Disclosure Control of Sensitve Information from OLAP Query Results

P.Kamakshi
Associate Professor
Department of Information Technology
K.akatiya Institute of Technology and Science
Warangal, Andhra Pradesh, India

Dr .A.Vinaya Babu
Dept. of C.S.E.
J.N.T.U, Hyderabad
Andhra Pradesh, India

***Abstract*— With the rapid development in technology organizations are able collect and store huge amount of business relevant information. Today the companies wish to convert such huge data into quality information. Organizations are looking forward for decision support system where the data collected from heterogeneous sources can be converted into digitized form and stored in a permanent storage space for business analysis purpose. Data warehouse provides the architecture to serve as a single integrated source of data which is different from operational database and stores current and historical records coalesced from multiple transactional systems. OLAP operation on Data warehouse database allows the client to analyze the data interactively and perform other business intelligence functions. The data in the data warehouse is collected from various sources like insurance companies, healthcare systems also [1] contains sensitive information, OLAP operation on such databases may reveal the information which is private to an individual. In this paper we discussed about the privacy violation of sensitive information by malicious user with valid OLAP queries. We proposed a framework to identify the privacy violation of sensitive information during OLAP operation and modify the query results with suitable privacy preservation technique.**

Keywords- OLAP, privacy, sensitive information, technology.

I. INTRODUCTION

The development in software, hardware, and computing and network technology facilitated different organizations to collect and store huge amount of business related information. The organizations realized that the growth of an organization and benefits can only be achieved by sharing the information from other organizations. Today Organizations come forward to store huge information in a single repository and analyze it to improve the service or performance, scientific analysis, research application, customer retention management. The data warehouse is a single integrated source which combines various databases across an entire enterprise. The multidimensional database structure is [3] used to store the data in the data warehouse. The data in the data warehouse y represents the business history of an organization. This historical data is used for analysis that supports business decisions at many levels of abstraction, from strategic planning to performance evaluation of a discrete organizational unit. Data in a data warehouse is organized to support analysis with OLAP rather than to process real-time transactions as in online transaction processing systems (OLTP). Multidimensional cube representation is used to store the data in the data warehouse. A subset of attributes in the database is used to construct a data cube. Few attributes are selected as measure attributes i.e. the attributes whose values are of interest and are aggregated according to the dimensions. Other attributes are selected as dimensions or functional attributes. Figure 1 below depicts 3-D cube model representing sales data having three dimensions i.e. year, product and stores location. Total sales are the measurement of interest. Figure 1 below shows the multidimensional model of data warehouse.

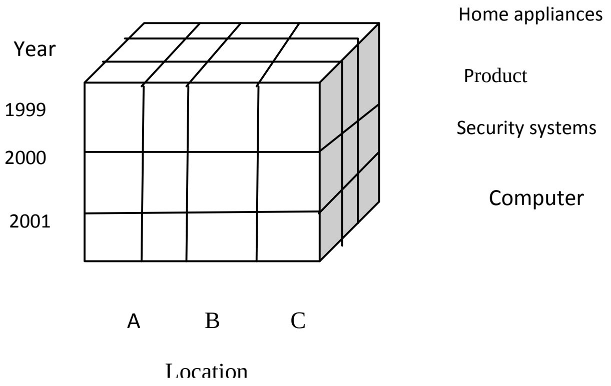


Figure 1. Multidimensional model of data warehouse

OLAP operations are use to perform complex operation on data warehouse data and detect novel patterns and relationships between data items which went unnoticed earlier. OLAP operation also provides a multidimensional presentation of data warehouse data, creating cubes that organize and summarize the data for efficient and quick analytical operation. OLAP technique facilitates the user to extract and view the organizational database from different aspects. For example a user can submit a request to analyze and show and compare all the companies who sold computer security products in America in the month of January with those for the similar type of products in the month of June. Various OLAP operations are rollup, drill-down, cube, pivot, slice and dice.

Multidimensional view of aggregate data is utilized by OLAP to provide quick access to strategic information for further analysis. The information ranges from basic navigation through slice and dice operation to more complex modeling. OLAP operations have the ability to provide information as and when for effective decision-making.

Generally the output of an OLAP is displayed in the form of a matrix. The rows and column represents the dimensions of the query. The analyst can understand the meaning of organizational data base contained in the data warehouse using multi-dimensional analysis. OLAP operations allow the analyst to traverse through the database through different operations like roll up, drill down, slice, dice and pivot.

II . PRIVACY ISSUES IN OLAP OPERATION

Over the past few years the significance of data warehouses and OLAP operation for an organization’s decision support system has rapidly grown. At the same time the problem with information privacy violation has been also developed. A data warehouse stores current and historical records consolidated from multiple transactional systems like health care system, banking, and insurance companies also contains sensitive information. The main aim of data warehouse is to provide access to all necessary data without any restriction on data access. But, as portion of the data is very sensitive and private to an individual, OLAP operation may reveal the pattern containing the [6] sensitive information creating security conflict. The revealed sensitive information can be misused without the knowledge of the actual data owner. The misuse of sensitive information may harm the data owner.

One straightforward solution to protect the privacy is to completely hide the sensitive information in the cell of multidimensional database or such data should not be included in the database. But such kind of suppression of information will not provide the valid results for analysis. There is immense need to provide the solution which can give results for analysis purpose but also protect privacy of an individual. In the proposed framework we investigate the privacy breaches caused by multi-dimensional queries [7] in OLAP (Online Analytic Processing) systems. We propose a novel technique to preserve the privacy of sensitive information during OLAP operation and give valid results for analysis purpose.

II. PROPOSED FRAMEWORK

The growth in technology has enabled the organizations to collect huge volume of information. Today most of the organizations wish to store the data and also share the data for mutual benefit and growth of organization. Many organizations depend on OLAP results to extract novel patterns which can help them to reduce costs, increase business expansion opportunities.

The proposed framework shown in figure 2 is very useful in a collaborative environment where group of parties want to share the data but restrict themselves from revealing the private information [8] about their clients or customers. This framework helps the parties to share the information but also protect the privacy of sensitive information.

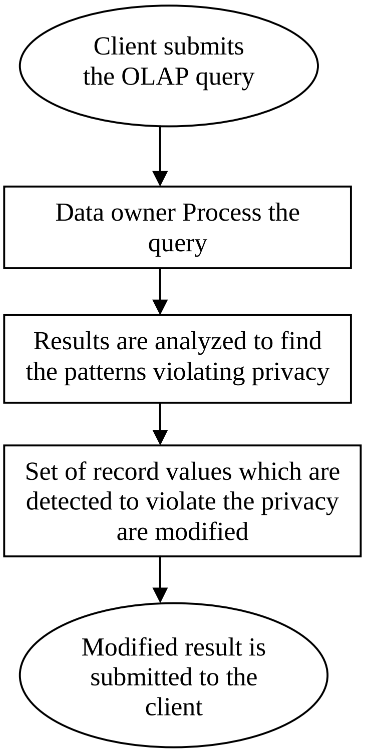
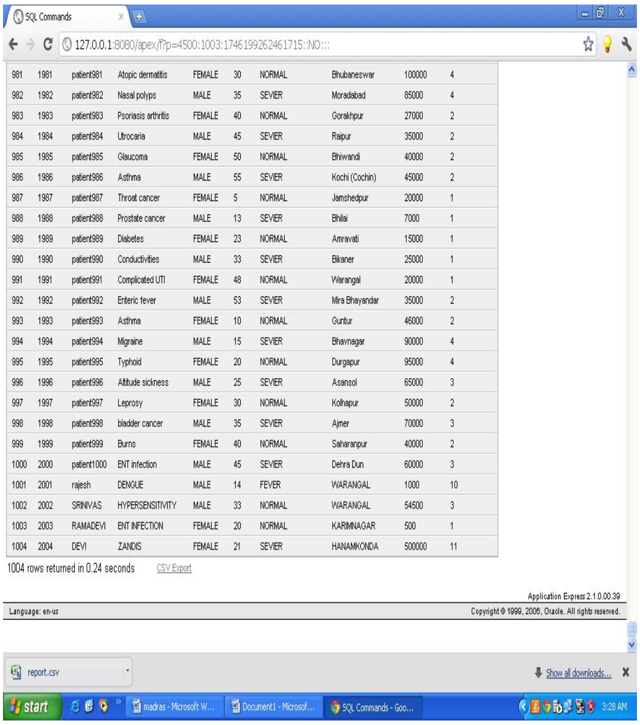

Client submits the OLAP query
Data owner Process the query
Results are analyzed to find the patterns violating privacy
Set of record values which are detected to violate the privacy are modified
Modified result is submitted to the client

Figure 2. The framework to preserve the privacy during execution of OLAP queries.

III. EXPERIMENTAL ANALYSIS

In this experimental work a standard health database of patient information contains nearly 1000 records. The different attributes Patient _id, name, disease, and gender, and age, type of disease, location, bill amount and category.



981	1981	patient981	Atopic dermatitis	FEMALE	30	NORMAL	Bhubaneswar	100000	4
982	1982	patient982	Nasal polyps	MALE	35	SEVIER	Moradabad	85000	4
983	1983	patient983	Psoriasis arthritis	FEMALE	40	NORMAL	Gorakhpur	27000	2
984	1984	patient984	Utrocaria	MALE	45	SEVIER	Raipur	35000	2
985	1985	patient985	Glaucoma	FEMALE	50	NORMAL	Bhiwandi	40000	2
986	1986	patient986	Asthma	MALE	55	SEVIER	Kochi (Cochin)	45000	2
987	1987	patient987	Throat cancer	FEMALE	5	NORMAL	Jamshedpur	20000	1
988	1988	patient988	Prostate cancer	MALE	13	SEVIER	Bhilai	7000	1
989	1989	patient989	Diabetes	FEMALE	23	NORMAL	Amravati	15000	1
990	1990	patient990	Conductivities	MALE	33	SEVIER	Bikaner	25000	1
991	1991	patient991	Complicated UTI	FEMALE	48	NORMAL	Warangal	20000	1
992	1992	patient992	Enteric fever	MALE	53	SEVIER	Mira Bhayandar	35000	2
993	1993	patient993	Asthma	FEMALE	10	NORMAL	Guntur	46000	2
994	1994	patient994	Migraine	MALE	15	SEVIER	Bhavnagar	90000	4
995	1995	patient995	Typhoid	FEMALE	20	NORMAL	Durgapur	95000	4
996	1996	patient996	Altitude sickness	MALE	25	SEVIER	Asansol	65000	3
997	1997	patient997	Leprosy	FEMALE	30	NORMAL	Kolhapur	50000	2
998	1998	patient998	bladder cancer	MALE	35	SEVIER	Ajmer	70000	3
999	1999	patient999	Burns	FEMALE	40	NORMAL	Saharanpur	40000	2
1000	2000	patient1000	ENT infection	MALE	45	SEVIER	Dehra Dun	60000	3
1001	2001	rajesh	DENGUE	MALE	14	FEVER	WARANGAL	1000	10
1002	2002	SRINIVAS	HYPERSENSITIVITY	MALE	33	NORMAL	WARANGAL	54500	3
1003	2003	RAMADEVI	ENT INFECTION	FEMALE	20	NORMAL	KARIMNAGAR	500	1
1004	2004	DEVI	ZANDIS	FEMALE	21	SEVIER	HANAMKONDA	500000	11

