ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Database Security Approach for Distributed Datasets: A Survey

Kalpana K. Palve, Prof R.W. Deshpande

Department of Computer Science and Engineering, Siddhant College of Engineering, Sudumbare, Pune University, Pune, MH, India.

ABSTRACT: Now a days there isa require of data characteristic security in disseminated database while preserving solitude. In the proposed work, we judge problem connected in publishing mutual data for anonymizingperpendicularly and parallel partitioned data. We deem the attack which might use a subset of the in general data. After in view of entire investigate work we formulate the distributed database classification in which first, we pioneer the notion of data solitude which guarantee the seclusion of anonymized data for dissimilar data contributor. Second, we current algorithms for exploiting the monotonicity of confidentiality constraints for checking data privacy professionally with the encryption representation using encryption algorithm. Third, we distribute the data to end user with the anonymization as well as security algorithm, and inspection the verification schema with TTP, which will give the guarantee to present high level safety to database. Experiments we use the hospital enduring datasets suggest that our advance achieves improved or comparable usefulness and competence than existing and baseline algorithms while fulfilling of proposed sanctuary work**.**

KEYWORDS: Distributed database, privacy, protection, security, SMC, TTP

I. INTRODUCTION

Privacy conservation techniques are mainly used to decrease the leakage of configuration about the particular creature while the data are shared and released to community. For this, the sensitive in sequence should not disclose. Data is getting modified first and then published for further procedure. For this a variety of anonymization method are followed and they are simplification, repression, permutation and perturbation. By various anonymization techniques data is modified which retain sufficient utility and that can be unconfined to other parties securely. Single association does not hold the absolute data. organizationrequire to share data for mutual remuneration or for publishing to a third gathering. For banking division want to integrate their consumer data for developing a scheme to provide improved services for its customers. However, the banks do not desire to indiscriminately reveal their data to every other for reason such as solitudedefense and commerce competitiveness.

Main objective is to publish an anonymized view of incorporated data, T, which will be resistant to attacks(fig1). Attacker runs the assault, i.e. a single or a group of exterior or internal entities that requirements to breach privacy of data using environment knowledge. mutual data publishing is carried out fruitfully with the help of trusted third party (TTP) or safe Multi Party Computation (SMC) protocols, which guarantee that information or data regarding particular creature is not disclosed wherever, that means it maintains confidentiality. Here it is unspecified that the data providers are partially honest. A more desirable advance for mutual data publishing is, first comprehensive then anonymize (fig 1)[1].

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

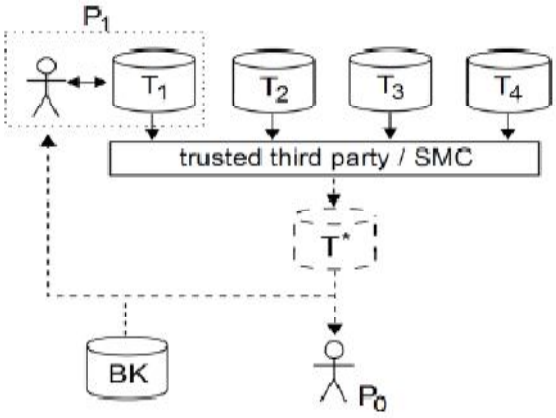

Fig1: Aggregate and Anonymize

In above diagram, T1,T2,T3 and T4 are database for which data is provided by supplier like provider P1 provide data for catalog T1. These distributed information coming from dissimilar providers get collective by TTP(trusted third party) or with SMC protocol. Then these aggregated data anonymizedadditional by any anonimizationmethod. P0 is the authenticate client and P1 trying to violate privacy of data which is provided by additional users with the assist of BK(Background knowledge). This kind of attack we preserve call as a "insider attack". We contain to defend our system from such a type of attack

