Overview of Privacy in Horizontally Distributed Databases

Kishori Pawar*
Padmabhooshan Vasantdada Patil Institute of Technology
Computer Engineering&Pune University
kishori.bhojane@gmail.com

Y. B. Gurav
Padmabhooshan Vasantdada Patil Institute of Technology
Computer Engineering&Pune University
ybgurav@gmail.com

Abstract— The main aim of privacy is to get the global result without affecting on security. Security and privacy (confidentiality) is of utmost importance in any kind of large scale data-mining, especially where the corporates are involved as parties. Here we overview & introduce a privacy-preserving algorithm for horizontally partitioned data distributed over two or more parties. Our base paper looks at implementing a secure protocol for mining of association rules in horizontally distributed database. We aim to extend this work by developing semi-honest model. This could be an ideal approach for a scenario where mining is difficult in a distributed database system due to the lack of trust demonstrated by databases in each other's association rules, leading honest nodes to lose privacy. Scientists working in this area have proposed their research work for various secure data-mining techniques. This paper reviews their work and gives an idea about the technique proposed above and how it can be helpful in maintaining the security & privacy.

Keywords— Association Rule; Data mining; Horizontally Distributed Databases., Privacy-Preserving, Semi-honest Model

I. INTRODUCTION

While considering data, data may be distributed among the various systems. Most of the businesses share their information along with their personal information for getting equal benefits. Sharing of this type of personal information arise the privacy issue. Though businesses share their private information but still they focus on to the data remains as a private only. This is known as secure mining. For the user the distributed database is like a single compartment it is not in scattered format. As data is increasing day by day we need to store it on different computer and whenever user want to access it, it work like a single unit though we are storing the data on different machines. The data on several computers can be simultaneously accessed and modified using a network. In a network each server is linked by its local database management system (DBMS), and each cooperates to maintain the consistency of the global database. To maintain the privacy of the data many scientist put their efforts so that we get data without losing the privacy of that related data.
Whenever we are concerning with data mining, Security is measure issue while extracting data. Privacy Preserving Data Mining concerns with the security of data and provide the data on demand as well as amount of data that is required. Sometimes it may happen that we get the information but not complete. The Privacy Preserving algorithm is concerns on the basis of its performance, data utility, and level of uncertainty or resistance. There are various techniques and tools for security are used. For mining the data many protocols were proposed by various scientists keeping common goal as to protect sensitive data. While studying about this problem of privacy preserving most of the scientist goes through by searching frequent item set and accordingly related association rule. Association rule indicates the association between various entity while fetching the data or getting the result. Suppose there is one man who wanted to buy bread at that time there is maximum possibility of buying the milk. This is association rule, where the things are connected with each other. Many business areas use this association rule for getting the benefits.

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were highly time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. In the horizontal distributed databases system, the data are stored on different machines. And this information belongs to the one particular related subject. Consider one example for proper understanding of this concept. Let there is one table which is having lots of records in rows and columns format. Some system may store some records which belong to the same column and other columns along with their information may store on the next machines. So the data is distributed along the various machines. If the data stored on different machines and is divided by rows means while storing the data on different machines some rows are stored on one machine and other are stored on different machines. i.e. partitioning the data according to rows is called horizontally partitioned (Distributed Databases) and if that data is stored and partitioning according to the column wise then it is called as a vertical partitioning (Distributed databases). Some may prefer to store or partition of data according to vertical and some may prefer to use horizontal partition. Most of the scientist uses semi-honest model where node may follow or not all protocol for accessing the data among distributed system.

Following figure illustrate these two partitions.

