International Journal For Technological Research In Engineering
Volume 6, Issue 9, May-2019 *ISSN (Online): 2347 - 4718*

A NOVEL DATA MASKING TECHNIQUES TO SECURE HEALTHCARE DATA

Siddartha B.K[1], Dhanya K.R[2], Harshitha K.H[3], Chaithra Nagesh B.N[4], Divya Shree K.T[4]
[1]Assistant Professor, [2345]8th Semester Students
Department of Computer Science and Engineering
B G S Institute of Technology, B G Nagara, Mandya-571448

Abstract: ***Healthcare sector area unit undergoing tremendous modification in analytical computing. The implementation of a Business Intelligence (BI) platform in numerous forms of sectors as well as attention, has become a crucial project. The common implementation is to extract sensitive information from production databases then load them into Associate in Nursing Information Warehouse. information masking techniques area unit wont to cut back the unintended revealing risk of sensitive information moreover on preserve the essential quality of information analytics and supply security from internal breaches (developers, researchers and testers). Reversible information masking techniques area unit wont to decrypts the sensitive information, preserve the information format, and maintain the standard of information utility (data analytics). during this paper a sensible intrinsically information masking framework (IMETU- establish, Map, Execute, Test, and Utilized) is planned specializing in the execution and testing modules. we offer finish to finish secret writing to supply security and fewer masking delay. The planned information masking algorithmic rule springs from the applied mathematics content of the extracted information set, that is sorted at sure levels (micro-aggregation) that area unit related to a numeric attribute. the mix of the connected applied mathematics info are going to be used at intervals a mathematical formula to get the new covert worth, then the applied mathematics variables can place along during a sequence and encapsulated to create a powerful try of public keys and secret keys. This strengthens the safety issue whereas introducing little overheads in performance and area as compared with previous secret writing techniques.***

Keywords: Data Warehouse, Health Data, Business Intelligence

I. INTRODUCTION

The aging demographics in Ontario and resultant health system pressures ar driving important health system funding reforms. Standardizing care of high-volume procedures to established best practices, tied to regional and structure funding, is mandating innovation, efficiency and collaboration between native system partners. One documented, high-volume health condition in older adults is hip fracture. Rehabilitative care is a vital element of associate integrated health care system, and recognized as crucial to facilitate the comeback of older adults to freelance living following health problem, injury and hospitalization. Understanding ability of patients to access rehabilitative care across the care time is important to coordinate amendment efforts in system amendment management. correct knowledge on volumes of patients with hip fracture, discharge tendencies, and patient outcomes may be difficult to get for system analysis and directional quality improvement. Business intelligence applications provide an opportunity to secure timely and valid data for these functions.

Business Intelligence is made public as "broad category of applications and technologies for gathering, storing, analyzing, sharing, and providing access to knowledge to assist enterprise users build higher business decisions" [1]. A BI platform consists of many interdependent components with a logical work flow as follows, and as shown in Figure1: [3].

A. The external data source: External data sources are variety types of databases, structured and unstructured flat files that considered the basis of the entire BI architecture as they feed the BI solution. However, they're not a part of the metal atmosphere however all the kinds of heterogeneous knowledge attributes ought to be completely understood by the designer and developers. [2]

B. Data Staging Area: The Staging area is an off-line copy of many operational data store systems with some of the following characteristics:

- Has a demoralized form of data mode
- Has a relational data model.
- Able to store historical and real-time data.

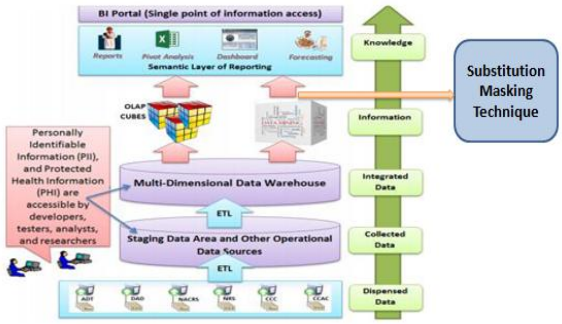

Fig 1: Business Intelligence Architecture

As shown in figure (1), the staging storage area and multidimensional data warehouse (DW) are considered the core components of the BI platform which are an integrated repository derived from multiple knowledge sources (operational and legacy) within the production surroundings. Saving sensitive knowledge into a central repository could be a serious privacy speech act threat once several internal users

of various security levels have access to the atomic number 83 services. Thus, knowledge privacy and therefore the reliability/utility square measure thought-about necessary problems that may be compromised within the atomic number 83 platform whereas mistreatment knowledge masking techniques. Most aid information science professionals take it without any consideration that the sharing of clinical/financial information at intervals the bismuth Platform for the aim of information analysis and analysis will have several advantages. However, the question is how to do so in a way that protects individual privacy and complies with the regulations as well as ensures the basic quality of data analytics which are useful and meaningful. Therefore, the right masking or de-identification techniques to protect sensitive data and preserve the data utility must be chosen to automate manual processes that lead to design a built-in data masking framework within the BI platform. In this paper, we propose a conceptual framework which is called IMETU. This framework consists of 5 modules (Identify, Map, Execute, Test, and Utilize). The first two modules are the most critical and important that we cover in this paper. Native encryption techniques (Database encryption, Symmetric, Asymmetric encryption) [11] are avoided due to complex calculations, increased storage space, and large overheads in query response time.

II. DYNAMIC DATA MASKING TECHNIQUES

A newer dynamic knowledge masking model is wherever knowledge is covert by associate in-line service because it goes into, or comes out from, a database. The results of a user query are intercepted and masked on the fly, with masked results transparently substituted before being returned to the user. For example, if a user queries too many credit card numbers, or if a query originates from an unapproved location, the returned data might be redacted [4]. This differs from view-based masking therein it happens outside the information repository and is offered for non-relational systems. The proxy may be an agent on top of the database, an appliance deployed 'inline' between users and data, to force all requests through the proxy. The huge advantage is data protection is enabled without need to alter the database; there are none of the additional programming and quality assurance validation processes. Another recent advance in dynamic knowledge masking is question substitution. In this variation the masking platform is smart enough to recognize a request for sensitive data. Such queries are intercepted and re-written to select information from a Securosis-Understanding and Selecting Data Masking Solution different (masked) column.

Fig 2: Substitution method

Currently several knowledge masking varieties square measure offered the subsequent square measure the necessary knowledge masking techniques [22]. 3.1 Substitution The Substitution technique replaces the prevailing knowledge with random values from a pre-prepared dataset i.e this technique consists of randomly replacing the contents of a column of data with information that looks similar but is completely unrelated to the real details [5]. For example, the surnames in a very client info may be alter by exchange the $64000 last names with surnames drawn from a biggish random list. Substitution is incredibly effective in terms of conserving the design and feel of the prevailing knowledge. The downside is that a largish store of substitutable information must be available for each column to be substituted. For example, to sanitize surnames by substitution, a listing of random last names should be offered. Then to sanitize telephone numbers, a list of phone numbers must be available [6]. Frequently, the ability to generate known invalid data (credit card numbers that will pass the checksum tests but never work) is a nice-to-have feature. Substitution knowledge will generally be terribly arduous to seek out in giant quantities - but any knowledge masking software package ought to contain datasets of ordinarily needed things. When evaluating knowledge masking software package the scale, scope and sort of the datasets ought to be thought-about. Another useful feature to look for is the ability to build your own custom datasets and add them for use in the masking rules [6].

III. HEALTHCARE ATTRIBUTES

Before navigating the main points at intervals the masking framework, we want to own a glance at the health knowledge attributes[7]. the aim of the "Personal Health info Protection Act (PHIPA, Ontario 2004)" is to determine rules for the gathering, use, and also the revelation of non-public health information (PHI) concerning people that shield the confidentiality of {the info|the knowledge|the data} and also the privacy of people, so as to facilitate the effective provision of care [8]. Personal Health info (PHI), suggests that "identifying info concerning a private in oral or recorded type that relates to the physical or mental state of the individual; relates to the providing of care to the individual; relates to payments or eligibility for care, additionally to, arrange of service at intervals the that means of the house Care and Community Services Act" [8]. Note that, characteristic info suggests that, info that identifies a private or that it's fairly predictable within the circumstances that it

				Reporting Time Intervals					
Data Attribute	Data Type	Sample	PHI/PII	Annually	Quarterly	Monthly	Weekly	Daily	Hourly
Health Card Number *	Numeric	1234 567 890	PHI	x	X	X	x	x	
Visit Number/ Encounter Number	Numeric	123456789	PHI	x	x	X	x	x	
Credit Card Number	Numeric	1234567890123	PII						
Full Name	Text	John Smith	PII					x	
Sex	Text	M	PII	x	x	X	x	x	x
Birth Date *	Date Time	1975/06/01	PII	x	x	X	x		
Postal Code *	Text	N0N1J2	PII	x	x	X	x	x	
Patient Municipality Code/ Resident Code *	Numeric	0501	PII	x	x	X	x	x	
Admit Date *	Date Time	2015/12/01	PHI	x	x	X	x	x	
Admit Time	Date Time	9:30 PM	PHI					x	x
Admission Category	Text	LIFE THREATENING CONDITION/URGENT/IMMEDIATE ASSESSMENT	PHI	x	x	X	x	x	
Facility (Hospital)	Text	(0966) BLUEWATER HEALTH	-	x	x	X	x	x	
Discharge Date *	Date Time	2015/12/10	PHI	x	x	x	x	x	
Discharge Time	Date Time	7:20 PM	PHI		x			x	x
Discharge Status	Text	DISCHARGED TO HOME WITH NO SUPPORT SERVICES	PHI	x	x	x	x	x	x

