



Journal Homepage: [-www.journalijar.com](http://www.journalijar.com)

## INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/12593  
DOI URL: <http://dx.doi.org/10.21474/IJAR01/12593>



### RESEARCH ARTICLE

#### A MECHANISM FOR DETECTING PARTIAL INFERENCES IN DATA WAREHOUSES

Johnson Grâce Yénin Edwige<sup>1,3</sup>, Adepo Joel<sup>2</sup> and Oumtanaga Souleymane<sup>1</sup>

1. Laboratoire de Recherche en Informatique et Télécommunication (LARIT), Institut National Polytechnique Félix Houphouët-Boigny, Abidjan, Côte d'Ivoire.
2. Université Virtuelle de Côte d'Ivoire, Abidjan, Côte d'Ivoire.
3. Ecole Supérieure Africaine des Technologies de l'Information et de la Communication (ESATIC), Abidjan, Côte d'Ivoire

#### Manuscript Info

##### Manuscript History

Received: 15 January 2021

Final Accepted: 19 February 2021

Published: March 2021

##### Key words:-

Data Warehousing, Security,  
Confidentiality, Inference Control,  
OLAP Queries

#### Abstract

Data warehouses are widely used in the fields of Big Data and Business Intelligence for statistics on business activity. Their use through multidimensional queries allows to have aggregated results of the data. The confidential nature of certain data leads malicious people to use means of deduction of this information. Among these means are data inference methods. To solve these security problems, the researchers have proposed several solutions based on the architecture of the warehouses, the design phase, the cuboids of a data cube and the materialized views of multidimensional queries. In this work, we propose a mechanism for detecting inference in data warehouses. The objective of this approach is to highlight partial inferences during the execution of a multidimensional OLAP (Online Analytical Processing) SUM-type multidimensional query. The goal is to prevent a data warehouse user from inferring sensitive information for which he or she has no access rights according to the access control policy in force. Our study improves the model proposed by a previous study carried out by Triki, which proposes an approach based on average deviations. The aim is to propose an optimal threshold to better detect inferences. The results we obtain are better compared to the previous study.

Copy Right, IJAR, 2021,. All rights reserved.

#### Introduction:-

Business intelligence offers companies the means to increase their revenues through decision support tools such as data warehouses. Unlike traditional databases that provide transactional data management, the data warehouse is dedicated to read-only. The interest of the data warehouse lies in the history of the data over many years, thus allowing a decisional analysis linked to the evolution of companies. In addition, users of a data warehouse have role-based access [1] to certain re-sources and perform multi-dimensional queries for decision support. The execution of these queries is carried out according to the direct accesses authorized to the user. For example, financial analysts use these queries in their decision making and they are subject to an access policy to preserve the confidentiality of sensitive data outside of their prerogatives. However, confidentiality is not always respected, as malicious users use circumvention methods to gain access to restricted information. This creates a security problem that needs to be addressed.

**Corresponding Author:- Johnson Grâce Yénin Edwige**

Address:- Laboratoire de Recherche en Informatique et Télécommunication (LARIT), Institut National Polytechnique Félix Houphouët-Boigny, Abidjan, Côte d'Ivoire.

Security in data warehouses covers a wide range of areas, including security in the operation of the OLAP server. This has been one of the areas of interest of several researchers in recent years. Despite the precautions taken in terms of access control policy, a malicious user can deduct from a combination of queries information that is not allowed to be accessed [2]–[5]. This phenomenon called data inference in multi-dimensional queries requires security measures to improve data confidentiality in data warehouses. OLAP servers to this day still have a real difficulty in supporting inference control, although efforts are being made in research [6]–[11]. There are two types of inferences: precise inferences and partial inferences. In precise inferences, the aim is to deduce exact values from authorized queries. In partial inferences, it is a matter of inferring approximate values that can lead to the desired result. This work is performed in the context of detecting partial inferences.

Indeed, we propose an approach to detect partial inferences that improves the level of confidentiality in data warehouses. This work is based on [2] and addresses the problem of partial inference in the context of Sum-type queries by proposing an optimal threshold for each subset resulting from the query.

