IOSR Journal of Engineering
May. 2012, Vol. 2(5) pp: 1091-1094

Preserving Privacy in Horizontally Partitioned Databases Using Hierarchical Model

[1]R. SRINIVAS, [2] Dr. K. PRASADA RAO, and [3]V. SREE REKHA

[1]*Asso.Prof., HoD, Sri Sai Aditya Inistitute of Science&Technology,Kakinada,India*
[2]*Professor, MCA Department.,CMRIT,Engineering College, Bangalore, India*
[3]*Asst. Prof., MCA Dept.,Montessori Mahila Kalasala, Vijayawada, India*

ABSTRACT

It is necessary to preserve information about the individuals in many organizations. So this paper is introduced to implement a distributed anonymization protocol to store data about the users on horizontally partitioned databases by providing secure query protocol to share data according to the specified query from the virtual databases. It concentrates on architecture querying heterogeneous distributed and also private databases**.**

Keywords: *Distributed Anonymization Protocol, Heterogeneous, Partitioned Database, Private Databases, Virtual Databases.*

I. INTRODUCTION

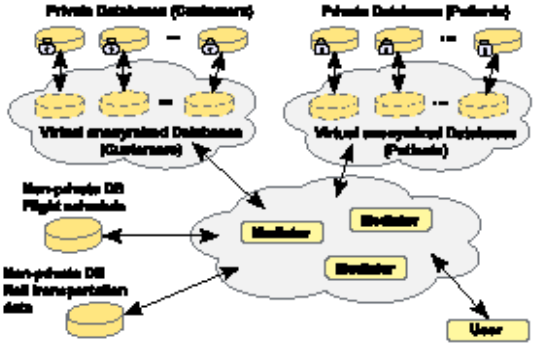
Nowadays, business organizations and government sectors need to maintain large amounts of data to be integrated in distributed databases for storing from scientific experiments, daily transactions in businesses etc. The data may contain personal information as well. For this reason, there is a rapid increase in identifying the value and also a chance in sharing the data through various distributed and mostly non trusted databases. The basic goal of the system is to retrieve the information present in distributed and heterogeneous databases by the investigators who need the share of data that includes personal information. The data sharing involves two limitations namely as privacy for individuals data and data confidentiality of data providers. While executing a query, the results are obtained from various databases by hiding individual's identifiable information.

II. BACKGROUND

The background information regarding privacy-preserving data sharing and integration of distributed heterogeneous data is provided in this section. Privacy preserving data sharing includes a large body of contributed work to anonymized data that transform a dataset to attain privacy principles such as k-anonymity techniques namely generalization, suppression, permutation and swapping of certain data values which are used to hide the individuals identifiable information. Though there are number of ways to preserve privacy data publishing for heterogeneous databases.

A new approach is used to implement data anonymization independently by the clients who query and get the necessary desired data based on the queries from the local anonymized databases by preserving privacy for the data sources. The major drawback with this approach is anonymization is done before the integration of data sources and will result data utility to suffer. Individual databases can access the anonymized data irrespective of ownership.

Another approach includes a third party who can be trusted by each of the data owners. They send their data to the trusted third party to integrate and anonymize later clients can query the centralized database. But by finding such a trusted third party is much difficult. That means if the server gets cooperated with the hackers or unauthorized clients then it leads to complete privacy loss by all the parties. Hence this paper is introduces a distributed data anonymization approach involving data owners in distributed protocols to implement a virtual integrated and anonym zed database for executing client queries. Remember, the anonym zed data is maintained at individual databases and combining the different databases including anonymization of data is processed through the distributed protocols. This approach is mainly used when multi-party computation (MPC) problem occurs rather than comparing with other large MPC problems. In MPC, first consider given number of participants every individual having private data and need to compute the value of a public function. A MPC protocol is said to be secure if and only if no participant can understand the description of public function and also the results of the function. There are also some works focused on data anonymization of distributed databases.



General architecture for secure data sharing

Architecture was designed for general purpose query infrastructure called DObjects depending on a distributed mediator-wrapper which includes a distributed processing engine that executes sub queries on system nodes in a dynamic and iterative way to minimize response time and throughput. DObjects are used for a distributed mediator-based system through which a federation of mediators and wrappers form a virtual system as a peer to peer network.

III. PRIVACY MODEL

This section describes about various privacy models for characterizing to achieve goals of privacy. The privacy goals are attained as two fold process as we discussed already about it in the previous section. At first check that the results of queries must not contain identifiable information. Then individual databases must not maintain data to each other apart from the outcomes of query.

