IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 12, Issue 3 (Jul. - Aug. 2013), PP 44-51
www.iosrjournals.org

Enabling Use of Dynamic Anonymization for Enhanced Security in Cloud

Ms. Swati Ganar[*1], Apeksha Sakhare[*2]
Department of Computer Science & Engineeing
G.H.Raisoni College of Engineering, Nagpur

***Abstract**: Cloud computing is a model that enables Convenient and On-demand network access to a shared pool of configurable computing resources where millions of users share an infrastructure. Privacy and Security are significant obstacle that is preventing the extensive adoption of the public cloud in the Industry. Researchers have developed privacy models such as k-anonymity, l-diversity, t-closeness. However, even though these privacy models are applied, an attacker may still be able to access some confidential data if same sensitive labels are used by a group of nodes. Publishing data about individuals without revealing sensitive information about them is an important problem. Data Anonymization is a method that makes data worthless to anyone except the owner of the data. It is one of the methods for transforming the data that it prevents identification of key information from an unauthorized person. We survey the existing methods of anonymization to protect sensitive information stored in cloud. Data can also be anonymized by using techniques such as, Hashing, Hiding, Permutation, Shifting, Truncation, Prefix-preserving, Enumeration, etc. We have implemented these methods also to see an anonymization effect and implemented a new method for anonymization.*
***Keywords*:** *Anonymization, Deanonymization, Data Hiding, Hash calculation, Data Shifting, Data Truncation, Data Enumeration, Data Permutation, IP prefix Preserving.*

Submitted Date 10 June 2013 **Accepted Date: 15 June 2013**

I. INTRODUCTION

Cloud computing is a model that enables Convenient and On-demand network access to a shared pool of configurable computing resources where millions of users share an infrastructure. It offers many potential benefits to small and medium-sized enterprises (SMEs). It provides many services for

- data processing
- storage and backup
- facilitate productivity
- accounting services
- communications
- Customer service and support.

Cloud computing is immune to security breaches, because it does not facilitate backup media, unsecured connection to hijack or eavesdrop.

But, the question of privacy or confidentiality arises whenever a user shares information in the cloud. Public or Private organizations publish their database on to the cloud for

Research purpose or some other purpose. This database contains sensitive information about many people. It is an information resource for research, analysis purpose. This database may help the Hospital to track its patients, a School to monitor its students or a Bank its customers. The privacy of this data must be preserved while disclosing it to third party or while placing it in long time storage. i.e. any sensitive information should not be disclosed. To reduce or eliminate the privacy risk, a method called Anonymization is used.

Anonymization is one of the privacy preserving techniques that manipulate the information, making the data identification difficult to anybody except the owners [1]. It is different from that of data encryption. Anonymization of data removes identifying attributes like names or social security numbers from the database. For example, the school will delete student ID and Bank will remove account number.

Anonymization has 3 primary goals [2]:

- To protect identities of specific user from being leaked
- To protect identities internal user from being revealed
- To protect specific security practices of organizations from being revealed.

Experts have developed different anonymization techniques, varying in their cost, complexity, ease of use, and robustness to achieve these goals [2][3]. Suppression [4] is very common method for anonymization. It is performed by deleting or omitting the data entirely. For example, an administrator in hospital tracking prescriptions will suppress patient's names before sharing data. In order to protect the sensitive values, Generalization [4] techniques can also be used. This technique replaces quasi identifier attributes with less specific values. It divides the tuples into quasi identifier groups (QI groups), and generalise values in every group to uniform format. For example, the data in microdata table is generalized using K-Anonymization technique. To effectively limit information disclosure, it is necessary to measure the disclosure risk of anonymized table.
Different techniques are required to anonymize qualitative and quantitive information. Some methods are as follows:

- Removing individual's name from document
- Blurring images to disguise face
- Modifying or re-recording audio files
- Modification in reports

A simple example of data anonymization is given below: the aim is to find turnover of some companies, whose names are kept secret. For this purpose, name of companies are changed in cloud based data. At the same time, some fictitious information is also added to cloud based data. Then a secure mapping table is generated to identify original and fictitious data. When the total turnover is calculated in cloud, the result achieved is incorrect. This incorrect result is then corrected by using secure mapping table [1].
The Anonymization procedure can be reversed and termed as Reidentification or Deanonymization. An adversary links the anonymized records to outside data, and tries to reidentify anonymized data.
Re-identification can be done in 2 ways:

- Adversary takes personal data and searches an anonymized dataset for a match.
- Adversary takes a record from an anonymized dataset and searches for a match in publicly available information.

