Survey on Privacy Preservation for Inference Control on OLAP

Maulik Joshi[1], Chetna Chand[2]

[1]*M.E. Student, Department of Computer Engineering, Kalol Institute of Technology & Research Center*

[2]*Assistant Professor,Department Of Computer Engineering,Kalol Institute of Technology & Research Center*

***Abstract-* Online analytical processing (OLAP) is working on analysis of multidimensional data cube and it is support decision making and knowledge discovery technique. Privacy preserving is important in OLAP because private information can be shown through user query. So adversarial inference of private information its main issue in OLAP. privacy preserving OLAP had focused on single aggregate function but it is not consider important class of privacy breaches and partial information has been generated .In this paper proposed technique provide protection for exact and partial disclosure in OLAP with more than one aggregate function with reducing processing time of OLAP cube.**

***Index Terms-* Privacy, preserving, OLAP, inference problem**

I. INTRODUCTION

OLAP is providing knowledge discovery and decision support technique in business intelligence system. We can apply query on multidimensional data cube. Among various ways of data analysis, OLAP is one of the most popular techniques. OLAP helps analysts to extract useful knowledge from large amount of data. Similar to any technology, OLAP is also double edged sword. Without sufficient security components, an OLAP system may become a powerful tool in the hands of malicious users in threating the privacy of individuals. Protection of private information is main issue in online analytic processing system; adversarial inference of private information from OLAP query answers is major privacy problem [1]. Privacy preserving OLAP had focused on single aggregate function but it is not consider important class of privacy breaches and partial information has been generated. In this proposed approach privacy protection in front of both exact and partial disclosure in OLAP systems using more than one aggregation function. Propose approach we consider SUM like function and MIN Like function [4]:

- MIN-like functions: MIN and MAX
- SUM-like functions: SUM, AVG, COUNT, MEDIAN, and STANDARD DEVIATION

It is provide guarantees that the privacy disclosure can not to exceed thresholds predetermined by the data owners. we will implement base paper algorithm with modification to make it faster and also reduce size & processing time of OLAP cubes. Our approach will efficient and can be implemented in existing OLAP systems with little modification. It is satisfies the simulate able auditing model and leaks no private information through query rejections.

II. PROBLEM STATEMENT

The data warehouse server stores private data and server may answer on the multidimensional aggregates of private data by users OLAP queries .However, it is a challenge to enable OLAP on private data without privacy breach of data owners.

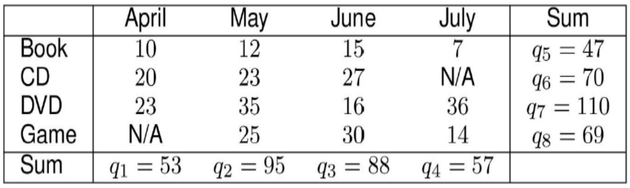
User may not access all individual data in data warehouse but privacy breach occurs when user will get certain information about private data point from OLAP queries but user don't have right to access that private data. So here using query answer user can infer certain information of private data point. This privacy breach become as the inference problem.

Now, example of inference problem as per base paper [1]:

In this example attribute Y is sensitive cell in collection 1 so user can not access information of this private data point.

User asks two queries:

1. What is the total no. of items in collection1?
2. What is the value of attribute X in collection 1?



	April	May	June	July	Sum
Book	10	12	15	7	$q_5 = 47$
CD	20	23	27	N/A	$q_6 = 70$
DVD	23	35	16	36	$q_7 = 110$
Game	N/A	25	30	14	$q_8 = 69$
Sum	$q_1 = 53$	$q_2 = 95$	$q_3 = 88$	$q_4 = 57$	

So first query answer is 47 and second query answer is 7.

Using this two query we identified our private data information such as in collection have 47 total number of items and attribute X contain 7 items out of 47 so we easily to get attribute Y information from this two query.

III. LITERATURE REVIEW

OLAP privacy obtained from data perturbation[5] and show that our perturbation provides guarantees against privacy breaches. Here develop algorithms for reconstructing counts of sub cubes over perturbed data. It is also identify the tradeoff between privacy guarantees and reconstruction accuracy and show the practically of our approach. The perturbation algorithm is everyone known; the actual random numbers used for hide sensitive data in the perturbation Technique. To allow clients to operate independently, it is use local perturbations so that the perturbed value of a data element depend only on its initial value and not on those of the other data elements. It is also proposed reconstruction algorithms both analytically and empirically

Another data perturbation technique called uniformly adjusted distortion [2], in this technique initially distorts one cell and then uniformly distributes this distortion in the whole data cube. This also provides accuracy with range sum queries and high availability. It presented a simple but effective distortion technique for privacy preservation in data cube. Distortion technique would not affect the response time of OLAP system queries as all calculations will be done at source side before the interaction of the users. Also it is applicable without divide a data cube into blocks as oppose to the state of the art technique.