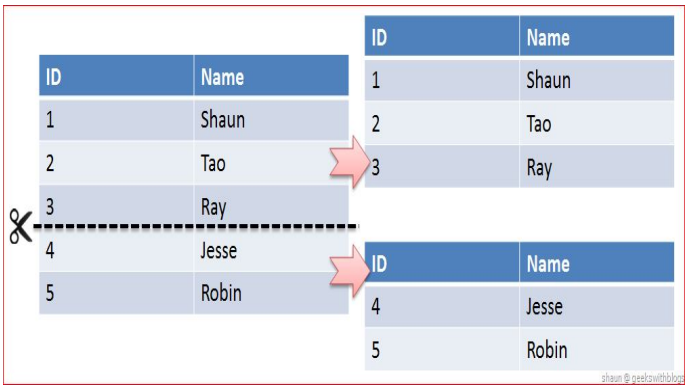

Fig. 1 Horizontal Partition of data

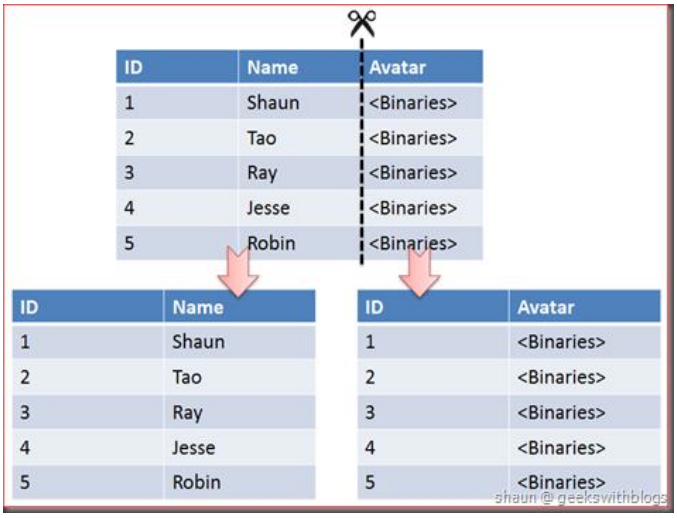

Fig.2. Vertical Partition of Data

Whenever we are considering the databases, data are distributed among the various parties and whatever result i.e. collective result of these all data either of the vertical partitioned or horizontal partitioned. As name specifies in distributed system data is not store in a single computer and as data is increasing day by day we need to store it on multiple storing devices. That's why we are using to store it on multiple machines. In this system though we are storing this data in scattered format, it is appearing like a single system for the user. User need not to know anything about which data is stored on which machine. User need to know only about the information that is he returning as a result. As the data is stored on different machines we need to access it simultaneously and also need to update the data or need to modify the data if any user changes that data within the network. Each database server in the distributed database is controlled by its local DBMS, and each cooperates to maintain the consistency of the global database. A database server is the software managing a database, and a client is an application that requests information from a server. Each computer in a system is anode. A node in a distributed database system can be a client, a server, or both. Two processes ensure that the distributed databases remain up-to-date called replication and duplication.

Replication involves using specialized software that looks for changes in the distributive database. Once the changes have been identified, the replication process makes all the databases look the same. The replication process can be complex and time-consuming depending on the size and number of the distributed databases. This process can also require a lot of time and computer resources.

Duplication, on the other hand, has less complexity. It basically identifies one database as a master and then duplicates that database. The duplication process is normally done at a set time after hours. This is to ensure that each distributed location has the same data. In the duplication process, users may change only the master database. This ensures that local data will not be overwritten.

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The

automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were highly time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

Data mining techniques are the result of a long process of research and product development. This volution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

II. SYSTEM ARCHITECTURE

System architecture describes the flow of data inside the system. It goes through various phases as shown in figure number 3. It is having initialization, in which the player is starting their role by holding some value in it. And then it will help to find out the next item. Next phase is generating candidate set, in which we are finding the key which appears repeatedly or you may say it which is intersection or common for both sites and players. Next phase is local pruning, in which we are trying to eliminate the unwanted result or extra data which will in turn help in mining the data. Next phase is Candidate key union, as word indicates it is based on the union of data of participating players. Next phase is local support computation, in which we are computing the local support that how much the participating player can support. Next phase is Broadcasting of the mining result in which we are going to display the result by merging the all result that we got from all participating player and then displaying it.

