IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10799

Data Anonymization Techniques

Miss. Shreya Mugal[1], Prof. K. K. Chhajed[2]

PG Scholar, CSE Department, P R Pote college of engineering and management , Amravati, India[1]

Assitant professor, CSE Department, P R Pote college of engineering and management , Amravati, India[2]

Abstract: Cloud computing is a subscription-based service to obtain network storage space and computer resources. Cloud provide users to store large volume of data and to perform application over cloud also provides greater flexibility of storing and computation of data but, Such applications can be processed, huge volume processing data sets are to be generated. For Storing some valuable intermediate datasets has been considered in order to avoid the high recomposing them. In this paper, the concepts involved in data anonymization will be addressed, some of the techniques used for this anonymization will be studied, which are the risks of re-identification associated and the analysys of some software tools that allows to perform these techniques.

Keywords: Cloud, Anonymiization, ARX.

INTRODUCTION

Cloud computing is a subscription-based service to obtain network storage space and computer resources. The Cloud helps to access the information from anywhere at any time but Internet connection is necessary to access the Cloud. e.g. Email client, If it is Yahoo!, Gmail, Hotmail, and so on, takes care of housing all of the hardware and software necessary to support client personal email account. Cloud also called as “pay-as-you-go” meaning that if technological needs change at any point to purchase more storage space from Cloud provider. It does not require a user to be in a specific place to gain access to Companies may find that cloud computing allows them to reduce the maintenance cost of data management, as they don't need to own their own servers and can use capacity leased from third parties. Cloud customers can save more capital investment of IT infrastructure, and concentrate on their own business. In accordance to it, the cloud-like structure allows companies to upgrade software more efficiently and quickly. Therefore, many companies and industries have been migrating or building their business platforms into cloud.What happens in large organizations and which has also been growing in smaller ones is that they carry out analyzes on their data in order to find patterns, trends and customer profiles But what is happening is that a lot of data is sold to other organizations or made available to the public for research purposes as a consultation service. These personal data may contain information that allows to identify the individual and being made public may violate privacy. With this, a very debated question has been how to maintain the privacy of this data without rendering it useless for analysis.

Cloud provide users to store large volume of data and to perform application over cloud also provides greater flexibility of storing and computation of data but, Such applications can be processed, huge volume processing data sets are to be generated. For Storing some valuable intermediate datasets has been considered in order to avoid the high recomposing them.

The advantages for Cloud-service clients:

- Ability to improve use by adding more capacity at peak demand
- Reducing costs
- Experimenting with new services
- Removing unneeded capacity

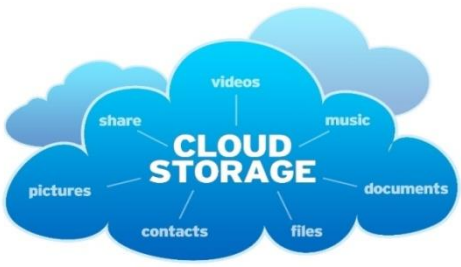

Figure 1: Cloud Storage

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10799

Cloud provide users to store large volume of data and to perform application over cloud also provides greater flexibility of storing and computation of data but, Such applications can be processed, huge volume processing data sets are to be generated. For Storing some valuable intermediate datasets has been considered in order to avoid the high recomposing them.

In this paper, the concepts involved in data anonymization will be addressed, some of the techniques used for this anonymization will be studied, which are the risks of re-identification associated and the analysys of some software tools that allows to perform these techniques. Finally we studied the Data anonymization tool namely as ARX.

I. DATA ANONYMIZATION TECHNIQUES

In order to correctly select the anonymization techniques to be useful, we must be aware of what is the purpose of this anonymization because the various techniques have various characteristics and may be more or less appropriate for certain purposes. The three most used ways to change data are to replace, modify or remove an attribute or a record. It is also important to be able to maintain the usefulness of the data and at the same time respect the privacy terms.In this section, we depict some of the data anonymization techniques and in what situations they should be applied.

A. Remove Attributes (Suppression)

In this technique, an attribute is removed from the dataset. This be supposed to happen whenever an attribute is not related or necessary for analysis or whenever it is impossible to anonymize it in any another way. In the example given in the show to Basic Data anonymization Techniques (Personal Data Protection Commission Singapore, 2018) for this technique, in which it was intended to investigation students' grades in an assessment test, the dataset was composed of three attributes: student name, trainer and grade. Figure 1 shows an example of the original dataset.

Student	Trainer	Test Score
John	Anna	93
Nicholas	Paul	86
Josh	Paul	54
Taylor	Anna	78

After suppressing the "student" attribute:

Trainer	Test Score
Anna	93
Paul	86
Paul	54
Anna	78

Figure 2 :– Original dataset. & anonymized data set

Suppression can also occur for a complete dataset record affecting several attributes. The main advantage of this technique is that, when permanently deleting an attribute or record, it becomes impossible to retrieve the information.