The rest of the paper is organized as follows: section 2 presents the state of the art of the work done on inference control in data warehouses; section 3 presents the proposed new threshold approach for partial inference control; in section 4 we perform an experiment and present and discuss the results obtained. A conclusion will conclude this work in section 5.

### **Literature survey:-**

The problem of data inference is addressed by several researchers [5], [12]–[15]. First present in traditional databases, it is automatically present in data warehouses. Inference control models are then put in place to protect sensitive data from unauthorized disclosure and to limit or even eliminate inference channels [15], [16].

### **Inference control in traditional databases:**

In [17], a database security mechanism is proposed, allowing the detection of inferences. It is a detection mechanism that uses the user's current query and the history of previous queries to alert in case of inference of potentially sensitive information. The data dependency is exploited to build a semantic inference model (SIM) that will be mapped to a Bayesian network to evaluate inference probabilities. The user will have the answer to his query according to a threshold based on the probabilities obtained. The limit of this approach is its applicability only to transactional databases. Also, it relies only on the dependency of the data.

The approach presented in [14] focuses on abstract databases to secure confidential information resulting from the combination of queries, prior knowledge, and dependencies of probabilistic data. The proposed inference control is based on the probabilistic logic programming language PROLOG from which the ANGERONA system was developed. It is a security mechanism that prevents the inference of information in the presence of probabilistic dependencies. However, it only takes care of precise inferences in the framework of probabilistic dependencies in the data.

Reference [13] proposes to respond to the needs of database audits with an inference control whose architecture requires the database server to consider traditional access control and the user platforms to apply inference control. Each user manages his inference control in a decentralized way. A generic protocol is established to formalize the interactions between the database server and the user platforms. This protocol also ensures the existence of inference control mechanisms whose security properties are formally proven. This decentralized management allows the database server to be decongested. However, as each user is obliged to take care of the inference control in a decentralized way, he can carry out DoS attacks in the database server.

To address that issue [5] has proposed another approach to inference control called private and self-applicable. that one is about preventing DoS attacks in databases. The approach forces the user to perform queries without inferences by applying an expensive inference control himself, or else a penalty is imposed if these constraints are violated. This penalty consists in depriving the user of access privileges to the database. Thus, the access control is incorporated into the inference control. A user can obtain the access key for the next query only if his current query is without inference. The purpose is to prevent the formation of inference channels. This approach is difficult in its application because for a query producing an unintended inference channel, a deprivation of access privilege is performed. In addition, data warehouses with their different configurations cannot be secured in the same way as databases.

### Inference Control in Data Warehouses:

Reference [18] presents an approach to inference control applicable in data cubes. Based on cardinalities obtained from cuboids, it consists in executing OLAP queries without inferences in summary data cubes without affecting the performance and availability of the OLAP server. The approach consists in decomposing a data cube into several non-compromising sub-cubes. Then the union of these sub-cubes is used to respond to user requests. The following conditions are used to define the non-compromise of a sum-only data cube:

- (i) The non-compromise of multidimensional aggregations can be reduced to that of unidimensional aggregations.
- (ii) Cuboids with a solid or dense core cannot be compromised.
- (iii) There is a narrow lower limit on the cardinality of a core cuboid so that it re-mains uncompromised.

This approach has a limit which is the rejection of all aggregations when a single sub cube is compromised. To address this limitation, [3] proposed an approach that applies to several types of queries such as Min, Max, and Sum. It is an inference control based on cardinality, which counts the number of cells that all queries have covered so far to determine whether a new query should receive a response. It is implemented during the materialization of the cuboids by calculating a cardinality  $t$  for each cell of the cuboids. The query restriction algorithm determines the cuboid corresponding to the query issued, then counts the number of cells  $t_0$  covered by this query. To finish, it compares the two cardinalities. The request is rejected if  $t \leq t_0$ . This approach only considers the precise inference and does not take into account all the dimensions of the data cubes.

The approach proposed in [16] reveals that the diversity of inference channels makes their resolution complex. The inference problem cannot be entirely solved, and existing technologies can only partially eliminate it. With the advent of big data, malicious users use their personal experience and knowledge to explore data. This proposed solution is based on an inference control module using a data dictionary containing prior knowledge, query history and the user's current query. The goal is to infer the result of the current query with the system knowledge. If it is possible for the user to infer information, then the result of the query is not returned.

