MAP REDUCE TOP DOWN APPROACH FOR SCALABILITY AND ANONYMIZATION

[1]ANJU M SUNNY, [2]MAGNIYA DAVIS
[1]M.Sc STUDENT, [2]ASSISTANT PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE
ST.JOSEPH'S COLLEGE, IRINJALAKKUDA

Abstract— Data is released in a published form for reuse by others, generally known as data publication or publishing. Up gradation of data to be a first class research output is the ultimate goal of this process. Sharing of delicate private data has become a cadre element of research. For privacy preserving and in order to afford increase of user data scalability a broad spectrum of techniques must be enforced, including data anonymization. K-anonymity, l-diversity, t-closeness etc. are the commonly used anonymization techniques for privacy preserving in data sets. In the existing system, generalization is the method used for k-anonymity. But it will not completely anonymize the sensitive data. In this paper, we put forward an integrated approach to anonymize large-scale data sets using the map-reduce framework. Top down Specialization (TDS) is done under the map-reduce framework. The data that is not anonymized is suppressed.

Keywords—Anonymization; MRTDS; Suppression; Generalization

I. INTRODUCTION

For individuals, organizations and agencies data publishing, data sharing has become a routine. Many Organizations and companies collect such data for research or other purposes. Privacy is one of the concerned issues accompanied with published data. Most of the published data have delicate data values which may be misused by the data recipient. Privacy preserving is in the context of preventing information disclosure of the personal data due to use by the legitimate user. If the intimate data like computerized health history and fiscal transaction history are assessed and mined by organizations like Disease Research Centre symbolic assistance are offered. They are usually deemed as extremely delicate. Misuse of such delicate published data by a third party can bring appreciable commercial disaster or harsh social prominence impairment to data owners.

Data anonymization is a prominent approach used for privacy preserving. Various approaches such as perturbation, cryptographical processes are also used for privacy preserving. Data anonymization refers to obscuring integrity or sensitive data for proprietors of data records. K-anonymity is a property possessed by certain anonymized data .A deliverance of data is said to have this property if each record is comparable to at least another k records. When the privacy of the data is preserved the scalability problem of the dataset should also be well addressed. The scale of datasets that need anonymizing is becoming a appreciable confrontation for conventional anonymization algorithms.

In this paper we propose Map Reduce Top down Specialization (MRTDS) approach for addressing the scalability problem of growing datasets. TDS is done based on a Taxonomy Tree i.e., generated manually. First we partition the dataset and then run MRTDS on these partitioned datasets in parallel to extract intermediary anonymization levels. Those intermediate anonymization levels are consolidated. Then run MRTDS on the entire dataset and get the final anonymization level. Then the original dataset is anonymized. As an enhancement we propose the suppression method. The data that are not anonymized is subjected for suppression. This will help in reducing the count of data that is not anonymized.

II. LITERATURE SURVEY

LeFevre et al. introduce scalable decision trees and sampling techniques for addressing the extensibility problem of anonymization algorithms [12]. Iwuchukwu and Naughton by fabricating a spatial index over datasets proposed an R-tree index-based approach which provides high efficiency. However, the above approaches aim at multidimensional generalization, thereby failing to work in the TDS approach [11]. Fung et al. introduced the TDS approach that produces anonymous datasets without the data ascertainment problem. A data structure Taxonomy Indexed PartitionS (TIPS) is exploited to improve the efficiency of TDS. But the approach is centralized, leading to its inadequacy in handling large-scale data sets. Several distributed algorithms are proposed to preserve privacy of multiple data sets retained by multiple parties [4]**.** Jiang and Clifton and Mohammed et al.proposed distributed algorithms to anonymize vertically partitioned data from different data sources without divulging privacy information from one person to another [3]. Jurczyk and Xiongand Mohammed et al.proposed distributed algorithms to anonymize horizontally partitioned data sets possessed by collective holders. However, the above distributed algorithms mainly aim at firmly assimilating and anonymizing multiple data sources [5]. Our research mainly focuses on the scalability issue of TDS anonymization, and is, therefore, orthogonal and complementary to them [10].