1004 rows returned in 0.24 seconds CSV Export

Application Express 2.1.0.00.39

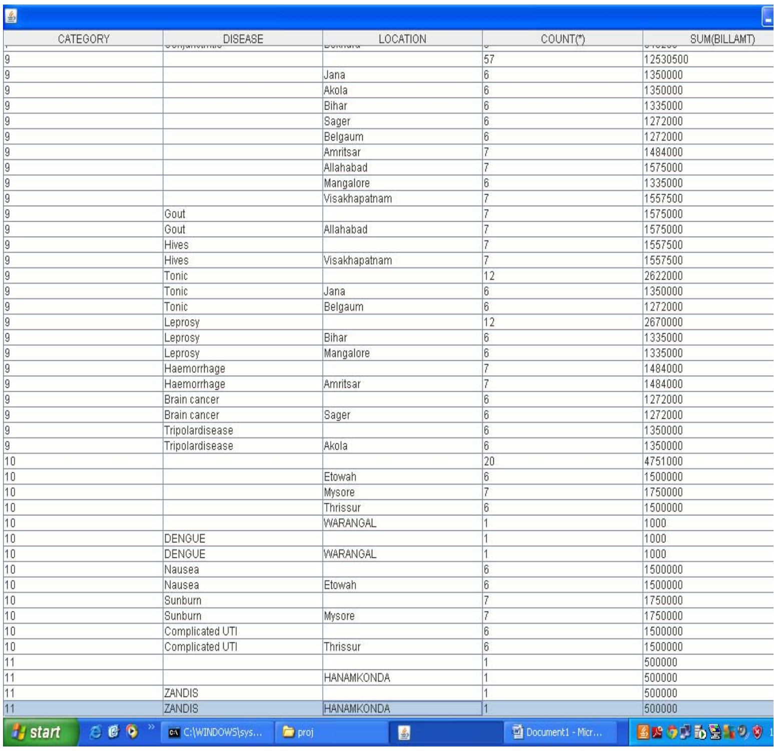
Language: en-us

Copyright © 1999, 2006, Oracle. All rights reserved.

Figure 3. The partial patient information database

The database administrator has the access to all information. But patient _id and name are not revealed for analysis purpose. The remaining attributes are used to perform various OLAP operations. We considered location as one of the sensitive attribute which can reveal the identity of an individual. The disease and bill amount are considered as the major attribute values to perform analysis. We performed cube operation on the given database .The cube operation gives the summarized information for a given query. From the output result it is found that in some cases there is only record from a particular location.

After getting such result having single record value, an intruder can easily [5] generate other queries and extract other information of that particular person The proposed framework modifies the location information with a default value and then the modified results are submitted to the client for analysis purpose. Other attributes values like disease, bill amount for particular disease, age with modified place value can be displayed for analysis purpose. Figure 3 shows the partial database considered for experimental work. Figure 4 below shows the output of cube operation showing the special record with count equal to 1.



CATEGORY	DISEASE	LOCATION	COUNT(*)	SUM(BILLAMT)
9			57	12530500
9		Jana	6	1350000
9		Akola	6	1350000
9		Bihar	6	1335000
9		Sager	6	1272000
9		Belgaum	6	1272000
9		Amritsar	7	1484000
9		Allahabad	7	1575000
9		Mangalore	6	1335000
9		Visakhapatnam	7	1557500
9	Gout		7	1575000
9	Gout	Allahabad	7	1575000
9	Hives		7	1557500
9	Hives	Visakhapatnam	7	1557500
9	Tonic		12	2622000
9	Tonic	Jana	6	1350000
9	Tonic	Belgaum	6	1272000
9	Leprosy		12	2670000
9	Leprosy	Bihar	6	1335000
9	Leprosy	Mangalore	6	1335000
9	Haemorrhage		7	1484000
9	Haemorrhage	Amritsar	7	1484000
9	Brain cancer		6	1272000
9	Brain cancer	Sager	6	1272000
9	Tripolardisease		6	1350000
9	Tripolardisease	Akola	6	1350000
10			20	4751000
10		Etowah	6	1500000
10		Mysore	7	1750000
10		Thrissur	6	1500000
10		WARANGAL	1	1000
10	DENGUE		1	1000
10	DENGUE	WARANGAL	1	1000
10	Nausea		6	1500000
10	Nausea	Etowah	6	1500000
10	Sunburn		7	1750000
10	Sunburn	Mysore	7	1750000
10	Complicated UTI		6	1500000
10	Complicated UTI	Thrissur	6	1500000
11			1	500000
11		HANAMKONDA	1	500000
11	ZANDIS		1	500000
11	ZANDIS	HANAMKONDA	1	500000

Figure 4. Showing the output result of cube operation

In figure 4 it is found that the disease DENGUE has only record in the city Warangal. An intruder after getting this result can generate other combination of queries which can identify the identity of actual person.

