ISSN:0975-9646

Privacy Preserving in Horizontal Aggregation Using Homomorphic Encryption

Patil Amrut A[#1], D. M. Thakore[*2]

[#1]*M.Tech Computer Department, Bharati Vidyapeeth University College of Engineering Pune, India*

[*2] *Head of Computer Department, Bharati Vidyapeeth University College of Engineering Pune, India*

Abstract— **Preparing a data set for analysis in data mining is a more time consuming task. For preparing a data set it requires more complex SQL quires, joining tables and aggregating columns. Existing SQL aggregations have some limitations to prepare data sets because they return one column per aggregated group. In general, significant manual efforts are required to build data sets, where a horizontal layout is required. Also many data mining applications deal with privacy for many sensitive data. Therefore we need privacy preserving algorithm for preserving sensitive data in data mining. Horizontal database aggregation is a task that involves many participating entities. However, privacy preserving during such database aggregation is a challenging task. Regular encryption cannot be used in such cases as they do not perform mathematical operations & preserve properties of encrypted data. This paper has two main approaches preparing Data set and privacy preserving in data mining. For preparing the data set we can use the case, pivot and SPJ method for preparing the horizontal aggregation and then employ a homomorphic encryption based scheme for data privacy during aggregations. Homomorphic encryption is the conversion of data into cipher text which allows specific types of computation operations to be performed on the data set and obtains encrypted result. The encrypted result is same as the result which is performed on the plain text. Although such schemes are already being used for protecting data privacy, no one has used these schemes in horizontal databases. This scheme must be modified & customized with relevant parameters and constraints for example key generation time, key generation strategy etc**.

***Keywords*— Data mining, Homomorphic encryption, Horizontal aggregation, Privacy preserving**

I. INTRODUCTION

Data mining is the process of analysing data from databases for different perspectives and summarizing it into useful information. Data mining information can be used to increase revenue. There are many existing operators and functions for aggregation in Structured Query Language. The most used aggregation is the sum of a column and other aggregation operators return the maximum, average, minimum or row count over groups of rows. All operations for aggregation have many limitations to build large data sets for data mining purposes. So in the large data set preparing the data for data mining is more time consuming task to prepare it for analysis and also the privacy is the issue in data mining.

Many times in the news we hear that confidential data leaks from databases. Few examples [4] the Homeland Security newswire listed that between 2009 and 2011, 8 million medical records were leaked. Another example is that last year a group of hackers infiltrated into the Sony Play station network and accessed about 77 million user profiles, of which most of it included credit card information. The attackers attack the database server and leak the confidential data.

To preserve such privacy homomorphic encryption can be used. Homomorphic encryption is the conversion of data into cipher text which allows specific types of computation operation to be performed and obtains encrypted result. The encrypted result is same as the result which is performed on the plain text. For example when one user adds two encrypted numbers and then another user could decrypt the result, without either of them being able to find the value of the individual numbers. RSA algorithm is used for an asymmetric encryption. The RSA algorithm was developed by Adi Shamir, Ron Rivest and Leonard Adleman at Massachusetts Institute of Technology (MIT) in 1977. The RSA is the initials of their surnames. RSA which is known of public-key cryptosystem and it is used for secure data transmission.

RSA website gives the mathematical details of the algorithm used in obtaining the public and private keys .The algorithm which involves multiplying two large prime numbers i.e. prime number is number divisible by one and only that number. Also it involves deriving sets of two numbers that constitutes the public key and another set that is private key, using some additional operations. The original prime numbers are not required .Once the keys have been developed they can be discarded. Only the public and the private keys are needed for encryption decryption. Only the owner requires knowing is its private key.

The keys has been developed, the original prime numbers are no longer important and can be discarded. Both the keys public and private are needed for encryption / decryption but only the receiver of a private key ever needs to know it. Using RSA system, the private key never needs to be sent across the internet. For encrypted text public key is used that has been decrypted using private key

Thus, if sender sends message to receiver, sender is able to find the receiver's public key from a central administrator and not the private key. Thus he encrypts a message using receiver's public key. When receiver

receives it, receiver decrypts it with its private key. The message so sent by the sender will be correctly decrypted only by the private key of the receiver. This preserves the privacy as only the authenticated receiver will be able to decrypt the message.

II. Related Work

Weimin Ouyang and Qinhua Huang [9] focused on the privacy-preserving sequential pattern mining in the following situation first is two parties, each having a private data sets, wish to collaboratively discover sequential patterns on the union of the two private data sets without disclosing their private data to each other. Therefore using Homomorphic encryption technology they put forward a novel approach to discover privacy-preserving sequential patterns based on secure two-party computation.

