www.ijiarec.com

ISSN: 2348-2079 MAR-2015

International Journal of Intellectual Advancements and Research in Engineering Computations

PRIVACY PRESERVATION IN BIG DATA USING ENHANCED TOP-DOWN SPECIALIZATION APPROACH

[1]S.Sivasankar, [2]T.Prabhakaran

ABSTRACT

Data size has grown a large in present years by the development of internet in large manner so that big data era arrives, with the cloud computing users able to store large amount of data in ease manner. Users now use both thestructured data and as well as unstructured data.In big data due to its large size all the tasks are consuming more amount of time. The internet users also share their private data like health records and financial transaction records for mining or data analysis purpose during that time data anonymization is used for hiding identity or sensitive intelligence so that data owners do not suffer with economical loss. Anonymizing large scale data within a short span of time is a challenging task to overcome that Enhanced Top – Down Specialization approach (ETDS) can be developed which is an enhancement of Two – Phase Top Down Specialization approach (TPTDS).

Key words: BigData, Cloud Computing, Data anonymization, ETDS, TPTDS.

INTRODUCTION

Data which contains the large volume, velocity and variety can be called as the big data. Cloud computing provides large amount of computation power and repository capacity which provides users to perform computation and data intensive applications without any cost to be spend for the infrastructure. Privacy of the user should be maintained in a good manner so only users can come forward to use our service and in privacy it includes with the identity preservation of the user, transaction history and during integration components policy to be protected with the cloud computing single work is shared with the lot of computers at that time privacy is to be ensured for the user [4]. In cloud computing security is measured by trust, confidentiality, integrity and availability. Trust means that providing user required service in correct and without any fault. Confidentiality means that only authorized users can access the protected data within cloud a large number of users, applications and devices are present so monitoring the access of the data is important. The different confidentialities are data confidentiality and software confidentiality, data confidentiality is achieved by user authentication. Integrity deals with the assets so that assets can be able to modify only by the users provided with the authority in integrity also data integrity and software integrity is present. Finally availability means that the data, software or resource which is requested by the user should be available in immediate manner to the user [5]. Security is to be provided for the intermediate data sets by the process of encrypting the data set instead of encrypting all the data sets privacy leakage upper bound approach identifies which data set must perform with the encryption, instead of encrypting all the data sets this approach saves the cost and time for the user. In the context of personal data privacy preservation is very important during the publication of personal data. There are simple and various kinds of techniques are present for the privacy preservation one such technique is the anatomy it is used for publishing sensitive intelligence or data to achieve privacy preservation quasi identifier and sensitive data are to be stored in table then by the grouping mechanism correlation is done with the data [13].

Private intelligence of the data owner should maintained with security so that only data owners does not suffer with the economical loss, in

Author for Correspondence:
[1]PG Scholar, Department of CSE,Nandha Engineering College, Erode, India, Email: Sivasankar550@gmail.com.
[2]Assistant Professor, Department of CSE, Nandha Engineering College,Erode, India. Email: t.prabhakaran@gmail.com.

the health related data several disease research centre collect the intelligence of the user then it analyze and mine those data without any prior permission from the user so the identity of the data to be hided with the data anonymization in that one suppression and generalization techniques are present in this Extended Top- Down Specialization approach generalization technique is used. Big data processing framework is very important in which knowledge comes from numerous and different sources.