B. Character Replacement

The substitution of characters consists of covering up characters of an attribute or value of the data by substituting those characters with a predefined symbol (for example, by X or *). This substitution can be partial, partially hiding a text or attribute, which may be sufficient to anonymize its content. Also, in the Guide to Basic Data Anonymization Techniques (Personal Data Protection Commission Singapore, 2018), an example can be found in which to make an investigation of a dataset where the post code was identified, the last 4 digits of the post code were replaced by the character 'X'. Figure given below illustrate the original dataset and the anonymized dataset.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10799

Before anonymisation:

Postal Code	Average No. of Orders/month
100111	2
200222	8
300333	1

After suppressing the "student" attribute:

Postal Code	Average No. of Orders/month
10xxxx	2
20xxxx	8
30xxxx	1

Figure 3: Example of replacing characters.

C. **Shuffling**

In this method, the data is arbitrarily mixed or reorganized and the values of the original attributes stay in the dataset but can be linked with another record. This technique can be used when it is intended to analyse only one attribute and it is not essential to relate it to the others.

For example, if we want to analyse the amount of sales in a given region, it is only necessary to use the attribute 'region' and the transformation does not effect the results because a certain region will occur the same number of times before and after the permutation.

However, this technique does not always provide anonymization of the data and it may be probable to reorganize it to its original form. Therefore, it must be used in combination with other techniques.

D. **Generalization**

Generalization is a further approach used by Google (Google, s.d.) and consists of generalizing the attributes in order to alter the respective scale or order of magnitude.

An example of this is to replace the "date" attribute (day/month/year) with the "year" attribute, removing the day and month. Like the addition of noise, this approach may prevent the individual from being identified, but it may not result in effective anonymization.

Table 1: Comparative Analysis of Techniques

Sr. no	Technique	Advantage	Disadvantage
1	Remove Attributes (Suppression)	On permanent deleting an attribute, retrieve the information impossible	Slow in nature
2	Character Replacement	As it is partial process, its processing is quicker	As substitution is partial there may change of revealing original data
3	Shuffling	As it analyze single attribute at a time, its processing is faster	this technique does not always provide anonymization of the data
4	Generalization	It is simpler technique	It always may not result in effective anonymization.

II. DATA ANONYMIZATION TOOL: ARX

ARX is a comprehensive open source software for anonymizing sensitive personal data. It supports a wide variety of (1) privacy and risk models, (2) methods for transforming data and (3) methods for analyzing the usefulness of output data. ARX is an open source framework developed in Java (ARX, s.d.). It permits to implement a number of of the techniques described, such as K-Anonymity and LDiversity, and also to implement a set of metrics to assess the loss of information.

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 7, July 2021
DOI 10.17148/IJARCCE.2021.10799

This software has a graphical tool with a simple and intuitive interface, shown in Figure which supports the import and cleaning of data, wizards for creating transformation rules, intuitive ways to adapt the anonymized dataset to the requirements and visualizations of risks and re-identification.

III. CONCLUSION

Various approaches are available to provide data anonymization to original data set in cloud. As cloud work on principle of pay-as-you-go principle it is necessary to reduce privacy preserving cost of data on to the cloud. Anonymization technique is helpful to reduce the cost by hiding intermediate data rather than whole. There is need much more confront technique which also improve the efficiency of data processing on to the cloud.

REFERENCES

[1] Bayardo, R. J. and Agrawal, R., "Data privacy through optimal kanonymization", In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE), pp.217–228, 2005.
[2] Johny Antony P & Selvadoss Thanamani Antony., "A Survey on Privacy Preservation in Big Data", International Journal of Engineering Science Invention Research and Development (IJESIRD) Vol 3, Issue 3, October 2016, ISSN 2349-6185
[3] Johny Antony P & Selvadoss Thanamani Antony., "A Review on Privacy Preservation in Big Data", International Journal of Modern Computer Science and Application (IJMCSA) Vol. 4, Issue 6, November 2016, ISSN 2321-2632.
[4] Johny Antony P & Selvadoss Thanamani Antony., "A Privacy Preservation Framework for Big data using differential privacy and overlapped slicing", International Conference on Big data infrastructure and cloud computing, Thiruvananthapuram, 9th October 2016, ISBN 9788192958040.
[5] Li, N., Li, T., and Venkatasubramanian, S., "t-closeness: Privacy beyond kanonymity and l-diversity", In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE), 2006.
[6] M. Nithya and Dr. T. Sheela., [2014], A Comparative Study on Privacy Preserving Data Mining Techniques, International Journal of Modern Engineering Research (IJMER), Vol 4, Issue 7, July 2014, ISSN 2249
[7] Mohammed, N., Chen, R., Fung, B., & Yu, P. S., "Differentially private data release for data mining", In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 493-501, ACM, 2011