Triki et al. [2] proposed an inference control approach for Min, Max and Sum queries. Inferences in Sum queries are resolved from the mean deviation, which includes the distribution of data around the average and resolving only partial inferences. Inferences in Min and Max queries are resolved with Bayesian networks for accurate inferences. However, no thresholding methodology has been proposed for the detection of partial inferences in Sum queries.

The basic hypothesis of their work is based on the composition of a data warehouse into several subsets of data that are homogeneous according to their structure. This subdivision is done by categorizing the data according to the dimensions specified in the Sum query. Each subset is noted:

$$S_i : \{(x_{ij}, n_{ij}) ; j = 1 \dots m_i \in \mathbb{N}\}, i = 1 \dots r \in \mathbb{N}; \quad (1)$$

With:

- $x_{ij}$  : the  $j^{\text{th}}$  measure of subset  $i$  of the data warehouse
- $r$  : the total number of subsets
- $m_i$  : the total number of measurements of subset  $i$
- $n_{ij}$  : the number of employees associated with the  $j^{\text{th}}$  measure of subset  $i$
- $\mathbb{N}$  : the set of natural integers

For each subset, the total number is determined by the Equation (2):

$$N_i = \sum_{j=1}^{m_i} n_{ij} \quad (2)$$

For each subset, the average is determined from the following Equation (3):

$$\bar{X}_i = \sum_{j=1}^{m_i} f_{ij} * x_{ij} \quad (3)$$

With  $f_{ij}$  : the frequency of each measurement in a subset.

From this average, the mean deviation is determined by the Equation (4):

$$\Delta \bar{X}_i = \sum_{j=1}^{m_i} f_{ij} * |x_{ij} - \bar{X}_i| \quad (4)$$

The term  $x'_{ij} = |x_{ij} - \bar{X}_i|$  represents the absolute value of deviations from the average. With the knowledge of the average, the contribution of each in the mean deviation makes it possible to predict the possibility of inferences [2]. This contribution is obtained by the Equation (5) :

$$\Delta x'_{ij} = f_{ij} * x'_{ij} = f_{ij} * (x_{ij} - \bar{X}_i) \quad (5)$$

Expressing Equation (5) in relation to the average gives an indicator (Equation (6)), expressed as a percentage, of the possibility of inferences following disclosure of the average [2].

$$\Delta x'_{ij} = \frac{f_{ij}}{\bar{X}_i} * (x_{ij} - \bar{X}_i) = f_{ij} * \left(\frac{x_{ij}}{\bar{X}_i} - 1\right) \quad (6)$$

### Proposed Approach:-

Our approach improves the detection of partial sum inferences. According to Triki et al [2], must be below a certain fixed threshold depending on several factors in the data warehouse for inference to occur. This threshold, set in the Triki example, was not obtained from a proposed methodology. However, setting a threshold for all the data in a warehouse concerned by the Sum query is not optimal. Indeed, since the Sum query is applied to subsets, each subset does not observe the same distribution. This results in standard deviations that vary from one subset to another. It should therefore be possible to establish a threshold for each subset observed.

It is consequently important to determine an optimal threshold above which no inferences can be made. To this end, we suggest a dynamic model to determine the threshold applicable to any subset of data. Indeed, each homogeneous subset  $S_i$  of the data warehouse is associated with an average, a variance, and a standard deviation. The study context requires us to work with a very large number of data. According to the central limit theorem [19], whatever the form of the distribution of a population with a very large population size is, its sampling has a Gaussian distribution. Its law tends towards a normal law. The estimators we will use in our approach are the average and the standard deviation. The standard deviation measures the dispersion of values in a statistical sample or probability distribution. It is defined by the square root of the variance or the root mean square of the deviations from the average. Consider the following Equation (7):

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{m_i} (x_{ij} - \bar{X}_i)^2}{m_i}} \quad (7)$$

With  $\sigma_i$  : the standard deviation of a subset

From all of the above and based on the distribution of the data around the average, the fluctuation interval for a Gaussian distribution gives a proportion of 68% of the values between  $\bar{X}_i - \sigma_i$  and  $\bar{X}_i + \sigma_i$ . Referring to equation (6), for  $x'_{ij} = \bar{X}_i + \sigma_i$ , we obtain the new threshold Equation (8):

$$sd_i = \frac{1}{m_i} * \left(\frac{\bar{X}_i + \sigma_i}{\bar{X}_i} - 1\right) = \frac{1}{m_i} * \frac{\sigma_i}{\bar{X}_i} \quad (8)$$