As to Map Reduce-relevant privacy protection, Roy et al investigated the data privacy problem caused by Map Reduce and presented a system entitled Airavat blending obligatory access control with differential privacy [7]. Further, Zhang et al.leveraged Map Reduce to automatically partition a computing job in terms of data security levels, protecting data privacy in hybrid cloud [8]. Our research exploits Map Reduce itself to anonymize large-scale data sets before data are further processed by other Map Reduce jobs, arriving at privacy preservation.

III. PROPOSED SYSTEM

A. Generalization

Data set D contain r number of data records. Each data records have m number of attributes. An attribute of a record is denoted as Attr, and the taxonomy tree of this attribute is denoted as TT. This attributes are arranged in taxonomy tree structure. Quasi-identifiers QID representing group of anonymous records. The top most value of the tree is T. When Top-Down specialization is applied to the taxonomy tree structure first it finds the best specialization, and then performs specialization again and finally update values for the next round. Values for the each specialization are analyzed. The highest IGPL value for specialization is regard as the best specialization. In specialization the data sets are split into two phases. The values are updated until k-anonymity.

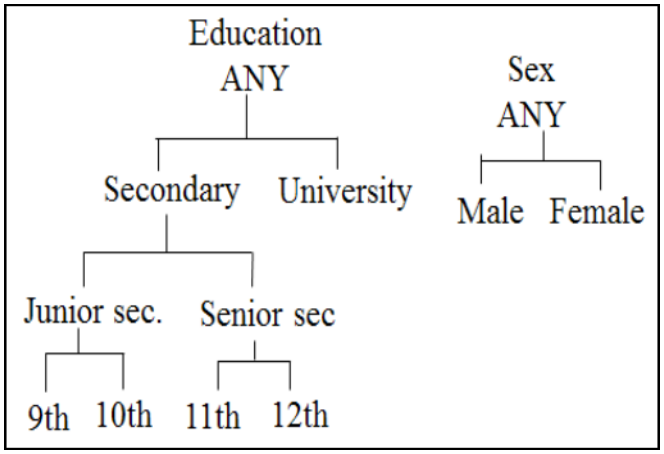

Fig.1. Taxonomy Tree of Attributes [9]

B. Two-Phase Top-Down Specialization (TPTDS)

TPTDS approach mainly has three components, namely data partition, anonymization level merging and data specialization. They are described in the below sections. The two phases of our approach are based on two levels of parallelization, job level and task level. Job level parallelization means that numerous Map Reduce jobs can be executed concurrently. Task level parallelization refers to that several mapper/reducer tasks in a Map Reduce job are executed concurrently over data splits. Our work is mainly based on task level parallelization.

In the first facet, an authentic dataset is partitioned into smaller ones. P is the number of partitions. The details of partition will be discussed in section. We run MRTDS a subroutine over each of the partitioned datasets in lateral. Intermediary anonymization levels are propagated by anonymizing partitioned datasets. An intermediate anonymization level means that further specialization can be performed without violating k-anonymity. MRTDS only leverages the task level parallelization of Map Reduce. Intermediary anonymization levels are the result of our first phase.

In the second facet, all intermediate anonymization levels are merged into one. The whole dataset D is further anonymized based on the merged anonymization level achieving k-anonymity finally. The output of this phase will be the final anonymization level. Ultimately, D is concretely anonymized according to final anonymization level. Algorithm 1 depicts the TPTDS approach.

ALGORITHM 1:TWO-PHASE TDS (TPTDS)

Input: Data set D, anonymity parameters k, ki and the number of partitions p.
Output: Anonymous data set D*.
1: Partition D into Di, $1 \leq i \leq p$
2: Execute MRTDS (Di, ki, AL0) $\rightarrow AL_i'$, $1 \leq i \leq p$ in parallel as multiple Map Reduce jobs.
3: Merge all intermediate anonymization levels into one, merge (AL1', AL2'… ALp') →ALI
4: Execute MRTDS (D, k, ALI) to achieve k-anonymity.
5: Specialize D according to AL*, Output D*.
Steps 1 and 2 comprise the first facet and rest of the steps contribute to second facet.