In This Paper[4] address issues related to the protection of private information in Online Analytical Processing (OLAP) systems, where a major privacy concern is the adversarial inference of private information from OLAP query answers. Most previous work on privacy preserving OLAP focuses on a single aggregate function or addresses only exact disclosure, which eliminates from consideration an important class of privacy breaches where partial information, but not exact values of private data is disclosed. We address privacy protection against both exact and partial disclosure in OLAP systems with mixed aggregate functions. In particular, it is propose an information-theoretic inference control approach that supports a combination of common aggregate functions and guarantees the level of privacy disclosure not to exceed thresholds predetermined by the data owners. It is demonstrate approach is efficient and can be implemented in existing OLAP systems with little modification. It also satisfies the simulatable auditing model and leaks no private information through query rejections. Through performance analysis, It is show that compared with previous approaches, This approach provides more effective privacy protection while maintaining a higher level of query-answer availability.

In Base Paper N-D algorithm [1] provided protection against Inference problem. This paper implement N-D algorithm For SUM LIKE Queries and given protection from privacy breaches. It is also demonstrate SUM Like Query in result. . If any query leads to inference problem then rejection of query or precisely give answer.

IV. COMPARISON WITH OTHER OLAP PRIVACY PRESERVING TECHNIQUES

Basic we have two types of technique for privacy preserving in OLAP:

A. Inference control
B. Data perturbation

Using two techniques we prevented from privacy breach so how to privacy preserving using methods describe below:

A. Inference Control

Inference control [1], [4] approach, which is based on three tier architecture. It given layer between user query and cube and in this layer contains predefined inference free aggregation function. If any query leads to inference problem then rejection of query or precisely give answer.

B. Data perturbation

Data perturbation [5] is one more approach for privacy preserving in OLAP. In this technique include noise with input source and when query issues for information, we will get answer with some estimation rather than exact value. So it is disadvantage of this technique.

Distortion technique [2] is part of data perturbation technique .This technique through replace to original value of cell with noise. In distortion technique, perform relative distortion (means 50%) rather whole cube distortion for avoid scalability.

V. ALGORITHM AS PER BASE PAPER

A. Inference Control Algorithm (for n-d Arbitrary Distribution) [1]

Require: h-dimensional query q on sub-cube ($a_1 \dots a_{n-h}$, ALL… ALL), l_o =0.

```
1: {When a query q is received.}
2: if function of q is MIN-like then
3: $l_0 \leftarrow f_{max}$ (q, |q|)/H(x).
4: else if function of q is SUM-like then
```

```
5: for $i \leftarrow 1$ to $n - h$ do
6: $\Theta_i \leftarrow (a_1, a_{i-1}$, ALL, $a_{i+1} . . . a_{n-h}$, ALL,. . . , ALL).
7: Find $T_i \leftarrow \{\Theta' | \Theta' \epsilon \Theta_k^s, \Theta'$ is subset of $\Theta i$ }
8: $\mu \leftarrow \min(\min\{\mu k\ (\Theta') | \Theta' \epsilon T_i\}, d_i + d_{n-h+1} + d_{n-h+2} + ......+ d_n - |T_i|, d_i/2, d_{n-h+1/2}, .......d_{n/2})$.
9: $t \leftarrow$ max(the maximum integer that satisfies $\Sigma \{\sigma_k(\Theta')\ (h-1/h)\ (1/h)\ |\Theta' \in T_i\} \geq t\ (d_i - \mu) + (t\ )\ (\mu-1)(d_1+d_2+.....+d_{h-h\ (t)}), 0)$ assuming 0, (0) =1.
10: $n_0 \leftarrow |q| - r_k(q) - t$.
11: $l_0 \leftarrow \max (l_0, (1 - l_p) x\ (f_{max}(q, n_0) - f_{min}(q, n_0 - 1))/H(x))$
12: end for
13: end if
14: if $l_p + l_0 \geq \min(l_k, l_{(q)})$ then
15: return Ø. {Reject query q}
16: else
17: if function of q is SUM-like then
18: $\mu_k(\Theta) \leftarrow \mu(\Theta)$
19: $T \leftarrow \{ \Theta' | \Theta' \epsilon \Theta_k^s, \Theta' \subseteq \Theta \}$.
20: $\sigma_k(\Theta) \leftarrow \sigma(\Theta) - \Sigma_{\Theta' \epsilon T} \sigma_k(\Theta')$
21: $\Theta_k^s \leftarrow \Theta_k^S$ U $(\Theta, \mu_k(\Theta), \sigma_k(\Theta))$
22: end if
23: $l_k \leftarrow \min (l_k, l(q))$.
24: $l_p \leftarrow l_p + l_0$.
25: return q. {Answer query q correctly}
26: end if
```

Let $r_k(q)$ be the number of cells in q that belong to G_k (i.e., known by C_k as preknowledge).

|q| is the number of all cells in q.

Let Q_k^s be the set of SUM-like queries in the query history Q_k, and $| Q_k^s|$ be the number of queries in Q_k^s.