### Experimentations and results:

The results below come from the analysis of the wage data of the carData package in R[20]. These data were collected from a university in the United States concerning the salaries of assistant professors, associate professors and professors over nine academic months from 2008 to 2009. The simulations were performed with Rstudio software Version 1.1.463 on a 05 cores HP computer with 08 gigabytes of Ram memory.

### Analysis of model data:

The following **Table 1** presents the variables of our dataset and their type.

**Table 1:-** Structure of the Salaries dataset of the carData package of R.

| Variables     | Type         |
|---------------|--------------|
| Rank          | Qualitative  |
| Discipline    | Qualitative  |
| Yrs.since.phd | Quantitative |
| Yrs.service   | Quantitative |
| Sex           | Qualitative  |
| Salary        | Quantitative |

As an example of a query, for the rest of our work, we assume that a user wants to know the average salaries of staff with the same grade and discipline. Let the following query be used:

R = SELECT Rank, Discipline, COUNT(Salary), AVG(Salary) FROM CUBESALAIRES GROUP BY Rank,Salary  
This query categorizes all these teachers into several homogeneous subsets presented in **Table 2**.

**Table 2:-** Homogeneous subsets formed from the specified query.

| Si | Rank      | Discipline | Size | Average(salary) |
|----|-----------|------------|------|-----------------|
| S1 | AsstProf  | A          | 24   | 73935.54        |
| S2 | AssocProf | A          | 26   | 83061.12        |
| S3 | Prof      | A          | 131  | 119948.27       |
| S4 | AsstProf  | B          | 43   | 84593.91        |
| S5 | AssocProf | B          | 38   | 101276.39       |
| S6 | Prof      | B          | 135  | 133393.76       |

To highlight the dispersion of the data around the average in each subset, the coefficients of variation  $\sigma_i / \bar{X}_i$  are calculated and presented in **Table 3**. Indeed, the distributions of the data around the average in each Si are different. And it is on this basis that we determine the dynamic thresholds.

**Table 3:-** Coefficients of variation of the subsets.

| Si | Rank      | Standard deviation | Average   | Coef. Var (%) |
|----|-----------|--------------------|-----------|---------------|
| S1 | AsstProf  | 4627,31            | 73935.54  | 6,26          |
| S2 | AssocProf | 10835,93           | 83061.12  | 13,05         |
| S3 | Prof      | 27853,23           | 119948.27 | 23,22         |
| S4 | AsstProf  | 7064,79            | 84593.91  | 8,35          |
| S5 | AssocProf | 10102,04           | 101276.39 | 9,97          |
| S6 | Prof      | 25813,69           | 133393.76 | 19,35         |

The coefficient of variation (CV) is the ratio of the standard deviation to the average [21]. An increase of the coefficient of variation leads to a large dispersion around the average. The above table indicates that the subsets S1 and S4 designating AsstProf teachers have the lowest coefficients of variation. This could be explained by the fact that they are at the beginning of their career. The data for these subsets tend to be closer to the average salary. The subsets S5 and S2, which designate teachers of the grade AssocProf, have their average coefficients of variation. Those designating teachers of the grade Prof, have the highest coefficients of variation. Their salaries tend to deviate from the average salary. It is therefore necessary to set a threshold for each subset to have optimal results.

### Experience, Results, and Discussion:-

We compare the Triki indices to the proposed threshold and observe significant results presented in the rest of this section. We recall that inference occurs when there is an index in a subset below the proposed threshold. **Table 4** presents the rates of inference that can be made. The analysis of the TIR% (S.F) column indicates that there are high inferences in all subsets when a fixed threshold of 10% is considered. Such a query will be rejected by the system. However, the TIR% (S.D) column shows that considering the threshold proposed by our approach, there is a subset for which no inference occurs. The interest in reducing the inference rates lies in the ability to authorize certain requests more easily to avoid binding rejections. Although we want to prevent inferences from being made, if the control system returns enough probability of inferences, we will have to execute fewer Sum queries. This would lead