The rest of this paper is organized as follows: Definitions are given in Section II. Section III reviews the related work, Section IV contains Problem definition. Anonymization techniques are presented in section V. Section VI consists of Methodology and Results. Finally, Section VII concludes the paper and gives suggestion for future work.

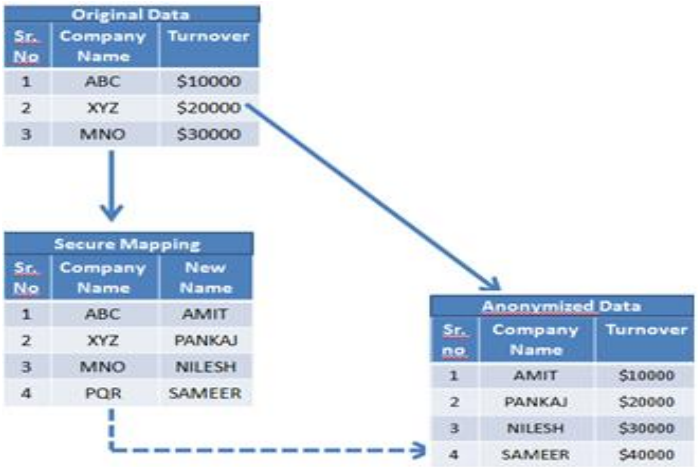

Fig.1:Data Anonymization in Cloud

II. DEFINITIONS

The database also called microdata is stored in a table which has multiple records. These records may be categorized as follows:

- Explicit identifiers
- Quasi identifiers
- Sensitive identifiers.

Explicit identifiers are the attributes which identifies an individual. For eg: Name, social security number etc.
Quasi identifiers are the attributes which can be linked with other information to identify an individual from population. For eg: gender, birth-date, zip code, diagnosis, etc.
Sensitive identifier is the attribute with sensitive value. Here the value of the attribute is not discovered to any individual.

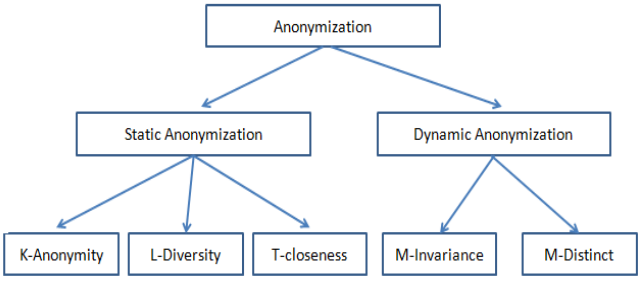

Fig. 2: Anonymization techniques

Above hierarchy in fig.2 shows the different techniques that are used for anonymization.
The sensitive information is being protected by using variety of techniques, including Data Swapping, Data Perturbation, Data recoding, Data Suppression, Data Aggregation, Data Generalization, Sampling and Rounding.

Most of the anonymization work has been done on static datasets. But, the real datasets are dynamic. So, dynamic anonymization is required. The dynamic datasets are made complex by using data updates. Above hierarchy shows the different techniques that are used for anonymization.
Data updates can be either external or internal [5]. External update leads to update of records in dataset. Whereas internal update leads to update of records attribute value.
There is always a correlation between old value and new value of a record. For example, a person's current salary in one particular organization is 4.5 lakhs per annum. After several years, even if we cannot determine her/his highest salary without complementary knowledge, we can conclude that it will not be lower than 4.5 lakhs per annum and will be one of {6 lakhs, 8 lakhs, more than 8 lakhs} with different nonzero probabilities.

