A BRIEF SURVEY ON PRIVACY PRESERVING DATA ANONYMIZATION TECHNIQUES ON CLOUD

Shital Chintawar 1, Ismail 2

1 PG Student, Dept of Computer Engineering, Alard college of engineering and management,
2 Professor, Dept of Computer Engineering, Alard college of engineering and management, Savitribai Phule Pune University, Pune

ABSTRACT:
In big data digital age, large amount of data collected, mined and published. Cloud computing is best suitable model to support big data requirements. But the cloud has potential risk of privacy breach. So cloud service users avoid to share microdata like electronic healthcare data, data containing financial transactions which can be used for mining and data analysis. Data anonymization techniques are widely used to encounter privacy concern. In recent years, many kinds of anonymization techniques are part of research. In this paper we briefly survey different anonymization techniques and approaches. Advantages and disadvantages of these techniques are analyzed.

Keywords: Data anonymization, privacy preservation, k-anonymity, cloud computing, MapReduce

[1] INTRODUCTION

Data sharing, data publishing become day today activity for individuals, organizations and agencies. Data publishing process is so complex due to volume of data. Most of the organizations are moving towards cloud to reduce the cost. Cloud systems provides massive computation power and storage capacity that enable cloud users to deploy applications without infrastructure investment. Inspite of these benefits of cloud, users hesitate to take advantage of cloud services because of concerns related to data security, data confidentiality. Here comes the privacy preservation and data protection comes into picture.

Privacy and security are most concerned issues in cloud computing. Personal data like health records, financial transactions, data released by govt. agencies like census is extremely sensitive. If this data analyzed and mined, can offer noteworthy benefits to community. For ex. Online cloud health service like Microsoft HealthVault where people gather, store, and share health information online. Data privacy can be hindered by malicious users of cloud, if there is any loophole in privacy protection measures. To combat with these threats of data security, data anonymization and data encryption techniques can be used before data sets are analyzed and shared on cloud. The encryption method is not suitable for big data sets. So data anonymization is more suitable option to address the scalability of data.
[2] DATA ANONYMIZATION CONCEPTS AND TECHNIQUES

Data anonymization is the process of masking sensitive data while preserving its format and data type. The masked data can be realistic or a random sequence of data. Output of anonymization can be deterministic, that is, the same value every time. Nature of output is dependent on the techniques used for anonymization. Anonymized data still look real in test environments and yields same results. In this process we remove or modify the identifying variables in micro data datasets. Identifying variables describes characteristics of person i.e. observable or registered. Direct identifiers, which are variables such as names, addresses, identity card numbers, social security numbers. They permit direct identification of an individual respondent, are not needed for statistical or research purposes. Thus they should be removed from the published dataset. Indirect identifiers are those characteristics that may be shared by several respondents, after combining these it may be possible to the re-identification of one of them. For example, the combination of attributes such as district of residence, age, sex, and occupation would be identifying if only one individual of that particular sex, age and occupation lived in that particular district. These variables are needed for statistical mining purposes, and should thus not be removed from the published data.

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

Anonymization data can be placed on cloud without worrying about others will capture it. Afterwards it can be mapped to original data in secure and trusted area. Following are the some anonymization techniques which will help us to provide security to data over cloud: K-anonymity, L-Diversity, T-closeness.

[2.1] K-ANONYMITY

The concept of K-anonymity was first introduced by Latanya Sweeney[7] in 2002. She had proposed this approach to solve the problem of identifying individual from person specific structured data released. Data released said to be k-anonymous when each record cannot be distinguished from at least k-1 individuals whose information is published in that data.

Each dataset in microdata will contain i) Personal identification attributes (Ex: name, address, ssn, UID) ii) Quasi identifiers (Ex: Gender, age, Zip code) iii) Sensitive attributes (Ex: Salary, Disease). Quasi Attributes are suppressed or generalized until each row become identical with k-1 other rows and we can say database is k-anonymous. It prevents database linkage up to some level. Accuracy of data is guaranteed in k-anonymity model.

There are two methods for K-anonymization, viz: Suppression, Generalization. In suppression some attribute values are replaced by *(asterisk) and in generalization individual values of attributes are replaced by broader category (ex: Age=30 can be replaced by age:[25-35]).

Consider the following data of patients (dummy).
Name and religion attributes are suppressed by * and age attribute is generalized in anonymized dataset.

Anonymized database will look as :

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Gender</th>
<th>Religion</th>
<th>State</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naina</td>
<td>28</td>
<td>Female</td>
<td>Hindu</td>
<td>Tamil Nadu</td>
<td>Cancer</td>
</tr>
<tr>
<td>Payal</td>
<td>23</td>
<td>Female</td>
<td>Hindu</td>
<td>Kerala</td>
<td>Viral infection</td>
</tr>
<tr>
<td>Nasreen</td>
<td>25</td>
<td>Female</td>
<td>Muslim</td>
<td>Tamil Nadu</td>
<td>TB</td>
</tr>
<tr>
<td>John</td>
<td>29</td>
<td>Male</td>
<td>Parsi</td>
<td>Karnataka</td>
<td>No illness</td>
</tr>
<tr>
<td>Smita</td>
<td>21</td>
<td>Female</td>
<td>Christian</td>
<td>Kerala</td>
<td>Heart-related</td>
</tr>
<tr>
<td>Vinod</td>
<td>24</td>
<td>Male</td>
<td>Buddhist</td>
<td>Karnataka</td>
<td>TB</td>
</tr>
<tr>
<td>Mrunal</td>
<td>20</td>
<td>Male</td>
<td>Hindu</td>
<td>Kerala</td>
<td>Cancer</td>
</tr>
<tr>
<td>Bindu</td>
<td>30</td>
<td>Male</td>
<td>Hindu</td>
<td>Karnataka</td>
<td>Heart-related</td>
</tr>
<tr>
<td>Joseph</td>
<td>27</td>
<td>Male</td>
<td>Christian</td>
<td>Karnataka</td>
<td>Heart-related</td>
</tr>
<tr>
<td>Leon</td>
<td>28</td>
<td>Male</td>
<td>Christian</td>
<td>Kerala</td>
<td>Viral infection</td>
</tr>
</tbody>
</table>

The above dataset has 2-anonymity in terms of age, gender and state because any combination of these attributes found in any row of the table, there are always at least 2 rows with those exact attributes.