Saleh I., Mokhtar A., Shoukry A. and Eltoweissy M. [8] proposed a new privacy-preserving protocol for association rule mining (P3ARM) over horizontally partitioned data. P3ARM is depended on a distributed implementation of the Apriori algorithm. To arbitrary assign polling sites to collect itemset supports in encrypted forms using homomorphic encryption techniques for each itemset a pair of polling sites are assigned. Polling sites are different for consecutive rounds of the protocol to reduce the potential for collusion. They show analysis and performance that P3ARM significantly outperforms a leading existing protocol. The P3ARM is scalable in the number of sites and the volume of data.

Raju R., Komalavalli R. and Kesavakumar V. [7] faced the multiple parties should collaboratively conduct data mining without breaching data privacy. The objective is to provide solutions for privacy-maintaining collaborative data mining problems. This privacy preserving Add and Multiply Exchange Technology is given and proposed as an approach of Add to multiply protocol based on Homomorphic encryption techniques is defined to exchange the data while keeping its private. They demonstrated data mining task using Privacy-maintaining Naive Bayesian classification. The solution is distributed data which no central and accesses all data for trusted parties.

Jianming Zhu [3] introduced new scheme for privacy-preserving collaborative data mining in distributed environment based on the Homomorphic encryption and ElGamal encryption system. This scheme used to compute the k-nearest neighbour search. It is provable secure and efficient and can prevent colluded attacker. Comparing with the previous issue on this work, they introduced method can be used in multi-parties who want to cooperatively compute the answers without revealing to each other their identity and their private data.

For satisfying privacy there are many Privacy Preserving Association Rule Mining (PPARM) algorithms proposed. Different authors to find the various methods like randomization, heuristics and perturbation and cryptography techniques for find privacy preserving association rule mining in horizontally and vertically partitioned databases. Kumbhar M.N. and Kharat R. [5] are analysis of different methods for PPARM performed and compared there results. For satisfying the privacy constraints in vertically partitioned databases, algorithms based on Homomorphic encryption, cryptography techniques, Secure Scalar product and Shamir's secret sharing technique are used. The combine advantage of both RSA public key cryptosystem and Homomorphic encryption scheme and algorithm that uses Paillier cryptosystem to compute global supports are used for horizontal partitioned databases. They are reviews the wide methods used for mining association rules over distributed dataset while preserving privacy.

The Semi-honest model provides weak security requiring small amount of computation, on the other hand, malicious models provides strong security requiring expensive computations like Homomorphic encryption. However, efficient computation of such set operations is desirable for practical implementation. Miyaji A. and Rahman M.S [6] build efficient and private set operations avoiding the use of expensive tools like Homomorphic encryption, zero knowledge proof, and oblivious transfer. In this they are constructed in game-theoretic model. In other way, instead of being semi-honest or malicious, the parties are viewed as rational and are assumed to act in their self-interest. The game-theoretic model satisfies computational Nash equilibrium.

In a data mining significant portion of time is dedicated to building a data set required for analysis. In a relational database, building such data set usually requires aggregating columns with SQL queries and joining tables. In existing SQL, aggregations are limited. They return a single number per aggregated group, producing one row for each computed number. These types of aggregations are helpful, but required a significant effort to build data sets suitable for data mining purposes. Where in data mining this tabular format is generally required.

Carlos Ordonez [2] had proposed very simple, yet powerful, extensions to SQL aggregate functions to produce tabular form aggregations that return a set of numbers instead of one number for each row. Horizontal aggregations helps for building answer sets in tabular form which is the standard form needed by most data mining algorithms. Here author was explained, two common data preparation tasks including transposition/aggregation and transforming categorical attributes into binary dimensions. They proposed two strategies to evaluate horizontal aggregations using standard SQL that are: 1) based only on relational operators 2) uses the "case" constructs.

III. Proposed System

This section consists of the description of the proposed system. Fig 1 shows the proposed system architecture. This architecture consists of following phases:

i. **Database:** It is an organized collection of data.
ii. **SQL Code Locking:** This step will automatically generate an efficient SQL code in order to evaluate horizontal aggregations. This code locking is necessary so as to get a consistent query evaluation.