III. RELATED WORK

Many techniques are available to anonymize the data. Some security models were also be used to improve data anonymization, such as k-anonymity, l-diversity, t-closeness etc.

Samarati [6] and Sweeney [7] introduced k-anonymity as the property that each record is indistinguishable from a defined number (k) if attempts are made to identify the data. For any data record with a set of attribute values, if there are at least k-1 other records that match those attribute values then, the dataset is said to be k-anonymized. K-anonymity can prevent only identity disclosure; it cannot prevent disclosure of attribute information.

Machanavajjhala et al. [8] introduced a new model, called l-diversity, which requires that there are 'l' different sensitive values for each combination of quasi identifiers. An equivalence class is said to have l-diversity if there are at least l "well-represented" values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity.

Similar to k-anonymity, l-diversity does not prevent attribute disclosure. And there are some attack that may occur on l-diversity such as, Skewness attack and Similarity attack. The information leakage occurs in l-diversity because it does not consider semantical closeness of sensitive values.

Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian[5][9] proposed a new privacy model known as *t*-closeness which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e. the distance between the two distributions should be no more than a threshold *t*). t-closeness uses Earth Mover Distance (EMD) to calculate the distance between two distributions [2]. And it also considers semantic closeness of attribute values. EMD can be calculated by using the solution of transportation problem. t-closeness prevents attribute disclosure but it cannot protect the dataset against identity disclosure. In [5], authors have used the Mondrian algorithm in which high-dimensional space is divided into regions and data points are encoded in one particular region by the region's representation.
To protect privacy of the database, some other techniques were utilized. These techniques are given below [10]:

- **Removing identifying information**: This is the simplest method of anonymization. Here, the field which is used to identify a specific individual is removed. The field to be removed may be Name, ID or some other field that is highly sensitive in context of data [11]. Depending on the context, which field is to be

removed is decided. For example, Patient name is removed from Hospital database because this field gives identification of many persons. Fig.3 displays an original Hospital database. Whereas fig.4 gives an anonymized version.

ID	QID			SA
Patient Name	Gender	Age	Zip code	Health Problem
Amit	Male	35	400071	Viral Infection
Pankaj	Male	37	400182	Viral Infection
Vishal	Male	39	400095	Heart problem
Sheetal	Female	54	440672	Flu
Pallavi	Female	58	440123	Heart problem
Nilesh	Male	54	440893	Viral Infection
Sagar	Male	41	400022	Flu
Mahesh	Male	46	400135	Flu
Sujata	Female	44	400182	Flu

Fig.3 : Original database

Removing a particular record from a database also gives good protection for sensitive data.

QID			SA
Gender	Age	Zip code	Health Problem
Male	35	400071	Viral Infection
Male	37	400182	Viral Infection
Male	39	400095	Heart problem
Female	54	440672	Flu
Female	58	440123	Heart problem
Male	54	440893	Viral Infection
Male	41	400022	Flu
Male	46	400135	Flu
Female	44	400182	Flu

Fig.4 : Removing Id field

- **Suppression**: Suppression consists of replacing value of variables with missing value. Or removing the fields. The aim of this method is to reduce the information content. In [12], depending on the violation of sensitive attribute, four different Suppression schemes have been suggested as follows:
  - Delete all violating sensitive values and replace them with unknown value
  - Delete all sensitive values
  - Delete minimum number of records which violate sensitive value
  - Delete all records

Suppression of 3 columns is shown in fig.5.

QID	SA
Gender	Health Problem
Male	Viral Infection
Male	Viral Infection
Male	Heart problem
Female	Flu
Female	Heart problem
Male	Viral Infection
Male	Flu
Male	Flu
Female	Flu

Fig.5: Suppressing 3 fields

- **Generalization**: This technique replaces quasi identifier attributes with less specific values. The rational of generalization is to partition the tuple into several Quasi-identifier fields, and generalize them in a uniform format. Privacy of data is preserved in generalized table iff it satisfies Generalization principle. For example, Birth date may be generalized to year of birth only. Fig.6 shows generalization of Zip code.