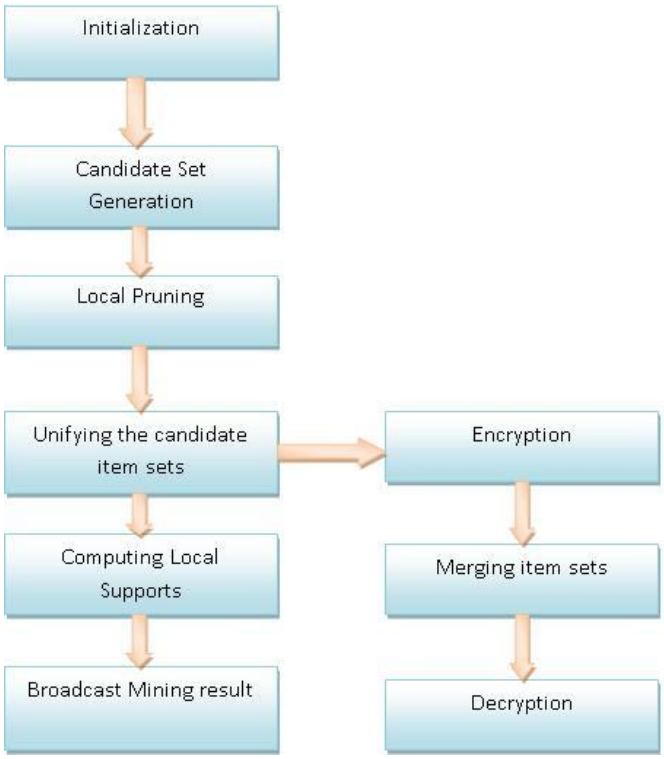

Fig. 3 System Architecture

III. METHOD AND PROCEDURE

A. *Parameter for generating synthetic database*

We are going to use following parameters for creating synthetic database for our work.

N= Number of databases in whole database
L=Number of items
At=Transaction average size
A_f=Average size of maximal potentially large itemsets
N_f=Number of maximal potentially large itemsets
C_S=Clustering size
P_S=Pool size
C_{or}=Correlation level
$M_{F=}$ Multiplying factor

B. *Title and Author Details Distributing Databases*

Once we get synthetic database D, next step is to distribute them among the players M. By splitting D over the M we will get D_m. such that $1 \leq m \leq M$. let $w_m$be the random number

C. *Comparison*

For experimental set up we use three different sets for comparing the result as used in [16] by using UNIFI-KC and UNIFI

- N— the number of transactions in the unified database,
- M— the number of players, and
- S— the threshold supports size.

IV. MATHEMATICAL MODEL DESIGN

Input: D-transaction database view as binary matrix N × L where each row is a transaction over some set of items $A=\{a_1,.....a_l\}$, and each column represents one of the items in A, The database D is partitioned horizontally between M players, denoted $P_1,.....P_M$, Player Pm holds the partial database Dm that contains Nm = |Dm| of the transactions in D, $1 \leq m \leq M$. The unified database is D= D_1 ∪.....D_M , and it includes N:=$N_m \sum^{M}_{m=1}$ transactions. An item set X is a subset of A. Its global support, supp(X), is the number of transactions in D that contain it. An itemset X is called s-frequent if supp(X) ≥ sN. It is called locally s-frequent at Dm if suppm(X) ≥ sNm, for each $1 \leq k \leq L$, let Fk s denote the set of all k-itemsets,

Process:

1. The Fast Distributed Mining algorithm

Step 1: Initialization

It is assumed that the players have already jointly calculated F_s^{k-1}. The goal is to proceed and calculate F_s^k.

Step 2: Candidate Sets Generation

Pm computes the set $F_s^{k-1,m} \cap F_s^{k-1}$. Then apply on that set the Apriori algorithm in order to generate the set of $B_s^{k,m}$ candidate k-itemsets.

Step 3: Local Pruning

For each X € $B_s^{k,m}$, Pm computes suppm(X). Then retain only those itemsets that are locally s-frequent denoted as $C_s^{k,m}$.