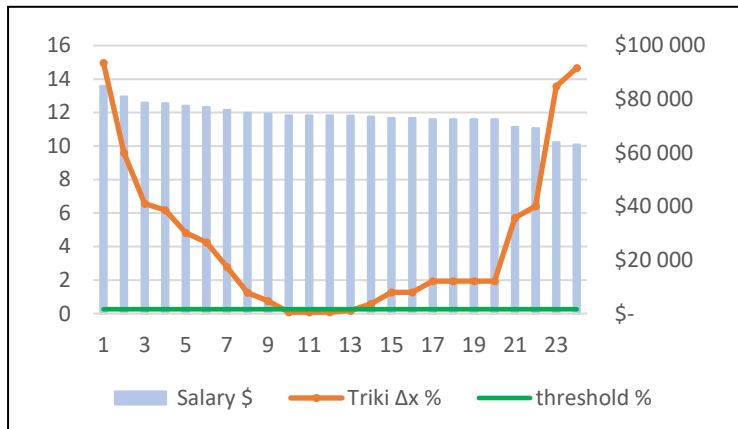
data analysts to be less efficient in their daily analyses. Proposing a threshold model that considerably reduces the inference channels to critical variables is advantageous. This reduction would be advantageous for the execution of a query.

**Table 4:-** Rates of inference observed by comparing Triki's thresholds to our approach.

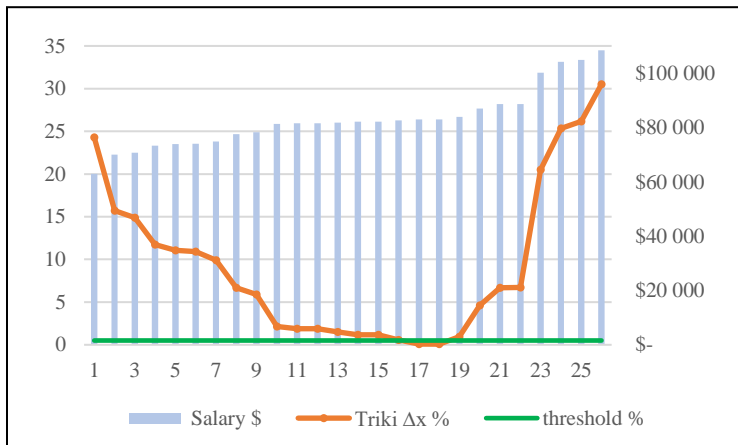
| Si | Threshold | C.V%  | TIR% (S.F) | TIR% (S.D) |
|----|-----------|-------|------------|------------|
| S1 | 0,2608    | 6,26  | 87,50      | 16,67      |
| S4 | 0,1942    | 8,35  | 81,40      | 4,65       |
| S5 | 0,2625    | 9,97  | 71,05      | 2,63       |
| S2 | 0,5018    | 13,05 | 61,54      | 7,69       |
| S6 | 0,1433    | 19,35 | 32,59      | 0,74       |
| S3 | 0,1773    | 23,22 | 28,24      | 0,00       |

with TIR (Realized Inference Rate) ; C.V (Coefficient of variation) ; S.F (Fixed threshold) ; S.D (Dynamic Threshold).

When coefficients of variation by subset are added, it is concluded that a subset with a high coefficient of variation has a high chance of producing no inferences. In this example, an increase in the coefficient of variation produces a decrease in the rate of inference. The following curves reveal that the threshold model we propose can detect the making of an inference. Each curve models the changes in the wage indices relative to the threshold obtained, for all subgroups.