3.1 Individual identifiability

k-anonymity is the most widely accepted models from existing models and serves as the basis for many others. For this reason, this model is used in our approach. Three ways are considered to define anonymization, attributes for a specified relational table T. Individuals are identified by attributes called as unique identifiers such as (X1,……Xd) that can be joined with external information to identify records again. The known identifiers called quasi-identifiers are removed completely from the released micro data. The k-anonymity model gives a way to prevent the duplicated entities by establishing unique entity for identifying each entity. All the tuples form a set in T containing similar values for the quasi-identifier set X1, ..., Xd termed as an Equivalence Class. T is k-anonymous with respect to X1, ..., Xd if every tuple contain size of atleast k is in an equivalence class. A k-anonymization of T is a transformation of the data T where it is k-anonymous.

3.2 Data Confidentiality

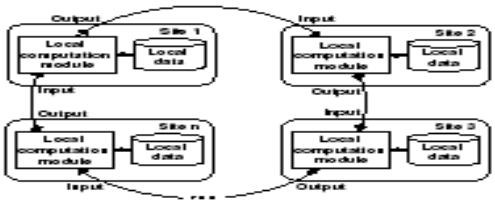
The semi honest model is introduced as the goal of secure MPC but is less inflexible. Apart from query results, data exposure between different parties must be reduced to guarantee absolute security where an individual database does not retrieve any data. This model is generally used in secure MPC problems.

IV. DISTRIBUTED ANONYMIZATION

The distributed anonymization approach is described in this section. Consider the data is split horizontally among n sites ($n > 2$) and each site maintain itsowns private database d_i. The quasi-identifier of each local database is standard in between all the sites. The sites engage in a distributed anonymization protocol where every site provides a local anonymized dataset a_i and their integration forms a virtual database that is guaranteed to be k-anonymous. Remember that whenever users query the virtual database, every individuals database process the query on a_i and then a distributed query protocol is used for attaining good results to guarantee that they are k-anonymous by itself.

4.1. Protocol Structure

Nodes are randomly mapped to a ring topology. Suppose that each node knows its predecessor and successor. Every node has local computation modules that process its part of the protocol independently & results of computation are passed along the ring.



Protocol Structure

Algorithm 1: Distributed anonymization algorithm leading site (i = 0)

1: function split (set d0)
2: Compute range of values for each attribute from quasi identifier in set $d = Sd_i$ (secure min/max protocol)
3: Choose best split attribute a with largest range of values
4: Find median value m of a in set $d = Sd_i$ (secure median protocol)
5: Send a and m to node 1
6: Split set d_0, create two sets, s_0 containing items smaller than m and g_0 containing items greater than m. Distribute median items among s_i and g_i.
7: Find sizeleft = jSsij and sizeright = jSgij (secure sum protocol)
8: if size left > 2 ¤ k then
9: Send split left = true to node 1
10: call split(s0)
11: else
12: Send split left = false to node 1
13: end if
14: if size right > 2 ¤ k then
15: Send split right = true to node 1
16: call split (g0)
17: else

18: Send split right = false to node 1
19: end if
20: end function split

4.2 Distributed Anonymization Protocol

The distributed algorithm for anonymization depends on the Mondrian algorithm using greedy recursive partitioning of the multidimensional quasi identifier domain space. Here the split attribute with highest normalized set of values and partitions of the data depending on the mean value of split attribute is chosen recursively. The process is repeated till no allowable split remains.

Algorithm 2: Distributed anonymization algorithm non leading node (i > 0)

1: function split (set d_i)
2: Read split attribute a and median value m from node (i ¡ 1) and pass them to node i + 1
3: Split set d_i into s_i containing items smaller than m and g_i containing items greater than m. Distribute medianitems among s_i and g_i.
4: Read split left from node i¡1 and pass it to node i+1
5: if split left then
6: call split (si)
7: end if
8: Read split right from node i ¡ 1
9: Send split right to node i + 1
10: if split right then
11: call split(gi)
12: end if
13: end function split

4.3 Secure Query Protocol

This section discusses about a distributed querying protocol that makes users to query this virtual database. Each database executes a query with its local randomized set when a query is received and the results are then combined using the response. Low overhead of the union protocol guarantees reasonable query response times. For a set of nodes with private data items, the problem is to perform computation among all the data items for minimizing data disclosure of the nodes to each other besides the final result.

4.4 Secure set union protocol

To minimize the data exposure set union protocol is used for randomization.

Algorithm 3: Secure set union protocol.

1: INPUT: xi: local subset contributing to union
2: Choose random t_i
3: tmax Ã max(t1::tn) (using secure max protocol)
4: if t_i = tmax then
5: Generate set of random items r
6: Send r [x_i] to successor
7: Receive X from predecessor
8: Result Ã (X ¡ r)
9: else
10: Receive X from predecessor
11: Send X [x_i] to successor
12: end if

The protocol is described as: While initializing for selecting, each node generates a random number t_i from ranges that are already defined. The highest value of t for every node is selected to be the initialized node. To find the highest value of T, secure max protocol is used. The main protocol round includes the leader node I that generates a random set r and adds its local subset xi to the current set. Then the intermediate result is passed to node i+1. Every node j adds its local subset xj to the intermediate result and then passes the result to node j + 1. The set is found by deleting random items r from the set when node i receives the result from its predecessor. A sketch of the algorithm is presented in Algorithm 3. A random set r is generated. The items are generated that are valid to other nodes and are not distinguished from real data. To get a range of valid domain set, numeric attributes are used to secure min/max algorithm.