Step 4: Unifying the candidate itemsets

Each player broadcasts his $C_s^{k,m}$ and then all players compute $C_s^k := U_{m=1}^M C_s^{k,m}$.

Step 5: Computing local supports

All players compute the local supports of all itemsets in C_s^k .

Step 6: Broadcast Mining Results

Each player broadcasts the local supports that he computed. From that, everyone can compute the global support of every itemset in C_s^k . Finally, F_s^k is the subset of C_s^k that consists of all globally s-frequent k-itemsets.

V. ALGORITHM

There are following some steps that are going to use in this scheme.

Step 1:

All the players generate keys using following key generation method. Key Generation: Let k be the security parameter that chooses two randomly k-bit prime numbers p and q. then set N=pq. Choose random base g Є B.

Step 2:

Player then jointly calculates Fsk-1.

Step 3:

Each player encrypts Fsk-1 using following encryption method.

Let c = cypher text.

Where c= gm rnmod N2

Where r= random value, r Є z*n

Step 4:

Each player Pm computes (k-1) item sets that are locally frequent in his site and also globally frequent Pm then computes Fsk-1,m∧ Fsk,m. He then uses this to generate Bsk,mof candidate k item set and encrypt bits using step 3 equation.

Step 5:

For each X Є Bsk,m, Pm computes suppm(X) and encrypt it using step 3 equation. He then retains only those item sets that are locally s frequent.

Step 6:

Each player broadcast his encrypted $C_s^{k,m}$ and then all player computes $C_s^k := U_{m=1}^M C_s^{k,m}$

Step 7:

Computing local support is now done by all players

Step 8:

Each player broadcast the local support that he computed and encrypts it before sending from that everyone can compete global support of every item set C_s^k.

VI. RELATED WORK

There are various methods that are used by different author for developing secure protocol for the mining of data. Most of the time data is distributed for the sharing purposes but the sharing or collectively analysis of this data is not possible for the security purpose. And everyone wants their data to be private always. If we got the data which does not use to get only private information or which does not effect on anyone's private data then that data mining will not all having the privacy issue. While privacy issue is concerns some paper focuses on two settings. Privacy preserving can be divided into following two categories

1. Perturbation and randomized based approach
2. Secure multiparty computation based approach

In second approach is based on the cryptographic tools for mining the data. But as it concerns with the multi-party, i.e. multiple user that's why the computation cost and communication cost is higher than first one. As the number of user increases the cost required to handle is also large. That's the main thing which is reduced by modifying the data mining algorithm for perturbation technique which will build classifier directly. Some businesses like hospital or bank need to preserve the personal information and they need to share the person specific record. There are some generalized techniques that were used at the cost of loss of information. And these techniques were used to solve the external linkage problem. There are mainly 2 things that come along with the result: Quasi identifier and sensitive attribute to protect our data we can do it by simply do not display the result together by quasi identifier and sensitive attribute. The main help of these, quasi identifier and sensitive attribute, is to support data mining tasks that consider both type of attribute. For improving the method for privacy we are transforming a part of quasi-identifier and personalizing the sensitive attribute values. Some paper implements clustering. Clustering is nothing but the grouping. That means while considering with the data, it is the proper grouping of inter related data. While sharing this data in a group , privacy is the measure issue. To preserve data privacy we may go through the use of synthetic data generation. So whenever we are applying the synthetic data generation technique IPSO families of methods are used. It uses to generate accurate result in cluster. Yao [2] was the first to propose a generic solution for this problem in the case of two players. Other generic solutions, for the multi-party case, were later proposed in [13], [14]. Some paper implements like, the inputs are the partial databases, and the list of association rules is the required output that hold in the unified database with no smaller than some defined support and confidence. [5], [16], [17], [1], [18].Kantarcioglu and Clifton studied and develop protocol. This protocol was computing the union of private subsets that are possessed by different players. In this main part of their protocol is a sub-protocol and hence it increases its cost. It implemented by hash function, obvious transfer and encryption. [19].

