Survey on Cloud security Using Data Anonymization Techniques

Anuja Phapale
AISSMS, Pune, India

Abstract- *Nowadays, cloud computing is most popular and modern technology of storing the large amount of information on the internet and accessing it from anywhere. Costs reduction, universal access, availability of number of applications and flexibility is a number of reasons for popularity of cloud computing. Users store sensitive information on cloud, providing security becomes important aspect. Data privacy is one of the most concerned issues in big data applications, because processing large-scale sensitive data sets often requires computation power provided by public cloud services. Data anonymization is a type of information sanitization whose intent is privacy protection. It is the process of either encrypting or removing personally identifiable information from data sets, so that the people whom the data describe remain anonymous.*

Keywords- Anonymization, Deanonymization, privacy, encryption, hashing, generalization

I. INTRODUCTION

In the current world, many organizations work on big data which may require huge storage. But these organizations cannot afford to create their own storage servers as it will be too expensive. The cloud computing provides services to such organization which are really fast and at low and affordable cost. Thus, cloud computing is a way to increase the capacity or add capabilities dynamically without investing cost on new infrastructure, for training new personnel, or licensing new software. Thus cloud computing extends Information Technology's existing capabilities. Cloud computing provides enormous computation power and storage capacity via utilizing a large number of commodity computers together, enabling users to deploy applications cost-effectively without heavy infrastructure investment. Cloud users can reduce huge upfront investment of IT infrastructure, and concentrate on their own core business. But as more and more information of individuals and organizations are placed on the cloud, concerns are beginning to grow about security and privacy of stored information.

Anonymization is a technique that can use to increase the security of data in the public cloud while still allowing the data to be analyzed and used. Data anonymization is the process of changing data that will be used or published in a way that prevents the identification of key information. Using data anonymization, key pieces of confidential data are hidden in a way that maintains data privacy. The data can still be processed to gain useful information. Anonymized data can be stored in a cloud and processed without concern that other individuals may capture the data. Later, the results can be collected and mapped to the original data in a secure area.

Importance of designing techniques to effectively anonymize data so that detailed results can be published and shared with others. The aim is that a malicious party should be unable to use this published data to infer anything private about the entities represented, while an honest party should still be able to perform a variety of ad hoc analyses and find results which are close to their true values on the original data.

II. DATA ANONYMIZATION AND ENCRYPTION

Data anonymization is the process of transforming data so that it can be processed in useful way, while preventing that data from being linked to individual identities. So it is to protect the privacy of the individual and to make it legal for governments and businesses to share their data without getting permission. Data anonymization methods include encryption, hashing, generalization, pseudonymization and perturbation.

Encryption involves transforming data to render it unreadable to those who don't have the key to decrypt it. Thus encryption can be useful tool for doing anonymization, particularly when hiding indentifying information in a set of data.

III. SYSTEM FEATURES

Anonymization is process of removing or modifying the identifying variables contained in the microdata dataset. Typically an identifying variable is one that describes a characteristic of a person that is observable, that is registered (identification numbers, etc.), or generally, that can be known to other persons. Identifying variables are those describe characteristic of a person.

- Direct identifiers, which are variables such as names, addresses, or identity card numbers.

- They permit direct identification of a respondent but are not needed for statistical or research purposes, and should thus be removed from the published dataset.
- Indirect identifiers, which are characteristics that may be shared by several respondents, and whose combination could lead to the re-identification of one of them. For example, the combination of variables such as district of residence, age, sex, and profession would be identifying if only one individual of that particular sex, age and profession lived in that particular district. Such variables are needed for statistical purposes, and should thus not be removed from the published data files.

Anonymizing the data will consist in determining which variables are potential identifiers and in modifying the level of precision of these variables to reduce the risk of re-identification to an acceptable level. The challenge is to maximize the security while minimizing the resulting information loss.

A. Data Reduction

Removing variables: The first obvious application of this method is the removal of direct identifiers from the datasets (race, religion, HIV, etc.)

Removing records: Removing records can be adopted as an extreme measure of data protection when the unit is identifiable in spite of the application of other protection techniques.

Global recoding: The global recoding method consists in aggregating the values observed in a variable into pre-defined classes.

Top and bottom coding: A special case of global recoding that can be applied to numerical or ordinal categorical variables. The variables "Salary" and "Age" are two typical examples. The highest values of these variables are usually very rare and therefore identifiable.