Table (1): Data Attributes of Discharge Abstract Database(DAD)

may be utilized, either alone or with different info to spot a private [8]. supported the definition of PHIPA, HIPAA ("Health Insurance movability and irresponsibleness Act"), or the other health privacy regulation, it's therefore crucial to safeguard care knowledge whether or not from patient

perspective or supplier perspective. it had been created primarily to modernized the flow of care info, stipulate however in person classifiable info maintained by the care and care insurance industries ought to shielded from fraud and address limitations on care insurance coverage[12]. during this work we have a tendency to use the foremost common information to grasp the sensitive health knowledge attributes that require to be obfuscated and guarded from revelation to internal and external users. Our approach is to investigate patient discharge knowledge (acute care) that is obtained from the "Discharge Abstract information (DAD)" system that was developed by the Ministry of Health of Ontario and also the "Canadian Institute for Health info (CIHI)" [9]. This information contains several knowledge tables that embrace elaborate patient level "abstract" for all variety of hospitalization services in one standardized supply as a patient journey at intervals hospitals' units. the information collections contains demographic, clinical, and body knowledge for all acute care discharges in Ontario [10].In order to make the classification element of the information masking framework, an intensive analysis has been conducted against the begetter knowledge set and its needed attributes from the enterprise news perspective (Utilization, Quality Improvement, Quality based mostly Procedure, Clinical Performance, Workload, etc.) and their news time intervals, as shown in Table (1). The sign (x) shows the usage of information attributes at intervals specific time intervals of the specified reports. Also, every knowledge attribute has been categorized on whether or not it holds sensitive knowledge (PHI or PII), or non-sensitive knowledge (-).

IV. ESTABLISH THE SENSITIVE INFORMATION

Based on the domain of tending business rules, that complies with the information privacy laws (such as PHIPA or HIPPA), we tend to conducted an intensive investigation of various business intelligence comes that area unit being designed for tending settings and mentioned these best practices with the domain consultants during this regard [7]. we tend to reached a agreement on the way to utilize reversible and irreversible information masking and de-identification techniques to guard sensitive health information which will not have an effect on the information usability. the final transformations for masking ways that we follow are:

• Health Card variety (HN): can convert to a novel patient symbol (Masked HN) that appointed to every patient that links their activity across all partner Health Service suppliers (HSP). For this sort of attribute, can apply unidirectional masking (irreversible) technique, like a way hashing with protective format.

• Chart variety and Visit variety: are hashed to a numeric format that the same as Health Card Number.

• Date of Birth (DOB YYYY/MM/DD): can convert to age bracket (Text) or to Age (Numeric) by applying further de-identification algorithmic program like Format protective encoding (FPE) or modulus primarily based variety alteration. • Address: can convert to Municipality (4 digits Residence Code), County (first a pair of digits of Municipality Code), Province, and postcode. we are able to apply de-identification algorithmic program like FPE or modulus primarily based transformations. postcode will be rolled up to FSA level (first three characters) by applying the covert out algorithmic program. alternative earth science data will be derived by exploitation the Municipality and/or postcode, as shown as follows: [10] • Activities' Date Stamps: (i.e., Admission Date Time, Discharge Date Time, Surgery Date Time) area unit replaced by yr, year, business Quarter, Calendar Quarter, and Month. Also, date travel masking will be applied to extend or decrease the date price among acceptable time vary (alteration among a similar month or a similar week). Additional de-identification algorithmic program will be used, like FPE or modulus primarily based transformations to keep up daily coverage values if required.

• Numeric Valued Attributes: like Length Of keep (LOS) in Acute patient, or Emergency Wait Time in Hours, etc. This may well be covert by numeric alteration among affordable vary, or exploitation the reversible masking ways like FPE or modulus primarily based transformation.