Attacks on K-anonymity : K-anonymity is vulnerable to homogeneity and background knowledge attacks. An attacker can predicate values of sensitive attributes when there is little diversity in values of those attributes. K-anonymization also does not guarantees about attacks using background knowledge.

[2.2] L-DIVERSITY

It is an extension to k-anonymity approach. L-diversity approach ensures that in each group of sensitive attributes have at least L different values. Main requirement of L-diversity is values of sensitive attributes in each group are well represented. To overcome the shortcomings of k-anonymity model, this approach can be used.
L-diversity Principle: An equivalence class is said to have L-diversity if there are at least L “well-represented” values for the sensitive attribute. A data is said to have L-diversity if every equivalence class of the data has L-diversity.

[2.2] T-CLOSENESS

T-closeness is a further fine-tuning of L-diversity group based anonymization that is used to preserve privacy in data sets by reducing the granularity of a data representation. This reduction result into a trade-off those results in some loss of effectiveness of data mining algorithms in order to gain privacy.

T-closeness Principle: An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. If all equivalence classes have t-closeness, we can say a table has t-closeness.

[3] SURVEY ON DATA PRIVACY PRESERVATION AND PRIVACY PROTECTION CLOUD COMPUTING ENVIRONMENTS USING MAPREDUCE

[3.1] A SCALABLE TWO-PHASE TOP-DOWN SPECIALIZATION APPROACH FOR DATA ANONYMIZATION USING MAPREDUCE ON CLOUD

This paper presented two phase top down specialization approach to anonymize large scale datasets using MapReduce framework. To make full use of the parallel capability of MapReduce on cloud, specializations essential in an anonymization process are split into two phases. In the first phase, original data sets are partitioned into a smaller group of datasets, and these data sets are anonymized in parallel, producing midway results. In the second one, the midway results of first phase are integrated into one, and again anonymized to achieve consistent k-anonymous data sets. MapReduce jobs are used to accomplish the concrete computation in both phases.

[3.2] WORKLOAD-AWARE ANONYMIZATION TECHNIQUES FOR LARGE-SCALE DATASETS

KRISTEN LeFEVR[1] has addressed problem of scalability. They introduced two techniques based on scalable decision trees and sampling in order to allow anonymization algorithms to be applied to datasets larger than main memory. Also they stated quality of data is best judged with respect to the workload for which the data will ultimately be used. These algorithms are highly efficient and result will be high quality data. But this approach fails in top down specialization approach and fails to work in multiple data sets.

[3.3] PRISM – PRIVACY-PRESERVING SEARCH IN MAPREDUCE

Prism is privacy preserving search system specially designed for cloud computing. It provides storage and query processing. Prism is designed to features like parallelism,
efficiency of MapReduce. If we want to use prism, we doesn’t require to modify the underlying system.

Use of prism will not introduce extra overhead on system. Disadvantage of prism is we could not secure public clouds using this approach.

[3.4] SEDIC: PRIVACY-AWARE DATA INTENSIVE COMPUTING ON HYBRID CLOUDS

Sedic is designed to protect data privacy in map reduce operations. Sedic offers a privacy-aware hybrid computing paradigm. Sedic schedules Map’s such that jobs on private clouds handle sensitive data while jobs on public clouds operate on nonsensitive data. It automatically extracts Combiner’s from Reduce functions that allow public clouds to process data.

[3.5] THE HYBREX MODEL FOR CONFIDENTIALITY AND PRIVACY IN CLOUD COMPUTING

The HybrEx model provides a unified way for an organization to employ their own infrastructure for sensitive, private data and computation, while incorporating public clouds for nonsensitive, public data and computation. HybrEx model uses partitioning of data and computation as a way to provide confidentiality and privacy. Here they discussed how this model can be used in one specific execution environment - MapReduce over Bigtable.

[3.6] A PRIVACY LEAKAGE UPPER-BOUND CONSTRAINT BASED APPROACH FOR COST-EFFECTIVE PRIVACY PRESERVING OF INTERMEDIATE DATASETS IN CLOUD

Upper bound privacy leakage constraint-based approach to identify which intermediate data sets need to be encrypted and which do not. Privacy-preserving cost of intermediate data sets can be significantly reduced significantly. But this process is highly complicated, efficient processing of data is quite challenging.

[4] CONCLUSION

A detailed survey is carried out on data privacy preservation and privacy protection cloud computing environments using MapReduce. Every approach has their own advantages and disadvantages. Centralized algorithms results in their inadequacy in handling large scale data sets. Although some distributed algorithms focus on secure anonymization of data sets from multiple parties, rather than the scalability aspect. TDS approach is highly scalable and efficient.
REFERENCES

[1] Xuyun Zhang, Laurence T. Yang, Senior Member, IEEE, Chang Liu, and Jinjun Chen, A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 2, FEBRUARY 2014 363