**Figure 1:-** The indices  $\Delta x$  (model of Triki and al) in relation to the proposed threshold: subset S1.



**Figure 2:-**The indices  $\Delta x$  (model of Triki and al) in relation to the proposed threshold: subset S2

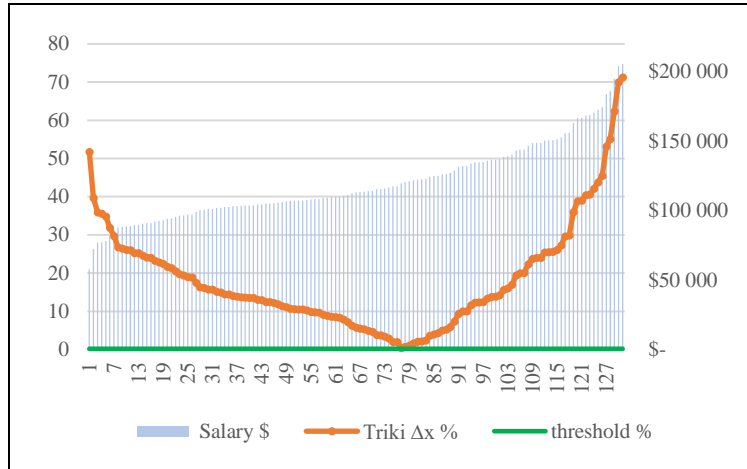


Figure 3:-The indices  $\Delta x$  (model of Triki and al) in relation to the proposed threshold: subset S3

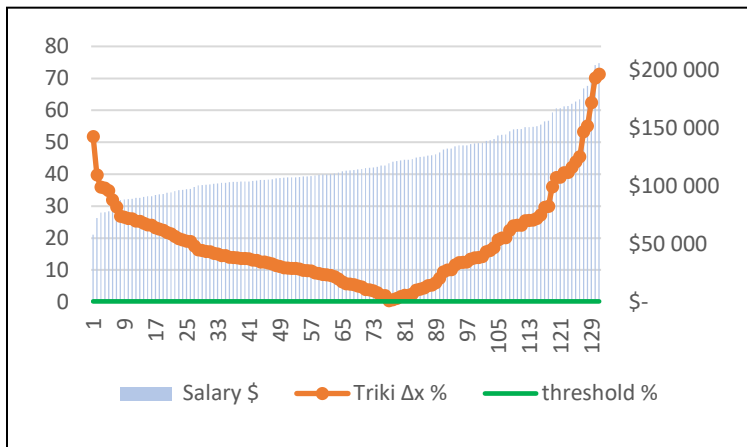


Figure 4:-The indices  $\Delta x$  (model of Triki and al) in relation to the proposed threshold: subset S4

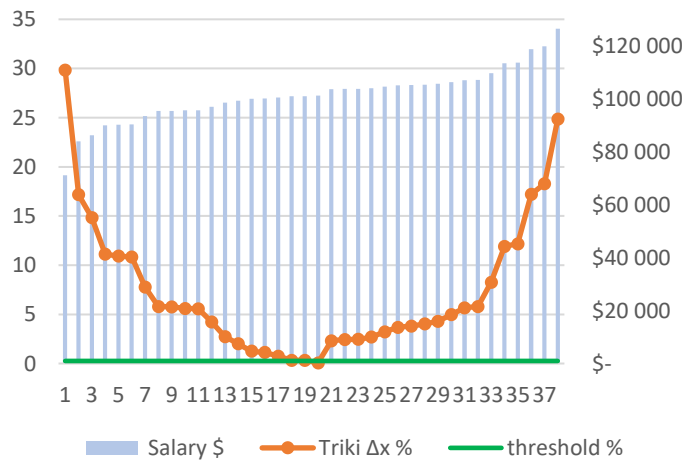
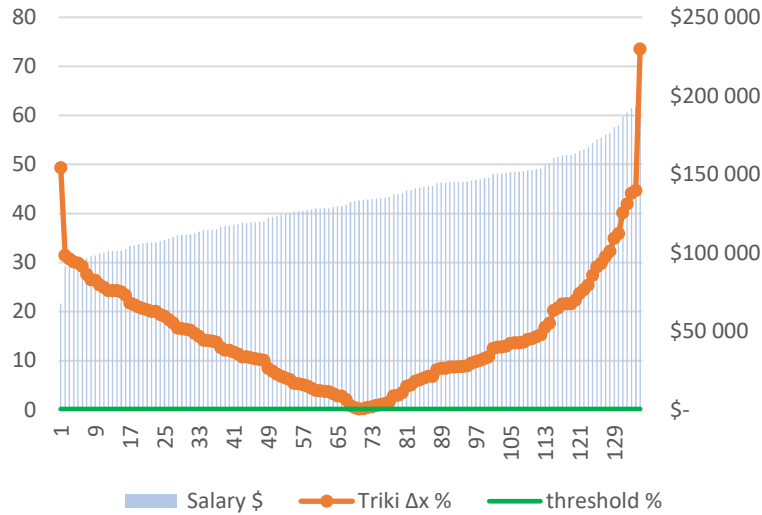