VII. CONCLUSIONS

In our dissertation we aim to develop privacy preserving protocol. We extend secure mining protocol used for distributed databases. We believe that this will improve security and privacy of mining operations in distributed database. We proposed a system that will help for secure mining of data in horizontally distributed databases. In our base paper we implement a secure protocol for mining of association rules in horizontally distributed database. We aim to extend this

work by developing privacy preserving protocol using homomorphic cryptography. This is because, in case of distributed databases when databases don't trust each other's association rules mining is difficult as honest nodes may lose privacy.

ACKNOWLEDGMENT

Sincerely thank the all anonymous researchers for providing us such helpful opinion, findings, conclusions and recommendations. I wish to thanks various people who contribute their work for privacy preserving and whose theory helped me to write this paper.

REFERENCES

[1] J. Vaidya, Clifton. "Privacy preserving association rule mining in vertically partitioned data " In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada: ACM.2002:639-644.
[2] C. Yao, "Protocols for secure computations," Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science, IEEE Press, New York, 1982.
[3] Gui Qiong, Cheng Xiao-Hui. Association Rule Mining Algorithm Based on Similarity Matrix of Transactions [J]. Journal of Guilin University of Technology, Vol. 28, No.4, Nov.2008, p p. 568-571.
[4] Mahmoud Hussein, Ashraf El-Sisi, Nabil Ismail, "Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous DataBase", Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, Volume 5178/2008, pp. 607 -- 616 (2008).
[5] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, and Michael Y. Zhu "Tools for privacy preserving distributed data mining," SIGKDD Explorations, Vol. 4, No.2pp1-7. (2003),
[6] A. Shamir;"How to share a secret," Communications of the ACM, vol.22 (11), pp.612–613, 1979.
[7] Jaiwei Han and Micheline Kamber, "Data Mining-Concept and Techniques", Morgan Kaufman Publishers, 2nd Edition , 2006.
[8] Moez Waddey , Pascal Poncelet, Sadok Ben Yahia, "Novel Approach For Privacy Mining Of Generic Basic Association Rules," In PAVLAD'09, November 6, 2009, Hong Kong, China, 2009 ACM.
[9] Xuan Canh Nguyen, Hoai Bac Le, Tung Anh Cao, "An Enhanced Scheme For Privacy-Preserving Association Rules Mining On Horizontally Distributed Databases," In IEEE 2012.
[10] N. V. Muthu Lakshmi1 & K. Sandhya Rani, "Privacy Preserving Association Rule Mining in Vertically Partitioned Databases," In International Journal of Computer Applications (0975 – 8887) Volume 39– No.13, February 2012.
[11] Xinjing Ge, Li Yan, Jianming Zhu, Wenjie Shi, "Privacy- Preserving Distributed Association Rule Mining Based on the Secret Sharing Technique," 2009 IEEE.
[12] Zhu Yu- quan, Tang Yang, Chen Geng, "A Privacy Preserving Algorithm for Mining Distributed Association Rules," 19-21 May 2011.
[13] A. Ben-David, N. Nisan, and B. Pinkas. FairplayMP - A system for secure multi-party computation. In CCS, pages 257–266, 2008.
[14] D. Beaver, S. Micali, and P. Rogaway. The round complexity of secure protocols. In STOC, pages 503–513, 1990
[15] Li Liu; Kantarcioglu, M.; Thuraisingham, B. Privacy Preserving Decision Tree Mining from Perturbed Data. In System Sciences, 2009.HICSS '09. 42nd Hawaii International Conference
[16] T. Tassa and D. Cohen. Anonymization of centralized and distributed social networks by sequential clustering. IEEE Transactions on Knowledge and Data Engineering, 2012.
[17] T. Tassa and E. Gudes. Secure distributed computation of anonymized views of shared databases. Transactions on Database Systems, 37, Article 11, 2012.
[18] S. Zhong, Z. Yang, and R.N. Wright. Privacy-enhancing k-anonymization of customer data. In PODS, pages 139–147, 2005.
[19] M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering, 16:1026–1037, 2004.
[20] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," in Proc. 7th Int. WWW Conf., Brisbane, Australia, 1998.

