would not otherwise be possible, Improve decision making, Understand human and social processes, Promote economic growth and Improve health and quality of life. In this connection we deploy variety of data on hadoop platform for its usability based on the customer needs. Hadoop is an Apache top-level project being built and used by a global community of contributors and users. Hadoop was created by Doug Cutting and Mike cafarella in 2005. Named in the remembrance of his son's toy elephant. The Apache Hadoop

framework is composed of the following modules: Hadoop Common which contains libraries and utilities needed by other Hadoop modules. Hadoop Distributed File System (HDFS) it is a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster. Hadoop YARN, it is a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications. Hadoop MapReduce, it is a programming model for large scale data processing. Below diagram shows the HDFS architecture which has name node and corresponding data nodes in HDFS Layer, similarly MAPREDUCE layer will have job tracker and task tracker.

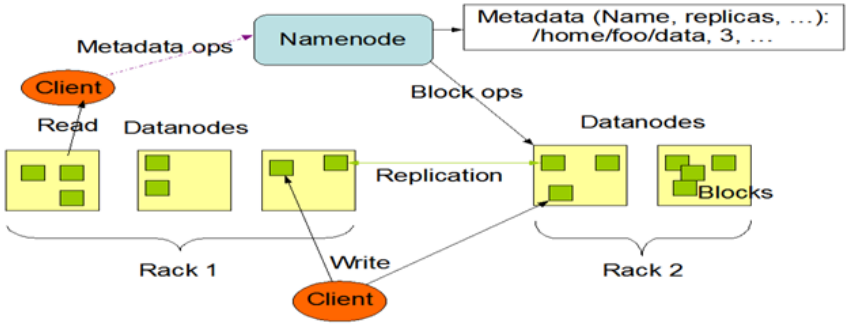


Figure 1. HDFS Architecture

3.1. Mathematical Model

In our technique we make use of two keys, where both are private (i.e. the keys are known only to the authorized personnel). Consider a table say 'T' which has 'm' rows represented as (R1, R2, …., Rm)and 'n columns represented as (C1, C2, …., Cn). Now let us consider a value say (Ri, Cj) that has to be masked where 'Ri' and 'Cj' represents the row value and column value repectively. Now the private keys used are K1 and K2. K1 is a random 128-bit random generated integer that remains constant for the whole table.K2 is another 128-bit random generated integer that remains constant for a single column, whose value ranges between the maximum and minimum value of the column(in the above case for column Cj). Now for masking the data we use the formula:

$$(Ri, Cj)` = (Ri, Cj) – (K2 \% K1) + K2 \quad (1)$$

Now for de-masking we use the formula:

$$(Ri, Cj)` = (Ri, Cj) – (K2 \% K1) + K2 \quad (2)$$

Equation (1) is used for masking the data and equation (2) is used for de-masking.

3.2. Proposed Algorithm

The proposed algorithm is used for data security using Enhanced Modulus based security with the concept of keys.

```
Begin algorithm
for each table n in the target database TD(1….N)
fetch K1 private key common for the entire table(n)
for each attribute j in the table
        fetch K2,j  private key common for entire attribute( j)
        for each item (Ri,Cj) in the table(n)
                fetch K3,i public key common for a tuple (i)
                find the length of the item (Ri,Cj) i.e len=strlen((Ri,Cj))
                if len NOT even
                        append zero to (Ri,Cj)
                for each pair of character/digit in (Ri,Cj)
                        convert character into integer and store in numb
                        apply masking formula for numb
                        append to (Ri,Cj)' //every loop append with previous values
                end for
        store (Ri,Cj)' in place of (Ri,Cj)
        end for
end for
end for
End algorithm
```

Figure 2. Algorithm for data security using E-MOD

4. Issues of Data Masking in Security

Business enterprises emphasized to covenant the following issues for a variety of data masking techniques such as (i) Risk minimization: No matter what security measures are taken, there is always a degree of risk involved in handling a large amount of sensitive data [3] [6]. Data breaches can damage a company's reputation, increase liabilities and invite legal suits (ii) accountability: Data breaches create negative publicity, harm current and future business, and damage organization's reputation and the client's confidence in it. It is crucial that the organization stays accountable to all stakeholders, customers and employees, and addresses their privacy needs effectively and (iii) regulatory norms: Confidentiality and privacy norms demand the protection of data against theft. Compliance to all norms is essential to prove the organization commitment to its prestigious customers [7] [8]. Data Masking Impediments. The points from a to e highlights the impediments of data masking methods in financial firms which are critical to business. (a). Data Utility: Masked data should look and act like real data. Data must be fit for proper testing and development, application edits and data validation (b). Data Relationships: Must be maintained after masking on Database level Referential Integrity (RI), Application level RI, Data Integration (Interrelated database RI).(c). Existing Business Processes: Must fit in with existing IT and refresh processes.(d). Ease of use: Must balance ease of use with need to intelligently mask data Usable data that does not release sensitive information and Knowledge of specialized IT/privacy topics and algorithmic importance should be pre-configured and built into the masking process.(e). Customizable: Solution/Process must be capable of being tailored to specific needs of the clients [9] [10].