ID	QID			SA
Patient Name	Gender	Age	Zip code	Health Problem
Amit	Male	35	400*	Viral Infection
Pankaj	Male	37	400*	Viral Infection
Vishal	Male	39	400*	Heart problem
Sheetal	Female	54	440*	Flu
Pallavi	Female	58	440*	Heart problem
Nilesh	Male	54	440*	Viral Infection
Sagar	Male	41	400*	Flu
Mahesh	Male	46	400*	Flu
Sujata	Female	44	400*	Flu

Fig.6: Generalized database (Zip field)

- **Aggregation**: This method gives aggregate statistics of database or field. Only summary statistics is given using aggregation. For example, it is possible to know that how many persons are suffering from flu using aggregation. Fig.7 gives an aggregated value from database.

Health Problem	Number of patients
Flu	4

Fig.7: Aggregated database

IV. PROBLEM DEFINITION

The data is often not protected when used. Regardless of where the data is stored or transferred, Data security is important. The data may be a raw data which contains more sensitive information. Most of the cloud users are worried since the cloud service provider may sell their data to another provider. In that case, some sensitive data of cloud users may be disclosed. Similarly, cloud service providers may use the data (images) for advertisement.

Given an anonymization, adding a new value performs s*pecialization on* the data, whereas removing some value performs *generalization*. Here, the aim is not just to anonymize the data, but to achieve a good anonymization with respect to its cost, complexity and robustness. To achieve this aim, we have implemented following techniques of anonymization and also proposed new technique for anonymization.

V. ANONYMIZATION TECHNIQUES

Different vulnerabilities are associated with different types of anonymizations. There are several techniques available to anonymize the data, such as encryption, substitution, shuffling, number and date variance and nulling some fields. We have implemented some anonymization techniques to obscure data in database.

1. DATA HIDING:

It suppresses a data value by replacing it with a value '0'. It is also called as Black marker anonymization. The advantage of hiding a record is that number of records is maintained after anonymization also. For example, while considering hospital database, an age of a patient may not be required for processing, so it is replaced with constant '0'.

2. Hash Calculation:

It finds a hash value of either one field or several fields. It takes a variable input and produces fixed size hash of input. The MD5 or SHA can be used. For example, hash of first name and last name can be calculated. The hash function is defined as,

$H(S) = H(\text{“}S1S2\ldots. Sn\text{”}) = S1 + pS2 + p^2S3 + \ldots. + p^{n-1}Sn$

$H(T) = H(\text{“}T1T2\ldots. Tn\text{”}) = T1 + pT2 + p^2T3 + \ldots. + p^{n-1}Tn$

Where,

S=first name , T=last name

P=prime number used for multiplication

3. Shifting:

Shifting shifts a field or data value by specific value. It adds some offset to data value. Shift value is the only key to shift function, so it is kept secret. For example, an offset value 10 is added in age field.

$$\int_{-\infty}^{\infty} \delta(t-A)f(t)dt = f(A)$$

Where t=10(shift amount for Age field).

4. Data Truncation:

It removes 'n' least significant bits from the numerical field [13]. Even if data at the end is lost, it preserves the information. For example, the telephone number of doctor is truncated, and only first 3 digits are displayed as shown below.

Decimal d1=new Decimal (07167485938, 0, 0, true, 8))

This formula will display only 3 digits i.e. 071(truncated from telephone number of doctor).

5. Data Permutation:

Permutation is a substitution technique. It replaces the original value by a new unique value. The selection of substitution value is random. These functions may result in noncollision. The formula for permutation is,

P(n,r)= nPr

In our case, a combination of first name and last name are permuted.