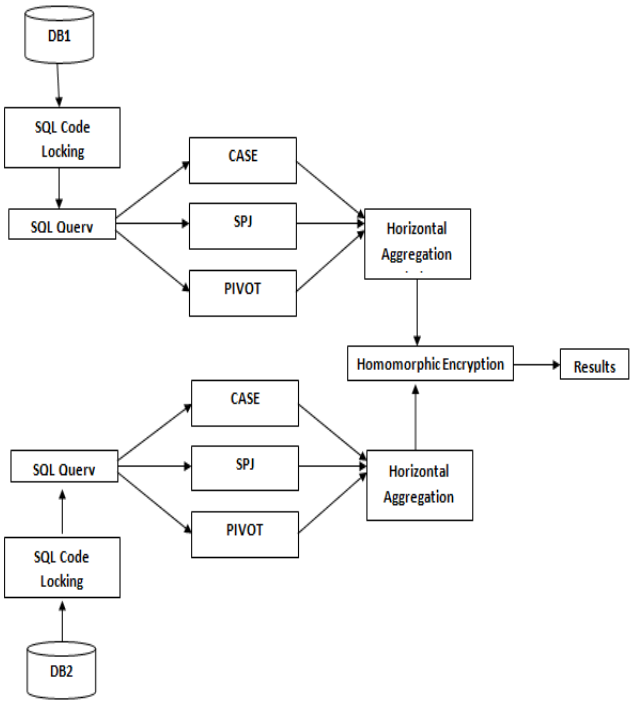

Fig 1: Proposed System Architecture.

iii. **SQL Query Evaluation:** SQL Query Evaluation can be done using following three methods:

a. **SPJ:** SPJ stands for select project join. This method is based on relational operators only. This method creates one table having vertical aggregation for each column of the result and then joins all the tables created to produce horizontally aggregated table.

b. **PIVOT:** PIVOT operator is considered to a built-in operator
in a commercial DBMS. It can help evaluating horizontal aggregations as this operator can perform transposition. This PIVOT method needs to determine internally how many columns are needed to store the transposed table and then can be combined with the GROUP BY clause.

c. **CASE:** It is exploiting the programming CASE construct. In this the case statement returns a value based on boolean expressions selected from a set of values. This is equivalent to doing a simple projection/aggregation query where each non-key value is given by a function that returns a number based on some conjunction of conditions.

iv. **Horizontal aggregation:** Using one of the methods described above, horizontal aggregated dataset is prepared. This dataset so prepared will considerably reduce the time required for data mining.

v. **Homomorphic Encryption**: It is this phase where data privacy is done. Homomorphic encryption converts data into cipher text which can be analysed and worked with as if they are in its original form. Also it allows homomorphic operations to be performed on the encrypted data and obtains results in encrypted form which when decrypted gives the same results as the original result.

IV. CONCLUSIONS

Research area of the data mining Privacy an ongoing research area and there are lots of issues that need to be addressed. Privacy preserving is an ongoing research area in data mining and lots of issues are there in data mining. For preserving privacy homomorphic algorithms are best suited as they are able to perform operations on the encrypted data. The result so obtained is same as the results obtained by performing the same operation on the original text.

REFERENCES

[1] Carlos Ordonez “Horizontal Aggregations for Building Tabular Data Sets”, ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2004.

[2] Carlos Ordonez, Zhibo Chen, “Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis”, IEEE Transactions on Knowledge and Data Engineering 2012.

[3] Jianming Zhu “A New Scheme to Privacy-Preserving Collaborative Data Mining”, Information Assurance and Security IAS '09 Fifth International Conference, 18-20 Aug 2009.

[4] Kathryn Stephens “Security and Privacy Implications of Healthcare Digitization”, NSCI September 22, 2011

[5] Kumbhar, M.N., Kharat, R. “Privacy preserving mining of Association Rules on horizontally and vertically partitioned data: A review” Hybrid Intelligent Systems (HIS), 12th International Conference, 4-7 Dec 2012

[6] Miyaji A., Rahman M.S. “Privacy-Preserving Set Operations in the Presence of Rational Parties”, Advanced Information Networking and Applications Workshops (WAINA), 26th International Conference, 26-29 March 2012.

[7] Raju R., Komalavalli R., Kesavakumar V. “Privacy Maintenance Collaborative Data Mining - A Practical Approach”, 2nd International Conference on, Emerging Trends in Engineering and Technology (ICETET), 16-18 Dec 2009

[8] Saleh I., Mokhtar A., Shoukry A., Eltoweissy M. “P3ARM: Privacy-Preserving Protocol for Association Rule Mining”, Information Assurance Workshop IEEE, 2006.

[9] Weimin Ouyang, Qinhua Huang ,“Privacy Preserving Sequential Pattern Mining Based on Secure Two-Party Computation”, Information Acquisition, IEEE International Conference , 2006.

[10] Weiwei Fang, Bingru Yang, DingLi Song, Zhigang Tang “A New Scheme on Privacy-Preserving Distributed Decision-Tree Mining “, Education Technology and Computer Science, ETCS '09, First International Workshop, 7-8 March 2009.