Figure 5:- The indices  $\Delta x$  (model of Triki and al) in relation to the proposed threshold: subset S5



**Figure 6:-**The indices  $\Delta x$  (model of Triki and al) in relation to the proposed threshold: subset S6

The orange curves designate the  $\Delta x$  indices from the model of Triki et al. The green curves designate the line representing the threshold. Those in blue in the form of a histogram represent wages. The  $\Delta x$  indices and the thresholds are expressed as percentages. Wages are expressed in dollars. We adopt an increasing representation of wages to show the trend of the indices in relation to the average.

Indeed, when the wage tends toward the average, the indexes move closer to the threshold curve because the average is an estimator of a wage for a given subset. We extracted the specific cases by subset that make inferences. For each of these sets, we note that the wages concerned are indeed very close to the average.

Table 5 highlights the critical wages that could be estimated from the average. Inferences are made only in these cases. Indeed, these are values very close to the average.

**Table 5:-** Critical Wages Producing Subset Inferences in Our Approach.

| Subset | Salary \$ | Triki $\Delta x$ % | Threshold % | Average   |
|--------|-----------|--------------------|-------------|-----------|
| S1     | 73800     | 0,18%              | 0,26%       | 73935,54  |
| S1     | 74000     | 0,09%              | 0,26%       | 73935,54  |
| S1     | 74000     | 0,09%              | 0,26%       | 73935,54  |
| S1     | 74000     | 0,09%              | 0,26%       | 73935,54  |
| S2     | 83000     | 0,07%              | 0,50%       | 83061,12  |
| S2     | 83001     | 0,07%              | 0,50%       | 83061,12  |
| S4     | 84500     | 0,11%              | 0,19%       | 84593,91  |
| S4     | 84716     | 0,14%              | 0,19%       | 84593,91  |
| S5     | 101210    | 0,07%              | 0,26%       | 101276,39 |
| S6     | 133217    | 0,13%              | 0,14%       | 133393,76 |

**Conclusion:-**

The objective of this study was to show that security in data warehouses against partial inferences of the Sum type can be achieved, by using the statistical method of the mean deviation, but also by proposing an adapted and optimal threshold. We started from the work of Triki et al [2] by showing the need to define a threshold per subset of data obtained. The results obtained indicate that considering the coefficient of variation and the integration of the standard deviation in the modeling of the threshold are efficient as solutions. Their use in our model contributes to better results. In our future work, we are planning to take into account one or more pieces of knowledge to propose a new approach of index computation adapted.