6. Data Enumeration:

Enumeration is also a substitution technique. It retains the chronological order in which events takes place. It is useful for applications demanding strict sequencing order. For example, salary field is enumerated while maintaining the order of execution.

$$f(x) := \begin{cases} -(x+1)/2, & \text{if } x \text{ is odd} \\ x/2, & \text{if } x \text{ is even.} \end{cases}$$

Here, x = Salary field

7. Ip Prefix-Preserving:

Since IP addresses are unique, it is possible to identify a person, an organization or a host. Therefore IP address anonymization is necessary. This method preserves the n-bit prefix on IP-address. Two anonymized IP addresses match on prefix of n-bits, if two real IP addresses match on prefix of n-bits. i.e. They share n-bit prefix if $a_1a_2....a_n=b_1b_2....b_n$ and $a_{n+1} \neq b_{n+1}$. The IP address is prefix preserved here.

Given $a = a_1a_2 \ldots a_n$, *let*

$F(a) := a_1{'}a_2{'} \ldots a_n{'}$

where $a_i{'}=a_i \oplus f_i{-}1(a_1, a_2, \ldots, a_i{-}1)$

And $\oplus$ *is a XOR operation for i=1,2,3...n.*

Here, F is prefix preserving anonymization function. It is a one-to-one function from $(0,1)^n$ to$(0,1)^n$

Prefix-preserving anonymization belongs to Typed Transformation, which uses single anonymized value for each unique value of original data [14]. The tool TCPdPriv uses prefix preservation anonymization.

CryptoPAN is an approach developed by Fan *et al.* for creating prefix preserving anonymized addresses without using prefix table [15].

VI. METHODOLOGY & RESULTS

Since most of the methods specified above has some drawbacks given as follows in table 1, there is a need to implement some new methods to prevent the security breach.

Anonymization Technique	Original Data	Anonymized Data	Drawbacks
Data hiding	12	0	Reduces utility of dataset
Data Shifting	9370207875	9370208874	May reverse the anonymization process
Data Truncation	0712345345	071	Results in loss of Information
Data Permutation	Swati Ganar	Reetu ganar Praful ganar	Number of original records are not preserved
Data Enumeration	12	36	Strictly used where order of execution is necessary

Table 1: Results of Anonymization techniques

Our method works on numerical attributes. The method calculates 'MOD n' of the numerical field and then displays the anonymized dataset. To find a 'MOD n' of a number, a minimum divisor 'n' is taken from the dataset. For example, to anonymize Age field from Hospital dataset, MOD will be calculated as below:

Age MOD n

65 MOD 20=5

Now, instead of 65, 5 will be stored in dataset, and thus Age field will be anonymized. Each time a new entry is stored in dataset, 'MOD n' is calculated from the Age field, based on minimum age of a Person. The result of this anonymization is shown below:

ADMIN VIEW

First Name	Age	Gender	Doctor Name
chandrashekhar	12	Male	Dr.kale
prafull	44	Male	kale
swati	25	Female	Rane
ritu	25	Female	Das
ritu	10	Female	fghfgh
punam	25	Female	Dr.pandey

HACKER VIEW

First Name	Age	Gender	Doctor Name
chandrashekhar	2	Male	Dr.kale
prafull	4	Male	kale
swati	5	Female	Rane
ritu	5	Female	Das
ritu	0	Female	fghfgh
punam	5	Female	Dr.pandey

Fig.8 Dynamic MOD

An attempt to make an Anonymization process dynamic has also be done as shown below.

ADMIN VIEW

First Name	Surname	Age	Gender	Doctor Name
chandrashekhar	Tiwari	12	Male	Dr.kale
prafull	nathille	44	Male	kale
swati	ganar	25	Female	Rane
ritu	gupta	25	Female	Das
ritu	detgtyr	10	Female	fghfgh
punam	marbate	25	Female	Dr.pandey