K – Anonymity is the technique for performing anonymization in data but it does not have much flexibility, so to achieve flexibility Mondrian multidimensional K – Anonymity is proposed which helps for achieving more quality during the time of anonymization takes place [15]. Big Data management is the important because the data management handled in effective manner then only more number of data can be stored the architecture for the big data is mainly concentrated in that one particularly components and layers are focused and the way in which data can be used [16]. The anonymization technique which is extended for scalable and incremental data to be anonymized is discussed in this spatial indexing technique is used for performing the anonymization with the scalable data. There are techniques and tools are available for the big data processing, MapReduce [19] is the programming model used for processing large scale data set, it is combined with the cloud to provide capable computation for applications and it is also used for generating large scale data set and the program which are written in this are parallelized in automatic manner. Privacy protection is also performed during the analysis of cluster for this purpose also anonymization is performed, it is very important to perform anonymization during the time of information sharing. A framework or predefined structure is very much needed for performing k-Anonymity in a perfect manner data security is also to be provided both for the data users and data providers and it is taken care by the distributed anonymization which is performed by several protocol set which are working in a distributed way [25].New computing platform is developed for performing operation with the large scale data during the time of data mining, graph analysis, model fitting and graph ranking. HaLoop [31] is the computing platform developed it overcomes the disadvantages of both Hadoop and MapReduce. In MapReduce the problem is that it does not support for the iterative programs. HaLoop allows the previously present applications it is present in Hadoop without any changes and also advances the performance by giving new inter iteration caching mechanism and scheduler has provide the user with advantages like automatic cache recovery and task re execution.

RELATED WORKS

Privacy preservation is very much important because it is the key concept while sharing the data. The issue with dealing large amount of data for the anonymization algorithm is considered in that one first is with the decision trees and sampling methods. Jiang and Clifton [24] introduced distributed algorithm which particularly perform the task with vertically separated data from various data origins and it is performed without attaching the private knowledge of one user to the other. The next is the distribution algorithm to hide the identity of data from horizontally separated data this approach mainly focus on the careful joining of large number of data sources and also hiding the data information. There are several Top-Down Specialization approaches [TDS] are present in which the first one is centralized TDS approaches and the next one is distributed TDS approach in the Centralized approach work is performed with centralize manner whereas in distributed there is no central control is present and the task is performed in distributed way .

CENTRALIZED TDS APPROACHES

Centralized TDS approaches which does not obey with the data structure of Taxonomy Indexed Partitions [TIPS] to increase the size of the data and capability for that unidentified data records and also keeping the unchangeable intelligence present in the TIPS. The data structure present makes the specialization to be fastest because indexing structure withdraws usually examining all the data sets and saving demographical outcome. The second is, the measure of metadata attainedto retain the demographical intelligence and linkage intelligence of document separations is approximately enormous measured among data sets themselves,

thereby consuming reasonable repository. In addition, the overheads acquired by controlling the interconnection architectureand updating the demographicintelligencewill be enormous at the time of data sets become massive. So in this waythere is a chance forcentralized approaches to suffer during the time of handling large scale data sets. The problems faced by centralized approaches are small performanceand scalability.

There is an assumption that all data measures shouldadapted in repositoryfor the centralized approaches [12]. Unfortunately, this expectationgenerally misses to hold in mostdata comprehensive cloud applications nowadays. In cloud environments, calculation is provisioned in the pattern of virtual machines (VMs). Usually, cloud compute services offer several types of VMs to the user. This makes, the centralized approaches are difficult in handling large-scale data sets on cloud it is also find to impossible when single VM alone is used but it performs the maximum computation and repository capacity.

DISTRIBUTED TDS APPROACH

To overcome the problems present in the centralized approaches a new approach namely distributed TDS approach is recommended. Distributed TDS approach [20] is to overcome the distributed anonymization challenge which particularly focuses on privacy protection against other parties, instead of focusing on large scale data or scalability problems. Moreover, this approach only employs knowledge gain, rather than its connection withprivacy loss, as the search criteria when finding out the perfect specializations. As indicated previously, a TDS algorithm without focusing the privacy loss will probably chooses the specialization that directs to a fast problem with theanonymity needs. In this way the distributed algorithm fails toproduce data sets which are made anonymous showing the identical datausage as centralized algorithm perform. The conclusion is observed that existing distributed algorithmsare not enough tosolve the scalability or the size of dataissue of TDS.

PRELIMINARY

BASIC NOTATIONS

Basic notations are described in this data set occupies the important place and it is denoted by D. Next is the record it is denoted by r and record belongs to the data set, record r has the form a set of values starting from value1 to value m where m used to denote the number of attributes. Sensitive value is denoted by sv. Attr is used to denote the attribute of record where as its taxonomy tree is denoted by TT and finally domain values is represented by DOM. The quasi identifier is denoted by qid, a group of anonymous records is represented by the quasi identifier.