II. LITERETURE SURVEY

We are studying dissimilarmethod which arebeforehand used for anonymization. We learningsolitude preserving data publish (PPDP) [2]and LKC[3] model gives improvedconsequence than conventional k anonymization model. And as well Two party protocol DPP2GA[4], It is solitary privacy preserving protocol not SMC since it introduce certain inference difficulty. Many system use k anonymization for provided that privacy. Attacker can assault on anonymizedscheme with the help of BK(background knowledge). L diversity helps to surmount this problem. In present research document [1], authors bring ina m privacy algorithm which confirm anonymization and L diversity. For this they believe generalization and bucketizationtechnique for maintaining anonimizedvision of data and also offer L assortment which help to increase solitude of data. This manuscript proposed a system in which we used a new knowledgei.e slicing algorithm with which we moreover used encrypted data which improves precautions. Slicing is the procedure which gives improved result than characteristicsimplification and bucketizationmethod. It gives better results for high dimentional data. It can perform permutation within bucket. In slicing we can pool resources sensitive attribute with some quasi identifier. On this sliced data we utilizeconfirmationalgorithms[1] which verifies that whether information is secured or not.

III. RELATED WORK

Due to dissimilar attacks attackers can assault on our system. For our scheme we consider certain insider attack like backdrop knowledge attack. Privacy fortification is impracticable due to the occurrence of the adversary's environment knowledge [6]. Second is relationship attack in which when an opponent is able to link a record possessor to a record in a published information table called record connection, to a sensitive quality in a published data table called quality linkage, or to the published data table itself called bench linkage. In this attack adversary may be acquainted with some victims data like QID etc. In some cases supplier himself can be an assailant. His own evidence which might be a subset of database. Maintaining safekeeping and isolation of article without with encryption have been a challenging difficulty in distributed arrangement. Various method and strategies are developed to construct maximum likelihood to make it probable. To conquer these tribulations we proposed a system. Problem meaning: Our main goal is to circulate an anonymized view of incorporated data, P* which will be resistant to attacks. We recover the security and

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

confidentiality with the help of slicing method, data privacy confirmation algorithm and protected data analysis with the assist of classifier.

Goals and objective

- ❖ We make use of slicing algorithm which gives improved result than generalization as well asbucketisation
- ❖ Binary algorithm apply to verify data privacy for each section of data with pruning strategy.
- ❖ Check EG monotonic used for checking the seclusion of equivalent group.
- ❖ We have to decrease estimation time of system.

Scope

- ❖ Proposed scheme is run on LAN system.
- ❖ Distributed system like infirmary patient data management, companiesemployer salary organization, banking system like individual information of bank account holders etc where we require to secure collaborative data.

IV. SYSTEM ARCHITECTURE

We first formally describe our problem setting. Then, we present our data-privacy definition with respect to a privacy constraint to prevent inference attacks by data-adversary, followed by properties of this new privacy notion. Let T = {t1, t2, . . .}be a set of records with the same attributes gathered from n data providers P ={P1, P2, . . . , Pn}, such that Ti are records provided by Pi. Let AS be a sensitive attribute with a domain DS. If the records contain multiple sensitive attributes then, we treat each of them as the sole sensitive attribute, while remaining ones we include to the quasi-identifier [10]. However, for our scenarios we use an approach, which preserves more utility without sacrificing privacy [11]. Our goal is to publish an anonymizedT* while preventing any data-adversary from inferring AS for any single record. An data-adversary is a coalition of data users with n data providers cooperating to breach privacy of anonymized records. When data are gathered and combined from different data providers, mainly two things are done, for anonymization process. To protect data from external recipients with certain background knowledge BK, we assume a given privacy requirement C is defined as a conjunction of privacy constraints: C1∧C2∧...∧Cw. If a group of anonymized records T* satisfies C, we say C(T*)=true. By definition C(Ø) is true and Ø is private. Any of the existing privacy principles can be used as a component constraint Ci. We now formally define a notion of data-privacy with respect to a privacy constraint C, to protect the anonymized data against data-adversaries. The notion explicitly models the inherent data knowledge of an data-adversary, the data records they jointly contribute, and requires that each QI group, excluding any of those records owned by an data-adversary, still satisfies C.