Let σ_k be the number of preknown cells covered by at least one query in Q_k^s.

That is, $\sigma_k = |\{x | x \in G_k, 3q \in$ such that $x \epsilon q\}|$.

Let μ_k be the minimum number of sensitive cells covered by a SUM-like query in the query history:

$$\mu_k = \min (|q| - r_k (q))\ q: q \in Q_k^s$$

In the algorithm, we use lp denote the current upper-bound estimate on lmax (Q_k).

When Q_k = Ø, the initial values of the parameters are μ_k =infinity, $\sigma_k = 0$, and $l_p = 0$.

With the algorithm, when a new query q is received, the data warehouse server computes an upper-bound estimate on lmax (q $|Q_k$) as l_0, and answers the query if and only if $l_p + l_0$ is less than the owner-specified threshold l.

If $l_p + l_o < l$ (i.e., query answer q can be issued to user Ck), then the data warehouse server updates the values of μ_k, σ_k,$| Q_k^s|$, and l_p in Steps 12-17.

Here, in this algorithm Query q apply on sub cube and query can be used with two different aggregate functions such as MIN Like and SUM Like. After that data warehouse computes an upper bound estimate l_o . if $l_p + l_o$ is greater than threshold value then query rejected but it's value less than threshold value ,so query accepted and query answer will be given to user.

As we can see, only l_p needs to be updated for MIN-like queries while all four parameters are updated for SUM-like ones.

Here, instead of maintaining μ_k and σ_k values for each user, we are maintaining μ_k and σ_k values for each sub cube of query history.

B. Modification made in the algorithm [1]

In line 9 of algorithm B (query evaluation of SUM like queries) we have modified the evaluation of't' such that

9. $t \leftarrow$ the maximum integer that satisfies $\Sigma \{\sigma_k\ (\Theta')\ (h-1) / h\ (1/h)\ |\Theta' \in T_i\}\ t\ (di-\mu) + (t)\ (\mu-1)\ (d1+d2+..... +d_h - h\ (t)$ assuming 0 (0) =1.

9.1. $t_{prev} = t$;

9.2. $t_m = \max (t_{prev} , 0)$;

Where, t_{prev} is t as calculated in the paper. t_m is modified t.

But base paper algorithm [1] has some limitation such as there are no specifications about n-dimensions, privacy measures are also not clear and the format of query specified only with SUM Like is demonstrated where there are other aggregate functions can be useful in some cases**.**

VI. CONCLUSION

This paper proposed different techniques for privacy preserving in OLAP. Through techniques we got control on inference problem but we should make more efficient method, for that purpose we will propose new approach it will be reduced size & processing time of OLAP cube with modification of base paper algorithm.

REFERENCES

[1] Rohit Goel, Mahesh Kumar," Implementation of Privacy Preservation of N-D Algorithms for Online Analytical Processing", IJIRCCE, Vol. 2, Issue 6, June 2014.

[2] Sara Mumtaz, Azhar Rauf, Shah Khusro," A Distortion Based Technique for Preserving Privacy in OLAP Data Cube",IEEE,2011.

[3] Gunwanti R. Bawane, Prof. Prarthana Deshkar," Integration of OLAP and Association rule mining",IEEE,2015.

[4] Nan Zhang, Member, Wei Zhao, Fellow," Privacy-Preserving OLAP: An Information-Theoretic Approach",IEEE, VOL. 23,2011.

[5] Rakesh Agrawal, Ramakrishnan Srikant, Dilys Thomas," Privacy Preserving OLAP", ACM,2005.

[6] Han Jiawei, Kamber M. Data Mining: Concepts and Techniques. San Francisco California: Morgan Kaufmann Publishers, 2001.

[7] Https:// www.wikipedia .org.

[8] Chaudhuri, S. and U. Dayal, "An overview of data warehousing and OLAP technology". ACM Sigmod record, 1997. 26(1): p. 65-74.

[9] Y. Li, H. Lu, and R.H. Deng, "Practical Inference Control for Data Cubes," Proc. IEEE Symp. Security and Privacy, Extended Abstract, pp. 115-120, 2006.

[10] L. Wang, S. Jajodia, and D. Wijesekera, "Securing OLAP Data Cubes Against Privacy Breaches," Proc. 25th IEEE Symp. Security and Privacy, pp. 161-175, 2004.

[11] Sung, S.Y., et aI., "Privacy preservation for data cubes",Knowledge and Information Systems, 2006. 9(1): p. 38-61.

[12] Hua, M., et aI., "FMC: An approach for privacy preserving OLAP.Data Warehousing and Knowledge Discovery", 2005: p. 408-417.

[13] Wang, L., D. Wijesekera, and S. Jajodia, "Cardinality-based inference control in data cubes". Journal of Computer Security, 2004. 12(5): p. 655-692.

[14] Wang, L., S. Jajodia, and D. Wijesekera, "Preserving privacy in on-line analytical processing data cubes". Secure Data Management in Decentralized Systems, 2007: p. 355-380.