5. Results and Discussions

The proposed methods provide flexibility around how the data will be masked and ensure that business rules of the enterprise application will not be impacted. After data segregation, the masking type will be decided based on the data such as substitution, replacement, multiplier, randomizer and shuffling, the same is illustrated below with example. Now the proposed technique E-MOBAT is compared with MOBAT taking the overheads - Time taken for execution and storage space overheads. The results prove to be a lot in the favor of E-MOBAT proved to be better than MOBAT.

1) Storage Space Overheads

Table 1. Storage Space Overheads

Size of Data (in MB)	MOBAT (in MB)	E-MOBAT (in MB)
2.5	3.015	2.19
5.0	6.223	4.30
10.0	12.25	8.22
20.0	24.42	16.50
40.0	48.44	32.90

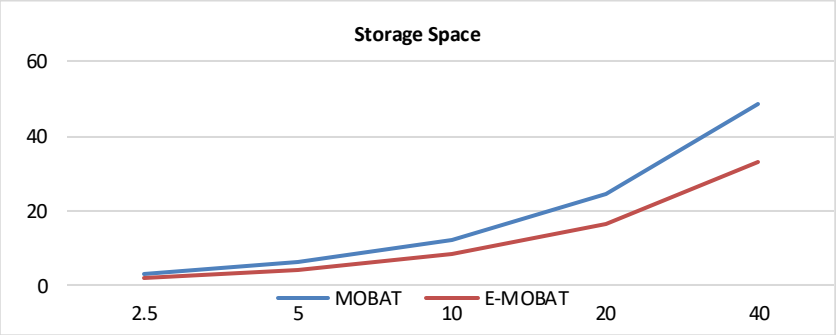


Figure 3. Storage Space Overheads

2) Execution Time Taken

Table 2. Execution Time Taken

Size of Data (in MB)	MOBAT (time in seconds)	E-MOBAT (time in seconds)
2.5	0.102	0.061
5.0	0.210	0.114
10.0	0.401	0.242
20.0	0.820	0.498
40.0	1.658	0.988

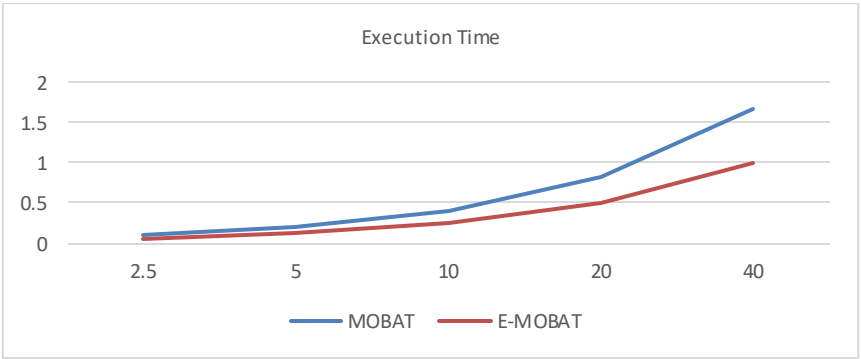


Figure 4. Execution Time Overheads

Thus the above results prove that the E-MOBAT technique has a better performance curve than the existing MOBAT inturn making it better than the already existing algorithms such as AES128, 3-DES, etc.The above graph shows the usage of primary memory (RAM) by both, the MOBAT and E-MOBAT to complete the execution of a query for varying number of values. The value in the x-axis represents the number of rows with each row having values from 9

different columns. The E-MOBAT uses the same amount of RAM as the MOBAT for smaller data-sets. But in real time scenarios involving millions of rows, the E-MOBAT functions at comparatively less RAM, than the MOBAT. The reduction in use of the Primary Memory is due to revision of the formulae used in the MOBAT. Proposed hybrid data masking method is a general approach that deals with the needs of security problems faced by various organizations when onsite-offshore business delivery model is used. Our hybrid data masking model framework ensures two principles while operation is carried out (i) Masking is not reversible. There is no way to reverse engineer the original data from the masked data and (ii) Masked data is usable. For example, when testing valid addresses the masked data must include valid zip codes not random numbers which fit the data type.

6. Conclusion

In current information business market, masking offers unique value beyond other data security tools both in its ability to preserve complex data relationships while protecting data, and its data management capabilities. Masking's combination of discovery, data set management, protection, and control over data migration is unique. No other data security product provides all these benefits simultaneously. Masking reduces risk with minimal disruption to business systems. These characteristics suit masking to meeting compliance requirements. The rapid growth we have seen in the data masking segment spurred by compliance, risk, and security demands has driven impressive innovate to capture increased customer demand. The proposed data security system is better then the MOBAT with the enhancement of Modulus operator and concept of keys for big data security and tested fir various data loads and compared the results between MOBAT and E-MOBAT with respect to storage and memory requirement, this is tested on MongoDB.