Local suppression: Local suppression consists in replacing the observed value of one or more variables in a certain record with a missing value. The result is an increase in the frequency count of records containing the same (modified) combination. A criterion is therefore necessary to decide which variable in the risky combinations has to be locally suppressed.

B. Data perturbation

Micro aggregation: Replace an observed value with the average computed on a small group of units (small aggregate or micro-aggregate), including the investigated one. The units belonging to the same group will be represented in the released file by the same value. Methods in micro-aggregation include individual ranking and multivariate micro aggregation.

When micro aggregation is independently applied to a set of variables, the method is called individual ranking. When all the variables are averaged at the same time for each group, the method is called multivariate micro aggregation. The easiest way to group records before aggregating them is to sort the units according to their similarity and the values resulting from this criterion, and to aggregate consecutive units into fixed size groups.

Data swapping: a perturbation technique for categorical micro data, and aimed at protecting tabulation stemming from the perturbed micro data file. Data swapping consists in altering a proportion of the records in a file by swapping values of a subset of variables between selected pairs of records (swap pairs).

Post-randomization (PRAM): induces uncertainty in the values of some variables by exchanging them according to a probabilistic mechanism. PRAM can therefore be considered as a randomized version of data swapping. As with data swapping, data protection is achieved because an intruder cannot be confident whether a certain released value is true, and therefore matching the record with external identifiers can "easily" lead to mismatch or attribute misclassification. The method has been introduced for categorical variable but it can be generalized to numerical variables as well.

Adding noise: Adding noise consists in adding a random value ε, with zero mean and predefined variance $\sigma 2$, to all values in the variable to be protected. Generally, methods based on adding noise are not considered very effective in terms of data protection.

Resampling: a protection method for numerical microdata that consists in drawing with replacement t samples of n values from the original data, sorting the sample and averaging the sampled values. Data protection level guaranteed by this procedure is generally considered quite low.

C. K-anonymity

K-anonymity has popularly used for data anonymization [1]. It is an effective way to Anonymized microdata. In a k- Anonymized dataset, each record is indistinguishable from at least k - 1 other record with respect

to certain "identifying" attributes. There are two common methods for achieving k-anonymity for some value of k. K-anonymity does not provide an efficient investigation for the multiple queries.

D. l-diversity

K-anonymity is vulnerable to attacks like homogeneity attack and background knowledge attack. Machanavajjhala et al. [2] introduced a new notion of privacy, called l-diversity, which requires that the distribution of a sensitive attribute in each equivalence class has at least "l-well represented" values. One problem with l-diversity is that it is limited in its assumption of adversarial knowledge. It is possible for an adversary to gain information about a sensitive attribute as long as information about the global distribution of this attribute. This assumption generalizes the specific background and homogeneity attacks used to motivate l-diversity.

IV. CONCLUSION AND FUTURE WORK

Due to the accelerating integration of computer and communication technology, internet has been established worldwide, and thus brings about various commercial services. Thereby, to transmit sensitive data is a great security concern. So, Cloud computing security is one of the major issues in the cloud computing environment as user does not want to lose their private/sensitive data stored on the cloud. This paper survey formal anonymization techniques and traditional encryption method used for securing data which is shared on cloud. As there are number of anonymization techniques and research is in process, there is fear of security concern. In future to provide more security by implementing re-encryption based protocol, even though encryption and decryption is time consuming process and requires more computation.

REFERENCES

[1] Padma L. Gaikwad 1, M. M. Naoghare, "Data Anonymization Approach for Data Privacy" International Journal of Science and Research (IJSR), December 2015.

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24, 2006.

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

[4] Bin Zhou, Jian Pei, Wo Shun Luk, "A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data" August 20, 2007.

[5] Graham Cormode, Divesh Srivastava," Anonymized Data: Generation, Models, Usage".

[6] Saranya M, Senthamil Selvi R," Data Anonymization Approach For Privacy Preserving In Cloud", International Journal of Computer Science & Engineering Technology (IJCSET), Apr 2015.

[7] Benjamin C.M. Fung, Ke Wang, and Philip S. Yu, Fellow, "Anonymizing Classification Data for Privacy Preservation", IEEE Transactions On Knowledge And Data Engineering, Vol. 19, No. 5, May 2007.

[8] Ashwin, Machanavajjhala, Johannes Gehrke, Daniel Kifer, "ℓ-Diversity: Privacy Beyond k-Anonymity".

[9] Ms. Apeksha Sakhare, Ms. Swati Ganar," Anonymization: A Method To Protect Sensitive Data In Cloud", International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013.

[10] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity".