Anonymization: A Method To Protect Sensitive Data In Cloud

Asst.Prof.Ms. Apeksha Sakhare

Department of Computer Science and Engineering
G.H.Raisoni College of Engineering, Nagpur
apeksha.sakhare@raisoni.net

Ms. Swati Ganar

Department of Computer Science and Engineering
G.H.Raisoni College of Engineering, Nagpur
swati_gnr@rediffmail.com

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. Data can also be anonymized by using techniques such as, Hashing, Hiding, Permutation, Shifting, Truncation, Prefix-preserving, Enumeration, etc. We survey the existing methods of anonymization to protect sensitive information stored in cloud.

***Keywords*— Anonymization, Deanonymization**

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.

Various anonymization techniques are used to achieve these goals[2][3].

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. And sensitive identifier is the attribute with sensitive value. Here the value of the attribute is not discovered to any individual.

Experts have developed different anonymization techniques, varying in their cost, complexity, ease of use, and robustness. 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: section II contains related work, Anonymization techniques are given in section III, section IV concludes with future work.

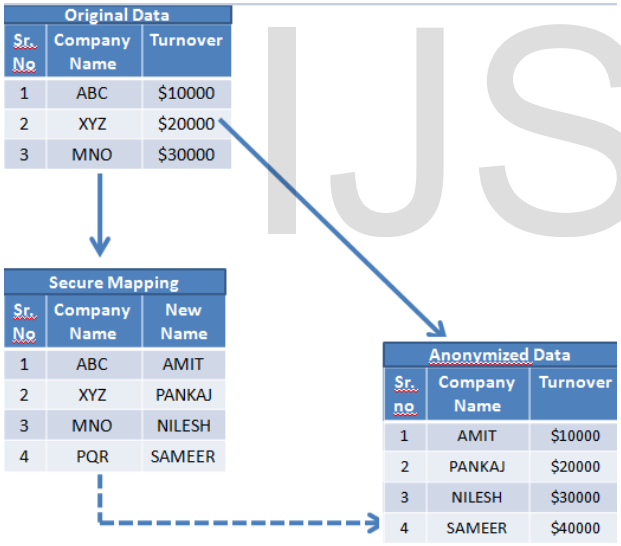

Fig.1:Data Anonymization in Cloud

II. 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 atleast k-1 other records that match those attribute values then, the dataset is said to be k-anonymzed. 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[9][10] 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.

To protect privacy of the database, some other techniques were utilized. These techniques are given below[11]:

- Removing identifying information: Here, the field which is used to identify a specific individual is removed. For example, Patient name is removed from Hospital database.

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.2 : Original database

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.3 : 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.

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.4: Suppressing 3 fields

- Generalization: This technique replaces quasi identifier attributes with less specific values. For example, Birth date may be generalized to year of birth only.

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.5: Generalized database (Zip field)

- Aggregation: This method gives aggregate statistics of database or field. For example, it is possible to know that how many persons are suffering from flu using aggregation.

Health Problem	Number of patients
Flu	4

Fig.6: Aggregated database

III. 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 discussed 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. 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.

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 that is kept secret. For example, an offset value 10 is added in age field.

4. DATA TRUNCATION:

It removes 'n' least significant bits from the numerical field. 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.

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. For example, 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.

7. IP PREFIX-PRESERVING:

This method preserves the n-bit prefix on IP-address. Two anonymized IP addresses match on prefix of n-bits, if and only if two real IP addresses match on prefix of n-bits. The IP address is prefix preserved here.

Prefix-preserving anonymization belongs to Typed Transformation, which uses single anonymized value for each unique value of original data[12]. 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[13].

IV. CONCLUSION AND FUTURE WORK

Inspite 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. In this paper, we surveyed few anonymization methods to protect sensitive data in cloud. Formal models of security for anonymization are also discussed. 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. 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. In future, the privacy preserving in cloud needs many efforts.

V. 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] 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
[11] Paul Ohm*, "Broken promises of privacy: responding To the surprising failure of anonymization", 57 ucla law review 1701, 2010
[12] Scott E. Coull*et al., "Playing Devil's Advocate*: Inferring Sensitive Information from Anonymized Network Traces", NDSS,2007
[13] 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.
[14] E. Boschi,Internet-Draft, B. Trammell, "IP Flow Anonymization Support, draft-ietf-ipfix-anon-06.txt", 2011