Table 1: Notations

Notation	Description
D	Data set
R	Record
m	Number of attributes
sv	Sensitive value
Attr	Attribute of record
TT	Taxonomy Tree
DOM	Domain values
QID	Quasi-identifier set

TOP-DOWN SPECIALIZATION

Top-Down Specialization is the continuous process which starts from the top values and ends with the low values. The values taken into here are domain values in the taxonomy trees of attributes. TDS process includes with the iteration and this iteration contains three main steps are performed they are finding the best specialization, performing specialization and updating the values of the search metric. Information gain per privacy loss (IGPL) is a measurement it considers both privacy and knowledge requirements. The highest value obtained from IGPL is taken as the special and it is used for each process.

TWO-PHASE TOP-DOWN SPECIALIZATION APPROACH

Two-Phase Top-Down Specialization (TPTDS) approach contains the two stages the first stage is the job level and the next stage is the task level in both the stages of the approach parallelization is achieved. Data partition, anonymization level merging and data specialization are the components of TPTDS. The basic plan of TPTDS is to achieve great scalability by creating a compromise between scalability and data utility.

ENHANCED TOP-DOWNSPECIALIZATION APPROACH

Enhanced Top-Down Specialization (ETDS) approach is the enhancement of existing Two-Stage Top-Down Specialization (TPTDS) approach in which privacy preservation can not be achieved when the scalability occurs. In this ETDS technique after the data partition, separated or partitioned data are clustered so that more number of data or scalable data can be protected with privacy. ETDS approach has all the components of TPTDS and the additional component is the clustering of partitioned data. In this approach also present with the job level and task level.

ALGORITHM: ENHANCED TOP-DOWN SPECIALIZATION (ETDS)

Input: Data set D, anonymity parameters k and number of partitions p

Output: Anonymous data set D^*

1: Partition D into D_i, 1<=i<=p

2: Execute ETDS

3: Merge anonymization levels into one

4: Specialize D and output D^*

The two levels are job level and task level and in both these levels works are carried out in concurrent manner or parallelization are attained. Job level parallelization means that diversify MapReduce activities can be accomplished. It helps to make the complete benefit of cloud framework properties on request. Task level parallelization meant by that diversify mapper/reducer work is performed. To achieve the scalability in a massive way parallelization of diversify jobs on data partitions is done but in this anonymization effectof the anonymization are not equal. Anonymous data sets are essential for that purpose in between data sets are combined and then complete data set is anonymized.In the first stage a whole complete data set D is separated or partitioned into tiny data sets. The next stage is the job level parallelization is completed. The calculation required in the TPTDS is made by MapReduce version of centralized Top Down specialization, which anonymizes the separated data to produce in between anonymization levels. In this intermediate anonymization level means further specialization can be carry out without disregarding k-anonymity. This method only influences the task level parallelization. In the second level all intermediate anonymization levels are combined into one. The components of ETDS approach are

- Data Partition
- Clustering the Partitioned Data
- Anonymization Level Merging
- Data Specialization

DATA PARTITION

Data partition means whole and complex data D is separated or partitioned into D_i that from this it is easier to say that all the data records of D_i is lesser than D. A data record can be evaluated as point in m dimension space and here m is the number of attributes. Random sampling technique is used for partition the data D.

CLUSTERING THE PARTITIONED DATA

After the partition of the data it is clustered so that it is helpful for the privacy protection of the large scale data or scalability occurs in ease manner, the partitioned data are clustered accordingly so that protecting privacy for that data is done and also large number of data is undergo with this operation.

ANONYMIZATION LEVEL MERGING

In this all the in between anonymization levels are combined into one. The combination of

anonymization levels is completed by merging cuts.

DATA SPECIALIZATION

In this original data set D is specialized in a factual manner by anonymization in a one pass MapReduce job. The Map function displays the anonymous records and also its count. Reduce function perform the task of combining the anonymous records and adds the number.