References

[1] Sonja Murdoch. "*Global Media Report 2015*". Global Industry Overview, Mckinsey Company. 2015: 1-29.
[2] Vishnu B, Manjunath T N and Hamsa C. "*An Effective Data Warehouse Security Framework*". IJCA Proceedings on National Conference on Recent Advances in Information Technology NCRAIT. 2014: 33-37.
[3] Clement Almeida, Harshitha K, Manjunath TN. "A Study on Column Segregation for Data Security". IJRCSIT. ISSN No.: 2319-5010 2014; 2(2).
[4] Manjunath TN, Ravindra S Hegadi. "Data Quality Assessment Model for Data Migration Business Enterprise". *International Journal of Engineering and Technology (IJET).* ISSN: 0975-4024, 2013; 5(1).
[5] Manjunath TN, Ravindra S Hegadi. "Statistical data Quality model for data migration business enterprise". *International Journal of Soft Computing.* 2013; 8(5): 340-351, ISSN 1816-9503.
[6] Ravikumar GK, et.al. "A Survey on Recent Trends, Process and Development in Data Masking for Testing". *IJCSI International Journal of Computer Science Issues*. 2011; 8(2).
[7] Ravikumar GK et.al. "*Design of Data Masking Architecture and Analysis of Data Masking Techniques for Testing*". IJEST11-03-06-217, 2011; 3(6): 5150-5159.
[8] Understanding and Selecting Data Masking Solutions - Creating Secure and Useful Data-Securosis, L.L.C. Data Masking: What You Need to Know What You Really Need To Know Before You BeginA Net 2000 Ltd. White Paper.
[9] Allen Dreibelbis, Eberhard Hechler, Ivan Milman, Martin Oberhofer, Paul van Run, Dan Wolfson. "Enterprise Master Data Management: An SOA Approach to Managing Core Information". Dorling Kindersley (India) Pvt. Ltd. 2008.
[10] Ralph Kimball and Joe Caserta. "The Data Warehouse ETL Toolkit". Wiley Publishing, Inc. Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications - Batini, Scannapieco. 2006.
[11] Kyung-Seok Ryu, Joo-Seok Park, and Jae-Hong Park. "A Data Quality Management Maturity Model". *ETRI Journal*. 2006; 28(2).
[12] Manjunath TN et.al. "Analysis of Data Quality Aspects in Data Warehouse Systems". *(IJCSIT) International Journal of Computer Science and Information Technologies*. 2011; 2(1): 477-485.
[13] Manjunath TN, Ravindra S Hegadi, Ravi Kumar GK. "Design and Analysis of DWH and BI in Education Domain". *IJCSI International Journal of Computer Science Issues*. 2011; 8(2), ISSN (Online): 1694-0814.545-551.

[14] Manjunath TN, Ravindra S Hegadi and Mohan HS. Article: "Automated Data Validation for Data Migration Security". *International Journal of Computer Applications.* 2011; 30(6): 41-46.
[15] Xiao-Bai Li, Luvai Motiwalla BY "Protecting Patient Privacy with Data Masking". *WISP*. 2009.
[16] Domingo-Ferrer J and Mateo-Sanz JM. "Practical Data-Oriented Microaggregation for Statistical Disclosure Control". *IEEE Transactions on Knowledge and Data Engineering*. 2002; 14(1): 189-2011.
[17] A Bonifati, F Cattaneo, S Ceri, A Fuggetta and S Paraboschi. "Designing data marts for data warehouses". *ACM Transactions on Software Engineering Methodologies*. 2001; 10(4): 452-483.
[18] Muralidhar K, R Sarathy and R Parsa. "An Improved Security Requirement for Data Perturbation with Implications for E- Commerce". *Decision Sciences*. 2001; 32(4): 683-698.
[19] Muralidhar K and R Sarathy. "Security of Random Data Perturbation Methods". *ACM Transactions on Database Systems*. 1999; 24(4): 487-493.
[20] Muralidhar K, R Parsa and R Sarathy. "A General Additive Data Perturbation Method for Database Security". *Management Science*. 1999; 45(10): 1399-1415.
[21] Muralidhar K, D Batra and P Kirs. "Accessibility, Security, and Accuracy in Statistical Databases: The Case for the Multiplicative Fixed Data Perturbation Approach". *Management Science*. 1995; 41(9): 1549-1564.
[22] Muralidhar K and R Sarathy. "*A Theoretical Comparison of Data Masking Techniques for Numerical Microdata*". 3rd IAB Workshop on Confidentiality and Disclosure - SDC for Microdata, Nuremberg, Germany. 2008.