Hence, after obtaining the OLAP query result, we identify of the combination of output which can violate the privacy of an individual. Fig. 5 below shows portion of the results where two separate records are found with count value 1.because name and patient id is unique it is completely suppressed. But as the place information when linked with other information which is not sensitive may reveal the identity of an individual, so the actual sensitive attribute location value is replaced with some default value related to the same hierarchy of location. Similarly, the age can be modified using additive perturbation. The additive perturbation adds small amount of noise [4] drawn from Gaussian distribution to the values under attribute Age. Finally the modified result is submitted to the client.

Disease	Gender	Age	Type of disease	Location	bill amount
Dengue	Male	14	Severe	ASDF	1000
hypersensitivity	Male	33	normal	ASDF	54500

Figure 5 The two specific records having count equal to1

Disease	Gender	Age	Type of disease	Location	billamount
Dengue	Male	16	Severe	INDIA	1000
hypersensitivity	Male	31	normal	INDIA	54500

Figure 6. The modified result to be submitted to the client for analysis

The ability to of an intruder to identify an individual will be difficult, because values under attribute location have been changed to next hierarchy. It will be complex to identify an individual at state level than at city level. Further, privacy is enhanced by modifying the age values. The modified results do not deviate from the desired summarized information and is valid for analysis purpose.

VI. CONCLUSION

On-Line Analytical Processing operation on data warehouse information enables analysts, knowledge workers to view their own organizational database from different point of view. As the decision support system integrated with data warehouse is utilized by number of users or groups, the possibility of retrieving personal information and re-identifying an individual increases. Hence, inspite of numerous advantages, OLAP operations [9] faces the privacy problem from an opponent who can deduce the private from OLAP query answers. Such type of privacy violation is more dangerous in case of medical, patient history, financial data bases which consists not only the statistical but also confidential information about an individual which should be known only to the privileged persons and relevant organization. Privacy can be protected by query restriction and by modifying the aggregate results before releasing it for analysis purpose. This paper proposes a framework which analyzes the OLAP query results, and identifies the record attribute values which one can discover the sensitive information stored in the cell and identify and individual's information. This model modifies the summarized information so that one cannot infer the private information stored in the cell.

REFERENCES

[1] Agrawal R, Srikant R (2000) Privacy-preserving data mining. In: Proc of the ACM SIGMOD conference on management of data, Dallas, TX, USA, pp 439–450

[2] L. Wang, D. Wijesekera, and S. Jajodia. (2003), OLAP Means On-Line Anti-Privacy, ISE Technical Report (2003).

[3] Chaudhuri S, Dayal U (1997) An overview of data warehousing and OLAP technology. SIGMOD Rec 26(1), pp 65–74.

[4] Kargupta H, Datta S, Wang Q, Sivakumar K (2003) On the privacy preserving properties of random data perturbation techniques. In: Proc of 2003 IEEE international conference on data mining. Melbourne,FL, pp 99–106

[5] Faloutsos C, Jagadish H, Sidiropoulos N (1997) Recovering information from summary data. In: Proc of the 1997 VLDB. Athens, Greece, pp 36–45

[6] R. Agrawal and R. Srikant, D. Thomas. Privacy Preserving OLAP. Proc. of SIGMOD 2005, Baltimore, Maryland, USA, 14-16 June, 2005, pp. 251-262.

[7] Yao Liu, Sam Y. Sung, Hui Xiong (2006) A cubic-wise balance approach for privacy preservation in data cubes, In proceedings of Information Sciences, pp .215–1240

[8] Torsten Priebe, Günther Pernul (2000) Towards OLAP SecurityDesign Survey and Research Issues. ACM. Proc. Third ACM ternationalWorkshop on Data Warehousing and OLAP (DOLAP 2000), McLean, VA, USA, pp. 33-40.

[9] Sam Y. Sung, Yao Liu, Hui Xiong, and Peter A. Ng (2006) Privacy preservation for data cube, Springer-Verlag London Ltd., Knowledge and Information Systems (2006) 9: pp.38–61