EVALUATION

EXPERIMENT SETTINGS

Experiment settings focus mainly with the cloud environment which is U-Cloud.U-Cloud is a cloud computing environment at the University of Technology Sydney (UTS). The system overview of U-Cloud is described in Fig. 1. The calculatingabilities of this system are located among several labs at UTS. On the top of hardware and Linux operating system, KVM virtualization software [30] is installed which helps to virtualize the architecture and provides unified computing and repositoryproperties. For the creation of virtualized data centres, OpenStack open-source cloud environment [31] is installed for management in global manner, scheduling the resources and interaction with users. For the creation of virtualized data centres, OpenStack open-source cloud environment [31] is installed for global management, resource scheduling and interaction with users.

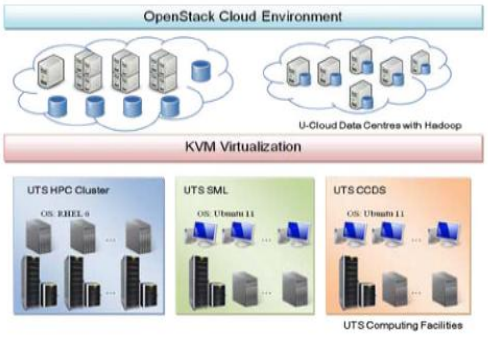

Fig.1. System view of U-Cloud

Further, Hadoop [32] is installed based on the cloud built through OpenStack to simplify enormous data.

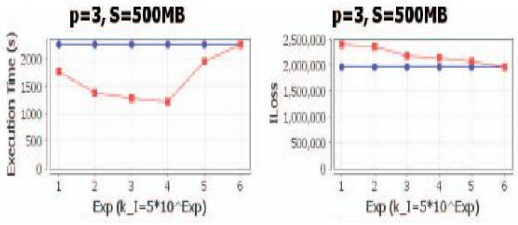

Fig.2. Execution time of ETDS and TPTDS

Experiment results shows that execution time of the TPTDS is more when compared with the EDS this is because in the EDS process partitioned data are clustered so this makes process to be completed in fast manner.

CONCLUSION

Privacy preservation technique only able to handle small scale data to handle large scale data effectively EDS approach is developed in this approach whole and complete data is taken and then it is partitioned next the partitioned data is clustered so this makes large scale data set privacy can be effectively preserved. This EDS approach performs faster than the TPTDS because this EDS approach able to perform with more data because of the clustering of partitioned data is done.

REFERENCES

[1]. S. Chaudhuri, “What Next?: A Half-Dozen Data management Research Goals for Big Data and the Cloud,” Proc. 31st Symp.Principles of Database Systems (PODS ’12), pp. 1-4, 2012.

[2]. M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A.Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M.Zaharia, “A View of Cloud Computing,” Comm. ACM, vol. 53,no. 4, pp. 50-58, 2010.

[3]. L. Wang, J. Zhan, W. Shi, and Y. Liang, “In Cloud, Can Scientific Communities Benefit from the Economies of Scale?,”

IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 2, pp.296-303, Feb.2012.

[4]. H. Takabi, J.B.D. Joshi, and G. Ahn, "Security and Privacy Challenges in Cloud Computing Environments," IEEE Security and Privacy, vol. 8, no. 6, pp. 24-31, Nov. 2010.

[5]. D. Zissis and D. Lekkas, "Addressing Cloud Computing Security Issues," Future Generation Computer Systems, vol. 28, no. 3, pp. 583-592, 2011.

[6]. X. Zhang, C. Liu, S. Nepal, S. Pandey, and J. Chen, "A Privacy Leakage Upper-Bound Constraint Based Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud," IEEE Trans. Parallel and Distributed Systems, to be published, 2012.

[7]. L. Hsiao-Ying and W.G. Tzeng, "A Secure Erasure Code-Based Cloud Storage System with Secure Data Forwarding," IEEE Trans.Parallel and Distributed Systems, vol. 23, no. 6, pp. 995-1003, 2012.