Complete Name	Age	Gender	Doctor Name
DzIr8c1eE6cWEfJXEahhg	0	R4k/ON3E0SlWbCR/9nCFg	R4k/ON3E0SlWbCR/9nCFg
Kr4jhUgIUUu00PKJ82e0RA	0	//c2rTL/UOO31pfJS0i0A	//c2rTL/UOO31pfJS0i0A
O/tOWNfAtUKapRqabXEl6w	0	eaPLp5CK8Ea55sQeMri5A	eaPLp5CK8Ea55sQeMri5A
k0JJnd1C0OVmfm09UVKoQ	0	vaeFHLswwkqL2S1qYD7ZZw	vaeFHLswwkqL2S1qYD7ZZw
GGqHcvO1yE6M1ZU20HgWQ	0	7hg8s4o/rUWHnBFCXbFAqw	7hg8s4o/rUWHnBFCXbFAqw
YLTqgB62kKlpDY4Q8mQTw	0	hu2euxLOk06mcI7MHsXr9g	hu2euxLOk06mcI7MHsXr9g

Fig.9 Dynamic Anonymization

Dynamic Anonymization uses setInterval() method in which the data is anonymized after each 10 seconds. This prevents an adversary from deanonymizing the dataset.

VII. CONCLUSION AND FUTURE WORK

In spite of the safeguards in place, Cloud computing faces privacy and security concerns. Cloud computing requires standard methodologies and technical solutions to assess privacy risks and establish adequate protection levels. A strong protection should be ensured by organizations, agencies for private data irrespective of the environment where the data is actually stored. Because loss of this sensitive data may create a negative impact for organizations.

Anonymization is a viable technique to secure cloud computing. It limits the misuse of sensitive data, but is not a complete solution to preserve confidentiality. In this paper, we surveyed few anonymization methods and implemented some techniques of anonymization to protect sensitive data in cloud. Formal models of security for anonymization are also discussed. Lots of techniques for anonymization have been implemented, but still there is a fear of security breach. Research for anonymization and deanonymization is in process. The techniques which are currently safe for anonymization may fail in future. Still data anonymization is a viable solution that is highly recommended for security in cloud. So, the available techniques of anonymization may be integrated to achieve better results. In future, the privacy preserving in cloud needs many efforts.

REFERENCES

[1] Jeff sedayao, "Enhancing cloud security using Data Anonymization", Intel white paper, June 2012

[2] R. Pang, M. Allman, V. Paxson, and J. Lee, "The Devil and Packet Trace Anonymization" ACM Computer Communication Review, 36(1):29–38, January 2006.

[3] A. Slagell and W. Yurcik. Sharing Computer Network Logs for Security and Privacy: "A Motivation for New Methodologies of Anonymization. In Proceedings of SECOVAL": The Workshop on the Value of Security through Collaboration, pages 80–89, September 2005

[4] Latanya Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression", 10 int'l j. On uncertainty, fuzziness & knowledge-based sys. 571, 572, 2002

[5] Information Commissioner's office, "Anonymization: managing data protection risk , code of practice", 2012

[6] P. Samarati, "Protecting Respondent's Privacy in Microdata Release," I6EE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.

[7] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.
[8] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in ICDE, 2006, p. 24.
[9] Ninghui Li Tiancheng Li, Suresh Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity", 2007.
[10] Paul Ohm*, "Broken promises of privacy: responding To the surprising failure of anonymization", 57 ucla law review 1701, 2010
[11] Ninghui Li, Member, IEEE, Tiancheng Li, and Suresh Venkatasubramanian, "Closeness: A New Privacy Measure for Data Publishing" ,IEEE transactions on knowledge and data engineering, vol. 22, no. 7, July 2010
[12] Junqiang Liu, Ke Wang, "On Optimal Anonymization for *l*+-Diversity", Data Engineering (ICDE), 2010 IEEE,pp. 213-224
[13] E. Boschi,Internet-Draft, B. Trammell, "IP Flow Anonymization Support, draft-ietf-ipfix-anon-06.txt", 2011
[14] Scott E. Coull*et al., "Playing Devil's Advocate*: Inferring Sensitive Information from Anonymized Network Traces", NDSS,2007
[15] J. Fan, J. Xu, M. H. Ammar, and S. B. Moon. Prefix preserving IP Address Anonymization: Measurementbased Security Evaluation and a New Cryptography-based Scheme. *Computer Networks*, 46(2):253–272, 2004.