Data Anonymization Technique for Privacy Preservation on Mapreduce Framework

1Chhaya S Dule, 2Girijamma H.A, 3Rajasekharaiha K M

1Asst. Prof Dept. of Computer Science & Engineering, Jyothy Institute of Technology, Bangalore,
Visvesvaraya Technological University (VTU), Belgaum, Karnataka, India
2Professor, Department of Computer Science and Engineering, RNS Institute of Technology, Bangalore,
Visvesvaraya Technological University (VTU), Belgaum, Karnataka, India
3Professor & HOD, Department of Computer Science and Engineering, Sri Sri Shivkumara Mahaswamy College of
Engineering, Bangalore,
Visvesvaraya Technological University (VTU), Belgaum, Karnataka, India

Abstract: Cloud Computing is a computing paradigm which enables flexible, on demand and low cost usage of
resources. These advantages are the causes of privacy and security issues, which arises because data owned by
different users are stored in cloud servers instead of under their own control. This paper introduces data
anonymization technique to hide sensitive data to avoid risk. The proposed method is Generalized method data
anonymization using Map reduce on cloud.

Keywords : Cloud Computing, Data Anonymization, MapReduce, Privacy preservation

I. INTRODUCTION

Cloud Computing is a revolutionary computing approach, which provides massive storage and computational capability. It provides user to implement applications cost effectively without investment and infrastructure. It enables users to have flexible, on demand access to computing resources via Internet. These advantages are causes of security and privacy problems, which emerge because data owned by different users are stored in cloud servers and users are losing their own control on data. There are different challenges which must be handled. First of all data confidentiality should be maintained. When sensitive information is stored on cloud servers which is out of users control increasing risk dramatically. Secondly personal information is at risk because ones identity is authenticated according to his information. Privacy is an important issue and has become vulnerable in these technologically advanced time. A key component is the protection of individually identifiable data. Many technique are proposed to protect privacy such as data perturbation, Query restriction, Data swapping, Secure Multiparty computation (SMC). In this paper we discussed different anonymization technique however the data set has been increased tremendously in the bigdata trend and this has become a challenge for anonymization of data set and for processing of large data set. We use Map Reduce integrated with cloud to provide high computational capability for application.

II. MAP REDUCE TECHNIQUE

Map Reduce is a programming model which computes large volume of data sets with parallel and distributed
algorithm on cluster. A Map Reduce program is composed of two user defined functions. A Map function performs sorting of data and Reduce function performs summary operation. Map Reduce is a framework for processing parallelizable problems across large datasets using large number of computers. Computational processing can occur on data stored either in file system (Unstructured) or in a database (Structured). It takes advantage of locality of data, processing data on or near storage assets to decrease transmission of data.

Map Step:
The master node takes input divides it into smaller sub problems and distribute them to worker nodes. A worker
node may do this again leading to multilevel tree structure. The worker node processes smaller problems
and passes answer back to its master node.

Reduce Step: master node collects answers to all subproblems and combines them to form output. Fig 1
represents Map Reduce operations.

![Map Reduce operation](image)

Fig 1: Map Reduce operation
Map Reduce allows for distributed computing of Map and Reduction operations. Each mapping operation is performed in parallel and independently although it is limited by number of processors near each source. Similarly set of reducers can produce the reduction phase. Provided all outputs of the map operation that share the same key presented to same reducers at the same time or if the reduction function is associative. This process appears inefficient compared to algorithms that are more sequential. Map reduce can be applied to larger datasets than commodity servers to sort petabyte of data. The parallelism also offers scalability and fault tolerance by recovering from partial failure of servers or storage during the operation. If one mapper or reducer fails, the work can be rescheduled assuming that input data is still available. A Map Reduce approach consists of 5 step parallel and distributed computation.

1. Prepare the Map() input – “The MapReduce system” designates Map processors, assigns the K1 input Key value each processor would work on, and provides that processor with all the input data associated with that key value.

2. Run the user provided Map() code – Map() run exactly once for each key value, generating output organised by key values K2.

3. “Shuffle” the map output to the Reduce processors – The map Reduce system designates Reduce processors, assigns the K2 key value each processor with all the Map generated data associated with that key value.

4. Run the user provided Reduce() code – Reduce() run exactly once for each K2 key value produced by the Map step.

5. Produce the final output – The Map Reduce system collects all the reduce output and sorts it by K2 to produce final outcome.

Fig 2 shows the steps in Map Reduce

Following Algorithm is used to MapReduce System.