[8]. N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," Proc.IEEE INFOCOM, pp. 829-837, 2011.

[9]. P. Mohan, A. Thakurta, E. Shi, D. Song, and D. Culler, "Gupt: Privacy Preserving Data Analysis Made Easy," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '12), pp. 349-360, 2012.

[10]. B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, vol. 42, no. 4, pp. 1-53, 2010.

[11]. B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and DataEng., vol. 19, no. 5, pp. 711-725, May 2007.

[12]. X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proc. 32nd Int'l Conf. Very Large Data Bases (VLDB'06), pp. 139-150, 2006.

[13]. K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-Domain K-Anonymity," Proc. ACM SIGMOD Int'lConf. Management of Data (SIGMOD '05), pp. 49-60, 2005.

[14]. K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," Proc. 22nd Int'l Conf. DataEng. (ICDE '06), 2006.

[15]. V. Borkar, M.J. Carey, and C. Li, "Inside 'Big Data Management': Ogres, Onions, or Parfaits?" Proc. 15th Int'l Conf. Extending Database Technology (EDBT '12), pp. 3-14, 2012.

[16]. K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Workload-AwareAnonymization Techniques for Large-Scale Data Sets," ACMTrans. Database Systems, vol. 33, no. 3, pp. 1-47, 2008.

[17]. T. Iwuchukwu and J.F. Naughton, "K-Anonymization as Spatial Indexing: Toward Scalable and Incremental Anonymization,"Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB '07), pp. 746-757, 2007.

[18]. J. Dean and S. Ghemawat, "Mapreduce: Simplified Data Processing on Large Clusters," Comm. ACM, vol. 51, no. 1, pp. 107-113, 2008.

[19]. N. Mohammed, B. Fung, P.C.K. Hung, and C.K. Lee, "Centralized and Distributed Anonymization for High-Dimensional Healthcare Data," ACM Trans. Knowledge Discovery from Data, vol. 4, no. 4, Article 18, 2010.

[20]. B. Fung, K. Wang, L. Wang, and P.C.K. Hung, "Privacy-Preserving Data Publishing for Cluster Analysis," Data and Knowledge Eng., vol. 68, no. 6, pp. 552-575, 2009.

[21]. N. Mohammed, B.C. Fung, and M. Debbabi, "Anonymity Meets Game Theory: Secure Data Integration with Malicious Participants,"VLDB J., vol. 20, no. 4, pp. 567-588, 2011.

[22]. L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'lJ. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

[23]. W. Jiang and C. Clifton, "A Secure Distributed Framework for Achieving k-Anonymity," VLDB J., vol. 15, no. 4, pp. 316-333, 2006.

[24]. P. Jurczyk and L. Xiong, "Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers," Proc. 23rdAnn. IFIP WG 11.3 Working Conf. Data and Applications SecurityXXIII (DBSec '09), pp. 191-207, 2009.

[25]. I. Roy, S.T.V. Setty, A. Kilzer, V. Shmatikov, and E. Witchel,"Airavat: Security and Privacy for Mapreduce," Proc. SeventhUSENIX Conf. Networked Systems Design and Implementation (NSDI'10), pp. 297-312, 2010.

[26]. K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan, "Sedic:Privacy-Aware Data Intensive Computing on Hybrid Clouds,"Proc. 18th ACM Conf. Computer and Comm. Security (CCS '11), pp. 515-526, 2011.

[27]. X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc.ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '06), pp. 229-240, 2006.

[28]. Y. Bu, B. Howe, M. Balazinska, and M.D. Ernst, "The HaloopApproach to Large-Scale Iterative Data Analysis," VLDB J., vol. 21, no. 2, pp. 169-190, 2012.

[29]. J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae, J. Qiu, and G. Fox, "Twister: A Runtime for Iterative Mapreduce," Proc.19th ACM Int'l Symp. High Performance Distributed Computing(HDPC '10), pp. 810-818, 2010.