In the proposed system in which input data is given in encrypted format(attribute name will be in encrypted format). Select point for slicing[12]. Check that input data against privacy constraint C for data privacy. Check further is slicing is possible or not. If slicing possible then do it and if not then decrypt data. Our final output T* are anonymized data which will seen only by authenticate user. Any adversary can not breach privacy of data. In this system we are using horizontal as well as vertical partitioning over database. Slicing algorithm provide better column partitioning. To understand this properly lets consider hospital management system for experiment. Let different departments are the providers who provides data from different sources. We consider disease as a AS (sensitive attribute) and age and zipcode are QI(quasi identifier).

Algorithms
Slicing Algorithm:

Definition 1: (Attribute separation and Columns).

In attribute separation, D(database) consists of several subsets, such that each attribute belongs to exactly one subset. Each subset of attributes is called a column. Specifically, let there be C columns C1; C2; . . . Cc, then U(c)i=1,C=D; and for any $1 \leq i_1 \neq i2 \leq c$, $Ci_1 \cap Ci2 = \Theta$. For simplicity of discussion, we consider only one sensitive attribute S. If the data contain multiple sensitive attributes, one can either consider them separately or consider their joint distribution [25]. Exactly one of the c columns contains S. Without loss of generality, let the column that contains S be the last

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

column C. This column is also called the sensitive column. All other columns{C1,C2…… Cc-1} contain only QI attributes.

Definition 2 : (Tuple Partition and Buckets).

In tuple partition, T consist of several subsets, such that each tuple belongs to exactly one subset. This tuples subset is called a bucket. Specifically, let there be b buckets. B1,B2……Bb then U^{b}_{i}-ıB$_i$=T and for any 1≤ ıı≠i2 ≤b, Bıı∩Bi2= Ɵ

Definition 3 (Slicing):

Given a microdata table T, a slicing of T is given by an attribute partition and a tuple partition.
For example, suppose tables a and b are two sliced tables. In Table a, the attribute partition is {{Age}, {Gender}, {Zipcode}, {Disease}} and the tuple partition is {{t1; t2; t3; t4}, {t5; t6; t7; t8}}. In Table b, the attribute partition is {{Age, Gender}, {Zipcode, Disease}} and the tuple partition is {{t1; t2; t3; t4}, {t5; t6; t7; t8}}.

Definition 4 (Column Generalization)

Given a microdata table T and a column Ci = (Xi1,Xi2,Xi3…….Xij) where Xi1,Xi2…..Xij are attributes, a column generalization for Ci is defined as a set of non overlapping j-dimensional regions that completely cover D[Xi1]* [Xi2] * ….. D[Xij] . A column generalization maps each value of Ci to the region in which the value is contained.
Column generalization ensures that one column satisfies the k-anonymity requirement. It is a multidimensional encoding and can be used as an additional step in slicing. Specifically, a general slicing algorithm consists of the following three phases: attribute partition, column generalization, and tuple partition. Because each column contains much fewer attributes than the whole table, attribute partition enables slicing to handle high-dimensional data. A key notion of slicing is that of matching buckets.

Definition 5 (Matching Buckets)

Consider sliced data and let (C1; C2; . . . ; Cc) be the columns. Let t be a tuple, and t[Ci] be the value of Ci of t. Let B be a bucket in the sliced table, and B[Ci] be the multiset of Ci values in B. We say that B is a matching bucket of t if for all T[C(i)]==B[C(i)] and 1 set if I set of C, 1≤i≤c,t[Ci]ЄB[Ci]
By using above slicing algorithm we can obtain anonymization and l diversity both. This two technique maintains he privacy of data.
Binary algorithm:

Data: Anonymize records DATA from providers P, an EG monotonic C, a fitness scoring function score F , and the n.
Result: if DATA is private secure C then True, else false
1. sites = sort_sites(P, increasing order, scoreF)
2. Apply slicing
3. while verify data-privacy(DATA, n, C) = 0 do
4. super = next_instance size(n− 1)&& (size_of_tupples (Σ) // identification of column
5. if privacy breached_by(Psuper, C) = 0 then
6. prune_all_sub-instances_downwards(Psuper)
7. continue
8. Psub = next_sub-instance_of(Psuper,n)
9. ifprivacy_is_breached_by(Psub, C) = 1 then
10. return 0 // early stop
11. whileinstance_between(Psub, Psuper) do
12.I = next_instance between(Psub, Psuper)
13. if privacy breached_by(P,C) = 1 then
14.Psuper = P
15. else

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

16. Psub = P
17. prune_all_sub-instances_downwards(Psub)
18. prune_all_super-instances_upwards(Psuper)
19.return 1

V. CONCLUSION

We consider a potential attack on collaborative data publishing. We used slicing algorithm for anonymization and L diversity and verify it for security and privacy by using binary algorithm of data privacy. Slicing algorithm is very useful when we are using high dimensional data. It divides data in both vertical and horizontal fashion. Due to encryption we can increase security. But the limitation is there could be loss of data utility.

Above system can used in many applications like hospital management system, many industrial areas where we like to protect a sensitive data like salary of employee. Pharmaceutical company where sensitive data may be a combination of ingredients of medicines, in banking sector where sensitive data is account number of customer, our system can use. It can be used in military area where data is gathered from different sources and need to secured that data from each other to maintain privacy. This proposed system help to improve the data privacy and security when data is gathered from different sources and output should be in collaborative fashion. In future this system can consider for data which are distributed in ad hoc grid computing. Also the system can be consider for set valued data.

VI. FUTURE WORK

We can implement the proposed architecture on hadoop base system with cube materialization and map reduce.
Also we can identify a subset of holistic measures that are partially algebraic and propose the technique of value partitioning to make them easy to compute in parallel. Design algorithms that partition the cube lattice into batch areas to effectively exploit both the parallel processing power of MapReduce and the pruning power of cube materialization algorithms. Further, and demonstrate the ability to surface interesting cube groups as part of the cube computation process. Experiments over real and synthetic data show that our MR-Cube algorithm efficiently distributes the computation workload across the machines and is able to complete cubing tasks at a scale where prior algorithms fail.

REFERENCES

[1] S. Goryczka, L. Xiong, and B. C. M. Fung, “m-Privacy for collaborative data publishing,” in Proc. of the 7th Intl. Conf. on Collaborative Computing: Networking, Applications and Worksharing, 2011.
[2] C.Dwork,”Differential privacy: A survey of results”, in Proc. of the 5th Intl. Conf. on Theory and Applications of Models of Computation, 2008, pp. 1
[3] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, “Centralized and distributed anonymization for high-dimensional healthcare data,” ACM Trans. on Knowledge Discovery from Data, vol. 4, no. 4, pp. 18:1–18:33, October 2010.
[4] W. Jiang and C. Clifton, “Privacy-preserving distributed k-anonymity,”inDBSec, vol. 3654, 2005, pp. 924–924.
[5] W. Jiang and C. Clifton, “A secure distributed framework for achieving k-anonymity,” The VLDB Journal Special Issue on Privacy Preserving Data Management, vol. 15, no. 4, pp. 316–333, 2006
[6] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam,“l-Diversity: Privacy beyond k-anonymity,” in ICDE, 2006,p. 24
[7] R. Sheikh, B. Kumar, and D. K. Mishra, “A distributed k-secure sum protocol for secure multi-party computations,” J. of Computing, vol. 2, pp. 68–72, March 2010 (2002)
[8] C. Dwork, “A firm foundation for private data analysis,” Commun. ACM, vol. 54, pp. 86–95, January 2011
[9] P. Jurczyk and L. Xiong, “ Distributed anonymization: Achieving privacy for both data subjects and data providers,” in DBSec, 2009, pp. 191–207
[10] C. M. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy-preserving data publishing: A survey of recent developments,” ACM Comput. Surv., vol. 42, pp. 14:1–14:53, June 2010.