**References:-**

1. El Ouazzani, A., Rhazlane, S., Harbi, N. and Badir, H. (2016) Dynamic Management of Data Warehouse Security Levels Based on User Profiles. 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt). <http://dx.doi.org/10.1109/cist.2016.7804961>
2. Triki, S., Ben-Abdallah, H., Feki, J. and Harbi, N. (2011). Sécurisation des entrepôts de données contre les inférences précises et partielles. *Ingénierie des systèmes d'information*, 16(6), 117–138. <http://dx.doi.org/10.3166/isi.16.6.117-138>
3. Zhang, N., Zhao, W., and Chen, J. (2004) Cardinality-based inference control in OLAP systems. Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP - DOLAP '04. <http://dx.doi.org/10.1145/1031763.1031776>
4. Wang, L., and Jajodia, S. (2008). Security in Data Warehouses and OLAP Systems. *Handbook of Database Security*, 191–212. [http://dx.doi.org/10.1007/978-0-387-48533-1\\_8](http://dx.doi.org/10.1007/978-0-387-48533-1_8)
5. Yang, Y., Li, Y., Weng, J., Zhou, J., and Bao, F. (2009) Self-enforcing Private Inference Control. *Lecture Notes in Computer Science*, 260–274. [http://dx.doi.org/10.1007/978-3-642-04642-1\\_21](http://dx.doi.org/10.1007/978-3-642-04642-1_21)
6. Bertino, E., & Cuzzocrea, A. (2009). Privacy Preserving OLAP over Distributed XML Documents. 2009 International Conference on Parallel Processing Workshops. <http://dx.doi.org/10.1109/icppw.2009.88>
7. Cuzzocrea, A., and Saccà, D. (2010) Balancing accuracy and privacy of OLAP aggregations on data cubes. Proceedings of the ACM 13th International Workshop on Data Warehousing and OLAP - DOLAP '10. <http://dx.doi.org/10.1145/1871940.1871960>
8. Cuzzocrea, A., Bertino, E. and Saccà, D. (2012) Towards a theory for privacy preserving distributed OLAP. Proceedings of the 2012 Joint EDBT/ICDT Workshops on - EDBT-ICDT'12. <http://dx.doi.org/10.1145/2320765.2320826>
9. Popeea, T., Constantinescu, A., Gheorghe, L., and Tapus, N. (2012) Inference Detection and Database Security for a Business Environment. 2012 Fourth International Conference on Intelligent Networking and Collaborative Systems. <http://dx.doi.org/10.1109/incos.2012.95>
10. Elouazzani, S. Ouederrou, S. Ibn El Ahrache, N. Harbi, H. Badir, and S. el Ibn Ahrache, "Preventive inference control between user profiles in data warehouse," 2019. Accessed: Jun. 14, 2020. [Online]. Available: <https://hal.archives-ouvertes.fr/hal-02054423>
11. Priebe, T., and Pernul, G. (2000) Towards OLAP security design --- survey and research issues. Proceedings of the 3rd ACM International Workshop on Data Warehousing and OLAP - DOLAP '00. <http://dx.doi.org/10.1145/355068.355313>
12. Dwyer, C. A. (2009) The Inference Problem and Pervasive Computing. SSRN Electronic Journal. <http://dx.doi.org/10.2139/ssrn.1508513>
13. Yang, Y., Li, Y., and Deng, R. H. (2007) New Paradigm of Inference Control with Trusted Computing. *Lecture Notes in Computer Science*, 243–258. [http://dx.doi.org/10.1007/978-3-540-73538-0\\_18](http://dx.doi.org/10.1007/978-3-540-73538-0_18)
14. Guarnieri, M., Marinovic, S., and Basin, D. (2017) Securing Databases from Probabilistic Inference. 2017 IEEE 30th Computer Security Foundations Symposium (CSF). <http://dx.doi.org/10.1109/csf.2017.30>
15. Farkas, C., and Jajodia, S. (2002) The inference problem. *ACM SIGKDD Explorations Newsletter*, 4(2), 6–11. <http://dx.doi.org/10.1145/772862.772864>
16. Zhao, H. Y., Meng, J., and Zhang, X. (2014) Research on Multilevel Secure Database Inference Channel. *Applied Mechanics and Materials*, 644-650, 3310–3313. <http://dx.doi.org/10.4028/www.scientific.net/amm.644-650.3310>
17. Yu Chen, and Chu, W. W. (2008) Protection of Database Security via Collaborative Inference Detection. *IEEE Transactions on Knowledge and Data Engineering*, 20(8), 1013–1027. <http://dx.doi.org/10.1109/tkde.2007.190642>
18. Wang, L., Wijesekera, D., and Jajodia, S. (2004) Cardinality-based inference control in data cubes\*. *Journal of Computer Security*, 12(5), 655–692. <http://dx.doi.org/10.3233/jcs-2004-12501>
19. Fischer, H. (2010). Conclusion: The Central Limit Theorem as a Link Between Classical and Modern Probability Theory. *A History of the Central Limit Theorem*, 353–362. [http://dx.doi.org/10.1007/978-0-387-87857-7\\_8](http://dx.doi.org/10.1007/978-0-387-87857-7_8)
20. Fox, J., and Weisberg, S. "An R Companion to Applied Regression, Third Edition. Sage, Thousand Oaks CA" 2019. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>
21. Brown, C. E. (1998) Coefficient of Variation. *Applied Multivariate Statistics in Geohydrology and Related Sciences*, 155–157. [http://dx.doi.org/10.1007/978-3-642-80328-4\\_13](http://dx.doi.org/10.1007/978-3-642-80328-4_13).