1. class MAPPER
2. method MAP(string t, integer r)
3. EMIT (string t, pair (r,1))

III. DATA ANONYMIZATION

Data anonymization approach is extensively studied and widely adopted for privacy preservation in non interactive data publishing and sharing scenario. Anonymity is a technique that organisation uses to increase security of data in public cloud. Key pieces of confidential data are obscured in a way that maintains data privacy. Data can still be processed to gain useful information by analyzing it. Data anonymization is process of changing data that will be published in a way that prevent identification of owner data. There are many approaches through which anonymization of data can be achieved such as K-anonymity, l-Diversity used to make data more private and anonymous to disable inferences from databases. By privacy preservation techniques organisations can benefit from user confidence, convenience and speed of access to information.

k-Anonymity

k-anonymity is a formal model of privacy. The goal of k-anonymity is to make each record indistinguishable from a defined knumber of other records if attempts are made to identify the data.
A set of \( k \) data is anonymized if for any record with given set of attributes there are at least \( k \) other records that match these attributes. Ex Consider a data set that contains two attributes Gender and Birthday. The data is anonymized if for any record \( k \)-1 other records have same Gender and Birthday. In general higher the value of \( K \) more privacy is achieved. \( k \) anonymity assigns properties in specific ways. The \( k \) anonymity data attributes are of three types

Key Attribute: Attribute which can identify n individual directly.

Quasi - identifier: Attribute which can be linked with external information to identify an individual.

Sensitive Attribute: Attribute or data that an individual sensitive about revealing.

There are two main ways through which \( k \) Anonymity can be achieved:

Bottom-Up Generalization:In Bottom-Up Generalization strategy, the data is initialized to its current state and generalizations are carried out over attribute values until \( k \) anonymity is not violated. In other words,to achieve required Anonymity , we have to climb up the Taxonomy Tree of an attribute.

Top-Down Specialization: In Top-Down Specialization strategy, all the attribute values are initialized to the root value of the hierarchy tree. The specialization is carried out iteratively over the attribute values, until the \( k \) anonymity is violated. The specialization is performed by replacing the parent attribute value by its child value in Taxonomy Tree.

Below is the Taxonomy Tree for the attribute SEX and AGE.

Fig 3: Taxonomy Tree for categorical attribute "Sex"

Fig 4: Taxonomy Tree for continuous attribute "Age"

**Definition:** \( k \) Anonymity: Let RT \((A_1,...,A_n)\) be a table and QIRT be the quasi-identifier associated with it. RT is said to satisfy \( k \) anonymity if and only if each sequence of values in RT[QIRT] appears with at least \( k \) occurrences in RT[QIRT]

As an example see the table below. If this table is to be anonymized with Anonymization Level (AL) set to 2 and the set of Quasi Identifiers as QI = \{AGE, SEX, ZIP, PHONE\}. The quasi-identifiers are identified by the organization according to their rules and regulations.

<table>
<thead>
<tr>
<th>NAME</th>
<th>AGE</th>
<th>SEX</th>
<th>ZIP</th>
<th>PHONE</th>
<th>DISEASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aman</td>
<td>20</td>
<td>M</td>
<td>190014</td>
<td>9419</td>
<td>Bronchitis</td>
</tr>
<tr>
<td>Balu</td>
<td>30</td>
<td>M</td>
<td>190001</td>
<td>9592</td>
<td>Lung Cancer</td>
</tr>
<tr>
<td>Charlie</td>
<td>40</td>
<td>M</td>
<td>192231</td>
<td>9823</td>
<td>STI</td>
</tr>
<tr>
<td>Dolly</td>
<td>50</td>
<td>F</td>
<td>190001</td>
<td>8988</td>
<td>Skin Allergy</td>
</tr>
<tr>
<td>Emalie</td>
<td>75</td>
<td>F</td>
<td>190002</td>
<td>8088</td>
<td>Skin Allergy</td>
</tr>
</tbody>
</table>

The NAME attribute here is "Sensitive", so we would like to “suppress” this attribute before anonymizing the above table. After, suppression the table will look like as below:

<table>
<thead>
<tr>
<th>AGE</th>
<th>SEX</th>
<th>ZIP</th>
<th>PHONE</th>
<th>DISEASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>M</td>
<td>190014</td>
<td>9419</td>
<td>Bronchitis</td>
</tr>
<tr>
<td>30</td>
<td>M</td>
<td>190001</td>
<td>9592</td>
<td>Lung Cancer</td>
</tr>
<tr>
<td>40</td>
<td>M</td>
<td>192231</td>
<td>9823</td>
<td>STI</td>
</tr>
<tr>
<td>50</td>
<td>F</td>
<td>190001</td>
<td>8988</td>
<td>Skin Allergy</td>
</tr>
<tr>
<td>75</td>
<td>F</td>
<td>190002</td>
<td>8088</td>
<td>Skin Allergy</td>
</tr>
</tbody>
</table>

Anonymizing the data through Top-Down Specialization, each attribute value will be initialized to the root of Taxonomy Tree and will look like as below:

<table>
<thead>
<tr>
<th>AGE</th>
<th>SEX</th>
<th>ZIP</th>
<th>PHONE</th>
<th>DISEASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0-100]</td>
<td>ANY</td>
<td>*****</td>
<td>*****</td>
<td>Bronchitis</td>
</tr>
<tr>
<td>[0-100]</td>
<td>ANY</td>
<td>*****</td>
<td>*****</td>
<td>Lung Cancer</td>
</tr>
<tr>
<td>[0-100]</td>
<td>ANY</td>
<td>*****</td>
<td>*****</td>
<td>STI</td>
</tr>
<tr>
<td>[0-100]</td>
<td>ANY</td>
<td>*****</td>
<td>*****</td>
<td>Skin Allergy</td>
</tr>
<tr>
<td>[0-100]</td>
<td>ANY</td>
<td>*****</td>
<td>*****</td>
<td>Skin Allergy</td>
</tr>
</tbody>
</table>

The data in the above table is highly privacy preserved, but the data utility is very low. The data is highly anonymized. We make a note here that Data Anonymization is not only the single goal that we are trying to achieve through Anonymization. We also make sure that data utility is high enough to make the information useful for mining.

The Top -Down Specialization Algorithm will iteratively specialize the attribute values till the \( k \) Anonymization is violated.

The given table after anonymizing it for \( k=2 \) will look like:

<table>
<thead>
<tr>
<th>AGE</th>
<th>SEX</th>
<th>ZIP</th>
<th>PHONE</th>
<th>DISEASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0-50]</td>
<td>M</td>
<td>1900**</td>
<td>9***</td>
<td>Bronchitis</td>
</tr>
<tr>
<td>[26-50]</td>
<td>M</td>
<td>190001</td>
<td>9***</td>
<td>Lung Cancer</td>
</tr>
</tbody>
</table>
The basic algorithm of Top-Down Specialization is as below.

Algorithm 1 TDS
1) Initialize every value in $T$ to the top most value.
2) Initialize $Cut_i$ to include the top most value.
3) while some $x \in UCut_i$ is valid and beneficial do
4) Find the Best specialization from $U \ Cut_i$
5) Perform Best on $T$ and update $U \ Cut_i$
6) Update $IGPL(\ x\ )$ and validity for $x \in U \ Cut_i$
7) end while
8) return Generalized $T$ and $Cut_i$

Initially $Cut_i$ contains only the topmost value $U \ Cut_i$ , in which case it returns the masked table together with $Cut_i$

In order to maintain the data utility we have to find the best specialization, otherwise the data utility will decrease considerably and can leave data futile for mining purpose. The metrics that are used to find the best specialization are to be chosen. We use $IGPL\ (\ Information\ Gain\ Privacy\ Loss\ )$ for each specialization $spec\ : \ p\ child\ (\ p\ )\ calculated\ as:\

$IGPL(spc) = IG(spcl) / PL(spcl)$ (1)

The term $IG(spcl)$ is the information gain after performing specialization, the $PL(spcl)$ can be computed via statistical information derived from data sets. Let $R$ denote the set of original records, containing attribute values that can be generalized to $x$ . $|R_x|$ is the number of data records in $R$ . Let $I(R_x)$ be the entropy of $R_x$ . Then, $IG(spcl)$ is calculated as:

$$I(R_x) = \sum_{i} P_i \cdot \log_2 \left( \frac{P_i}{P} \right)$$ (2)

Let $| (R_x, sv) |$ denote the number of data records with sensitive value $sv$ in $R_x$ . $|R|_x$ is computed by:

$$|R_x| = \sum_{sv} \left( \frac{|(R_x, sv)|}{|R_x|} \right)$$ (3)

The anonymity of a data set is defined by the minimum group size out of all QI-groups, denoted as $A$ , i.e.

where

$|QIG\ (\ qid)\ |$ is the size of $QIG\ (\ qid\ )$ . Let $A_{spec}$ denote the anonymity before performing $spec$ , while $A_{spec}$ be that after performing $spec$ .

Privacy Loss caused by $spec$ is calculated by:

$$PL(\ spec\ ) = A_{spec} - A_{spec}$$ (4)

$I-$ Diversity:

$K$ anonymity is vulnerable to a number of attacks . There are two attacks possible.

Homogeneity Attack : If an attacker knows that Balu has an entity in the data and Balu is in his 30 , The attacker will know that Balu has Lung Cancer.

Background Knowledge Attack : If an attacker knows that 21 years old Aman has Bronchitis then attacker knows that Dolly has Skin allergy.$I-$ Diversity another privacy model which can handle both Homogeneity and Background Knowledge Attack and improve the anonymity. Beyond $K$ anonymity,$I-$ Diversity requires there are $I$ different sensitive values for each combination of quasi-identifier.

CONCLUSION:

This paper incorporates the study of Data Anonymization technique used for privacy preservation on cloud on Map reduce Framework to avoid risk which arise because data owned by different users are stored in cloud servers instead of their own control.

REFERENCES:


[8] Benjamin C.M Fung ,Ke Wang and
S.Yu,"Anonymizing Classification Data for Privacy Preservation " IEEE Transactions on Knowledge and Data Engineering ,vol 19 no 5 ,2007.