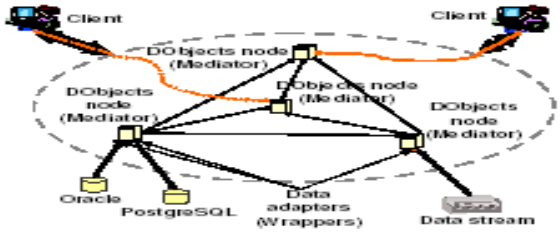


Figure: DObjects system architecture

V. DATA SHARING ARCHITECTURE

The overview of DObjects include distributed mediator-based framework that is used for data integration and query layer for privacy-preserving data sharing in the complete architecture. The framework is used to integrate distributed anonymization and secure query protocols with the help of DObjects framework.

5.1 DObjects Data Services

The system does not include centralized services and uses computing for sharing resources. Every node in the system act as a mediator and provides its computational power that can be used by others during query execution. Nodes are pulled such that they can pull data from external data sources and transform it to a standard format that is expected while building query responses.

5.2 Data model

The data model represents the data as objects and use fixed entities. From users view, query responses are objects of their interested type. In the set of attributes, objects are categorized into two groups: simple and referential. Simple attributes include numbers or strings whereas referential attributes involve an object -oriented language and allow the definition of collecting relations in between data objects.

5.3 Data operations and query languages

DObjects involves all standard data operations. Users can query, create, delete and update constant entities whenever they are interested to perform some operations. User creates queries by building a hierarchy of objects. Every query is created for a given fixed entity type and specifies the types of attributes implemented in the queries.

5.4 Privacy-Preserving Data Sharing

All the components of a system are discussed in the previous topics about security preserving issues using data sharing. Building components include distributed anonymization protocol and distributed querying protocol which guarantees privacy of the individuals in the published data as well as data custodians implementing confidentiality, and the DObjects framework providing a scalable and transparent interface for query execution.

5.5 Architecture implementation proposal

DObjects are used for clients as an abstraction of centralized system for submitting queries in the form of OQL. The basic goal is to provide an efficient access to data requiring/not requiring anonymization. The frameworks are responsible for executing query components and retrieve the outcomes. First DObjects node which receives OQL query performs initial query into fragments. The secure protocol is implemented by considering every DObjects node within virtual groups. Consider a group of cooperating hospitals. The patient databases are maintained independently by each of the hospitals. These are used as a set of horizontally partitioned data distributed among different heterogeneous databases.

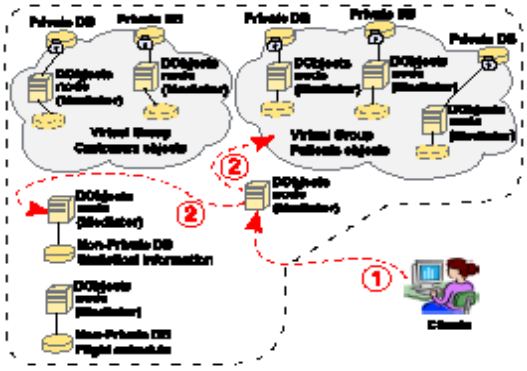


Figure: Implementation of architecture for secure data sharing

Select data.mortality, data.morbidity, patients.age, patients.nation, patients.disease from country c, c.data data, c.patients patients where data.humanity < 0.8 and patients.age > 60

Query: Sample DObjects query

The second referential object may be a list of patients that visit hospitals in each county. As we know in HIPAA act, the personal health information is protected. Therefore the distributed anonymization protocol is replaced in place of independent anonymization to getter better results.

VI. RESULTS

The different dimensions in which the research work continues is namely in developing a protocol toolkit incorporating more privacy principles & algorithms by using different network topologies & optimization techniques in order to further improve the efficiency of the protocols and also in investigating game theoretic approaches.

Final planning is to implement DObjects data services within virtual groups providing a working prototype for secure data sharing platform to get a novel architecture with new ideas in this area of research.

V. CONCLUSION

A conclusion section must be included and should indicate clearly the advantages, limitations, and possible applications of the paper. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

References

[1] S. Zhong, Z. Yang, and R. N. Wright. Privacyenhancing k-anonymization of customer data. In PODS, 2005.

[2] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In IEEE ICDE, 2006.

[3] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on anonymized tables. In ICDE, pages 116–125, 2007.

[4] P. Jurczyk, L. Xiong, and V. Sunderam. DObjects: Enabling distributed data services for metacomputing platforms.