• alternative matter information attributes: they're thought of as indirect identifiers and would like to use the de-identification algorithms for the whole attribute like FPE or Pseudonymization ways

V. IMPLEMENTATION

The implementation of the AES-128 coding and cryptography formula with the assistance of MATLAB software system is

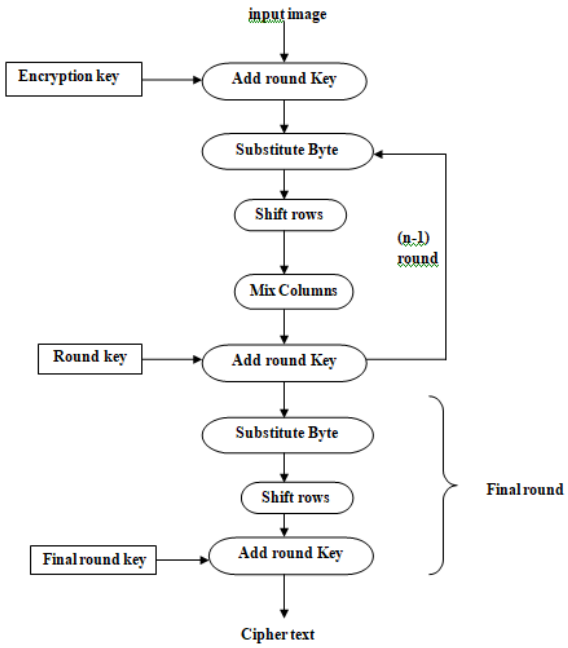

Fig.3: Flowchart of AES Encryption algorithm

multidimensional language of AES coding formula done. during which the input is a picture and therefore the key in positional notation format and therefore the output is that the same as that of input image. For coding method 1st, dividing image and creating it 4*4 computer memory unit state i.e.

matrix format. Calculate the quantity of rounds supported the key Size and expand the key victimization our key schedule. And their area unit (n-1) rounds performed that area unit substitute computer memory unit, shift rows, combine columns and add spherical key. the ultimate spherical "n" doesn't incorporates combine column within the iteration. Figure three shows the flow of formula.

VI. CONCLUSION

This analysis has highlighted the unusefulness of the normal knowledge masking techniques in terms of knowledge utility at intervals the bismuth analytic platform, taking into thought the importance of knowledge privacy. Also, this analysis made public the most modules of the planned intrinsic knowledge masking framework (IMETU – establish, Map, Execute, Test, and Utilize). moreover, this analysis has targeted on the "Execute" and "Test" of a replacement planned knowledge masking technique (COBAD) supported the applied mathematics content derived from the loaded dataset at aggregate levels. The strength of victimization AES technique depends on the subsequent factors: • The encapsulation method of constructing a public key, supported a sequence of the applied mathematics variables that area unit being chosen, supported the complexness of the de-identification calculation, depends on the scale of the generated public key (Simple 128-bit, Medium 192-bit, onerous 256-bit). • employing a pre-stored personal key (128-bit) within the masking formula can increase the protection of the sensitive knowledge attributes within the DW.

REFERENCES

[1] Gartner Group, Feb 2012, http://www.gartner.com/technology/reprints.do?id=1196VVFJ&ct=120207&st=sb.
[2] O. Ali, A. B. Nassif, L. F. Capretz, "Business Intelligence Solutions in Healthcare A Case Study: Transforming OLTP system to BI Solution", in 23rd IEEE International Conference on Tools with Artificial Intelligence, Florida, USA, 2011, pp. 393-398.
[3] R. Guro. Components of business intelligence. The Business Intelligence Guy [Online]. 2011. Available: http://www.the-businessintelligenceguy.com/components-of-business-intelligence-bi/ .
[4] "Understanding and Selecting Data Masking Solutions: Creating Secure and Useful Data Securosis", L.L.C. 515 E. Carefree Blvd. Suite #766 Phoenix, AZ 85085 T 602-412-3051 info@securosis.com ,August 10,2012 - www.securosis.com
[5] Ravikumar G K1 ,Manjunath T N2, Ravindra S Hegadi3,Umesh I M4 "A Survey on Recent Trends, Process and Development in Data Masking for Testing" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011 ISSN (Online): 1694-0814 www.IJCSI.org
[6] AMBROSIA, V. G., BUECHEL, S. W., BRASS, J. A., and PETERSON, J. R., 1998, An integration of remote sensing, GIS, and information distribution.
[7] O. Ali, A. Ouda, " A Classification Module in Data Masking Framework for Business Intelligence Platform in Healthcare", IEEE 7th Annual Conference (IEMCON), Vancouver, Canada, 2016.
[8] Text of the regulation, "Personal Health Information Protection Act", Ontario Government, CHAPTER 3 Schedule A, e-laws Ontario Government, http://www.e-laws.gov.on.ca , 2004.
[9] Canadian Institute for Health Information (CIHI), "Discharge Abstract Database (DAD) Metadata / Data Elements", https://www.cihi.ca/en/types-of-care/hospital-care/acute-care/dadmetadata
[10] Training materials, "IntelliHEALTH – Inpatient Discharge User Guide", Ontario Ministry of Health and Long – Term Care, Version 1.0, September 2010.
[11] https://study.com/academy/lesson/database-encryption-techniques-application.html.
[12] https://en.m.wikipedia.org/wiki/Health_Insurance_Portability_and_Accountability_Act.