C. Data Partition

The dissemination of data records in the partition must be identical to the authentic data. Random sampling technique is adopted for partition. Specifically, a random number rand, $1 \leq rand \leq p$, is spawned for each data record. A record is accredited to the partition Drand. The number of Reducers should be equal to p, so that each Reducer handles one value of rand, exactly producing p outcome files. Each file contains an incidental sample of D.

Algorithm 2: data partition map-reduce

Input: Data record (ID_r, r), r∈D, partition parameter p.
Output: Di, $1 \leq i \leq p$.
Map: Generate a random number rand, where $1 \leq rand \leq p$; emit (rand, r).
Reduce: For each rand, emit (null, list(r)).
Once partitioned datasets are obtained, we run MRTDS on these datasets in parallel to acquire intervening anonymization levels.

D. Anonymization Level Merging

All intervening anonymization levels are amalgamated into one in the second phase. The merging of anonymization levels is completed by amalgamating cuts. To assure that the merged intervening anonymization level at no time breaches privacy requirements, the more familiar one is selected as the merged one. MRTDS can further anonymize the absolute data sets to yield final k-anonymous data sets in the second phase.

E. Data Specialization

An authentic data set D is explicitly specialized for anonymization in a one-pass Map-Reduce job. After obtaining the merged intervening anonymization level, we run MRTDS on the entire data set D, and get the final anonymization level. Then, the data set D is anonymized by renewing original attribute values in D with the responding domain values in AL*.The Map function expels anonymous records and its count. The Reduce function simply conglomerates these anonymous records and counts their number. An anonymous record and its count represent a QI-group. The QI-groups constitute the final anonymous data sets.

Algorithm 3: Data Specialization Map & Reduce

Input: Data record (IDr, r), r∈D; Anonymization level AL*.

Output: Anonymous record (r*, count).

Map: Construct anonymous record r*=p1, <p2… pm, sv>, pi, 1≤i≤m, is the parent of a specialization in current AL and is also an ancestor of v_i in r; emit (r*, count).

Reduce: For each r*, sum← ∑count; emit (r*, sum).

F. MRTDS (Map Reduce Version of TDS)

To explicitly conduct estimation it is invoked in both the phases. Generally, a Map-Reduce program consists of Map and Reduce functions, and a Driver that coordinates the macro execution of jobs.

- *MRTDS Driver*

Usually, a single Map-Reduce job is deficient to attain a conglomerated task in many applications. Thus, a group of Map-Reduce jobs are harmonized in a driver program to attain such an objective. MRTDS consists of MRTDS Driver and two types of jobs, i.e., IGPL Initialization and IGPL Update. The driver arranges the execution of jobs.

Step 1 initializes the values of information gain and privacy loss for all specializations, which can be done by the job IGPL Initialization. Step 2 is iterative. First, the best specialization is selected from valid specializations in current anonymization level. A specialization spec is a valid one if it satisfies two conditions. One is that its parent value is not a leaf, and the other is that the anonymity. i.e., the data set is still k-anonymous if spec is performed. Then, the current anonymization level is modified via performing the best specialization i.e., removing the old specialization and inserting new ones that are derived from the old information gain of the newly added specializations and privacy loss of all specializations need to be recomputed, which are accomplished by job IGPL Update. The iteration continues until all specializations become invalid, achieving the maximum data utility.

The iteration of Map-Reduce jobs is controlled by the Anonymization level, AL in Driver. AL is dispatched from driver to mappers and reducers. The value of AL is modified in driver according to the output of the IGPL initialization or IGPL update jobs.

- *IGPL Initialization and Update Job*

The main task of IGPL Initialization is to initialize information gain and privacy loss of all specializations in the initial anonymization level AL. It has both the Map and Reduce functions. The input of the mapper is the data record and the anonymization level. Output is the intermediate key-value pair .i.e. ., key and count. The input of the reducer is the key-value pair .i.e., key and list (count). Output is the information gain and anonymity for all specializations.

IGPL update requires less computation and consumes less network bandwidth. It has both mapper and reducer. Since it executes iteratively it dominates the scalability and efficiency of MRTDS. Compared with IGPL initialization only a part of data is processed and less bandwidth is consumed. Combiner, an optimization technique is used to reduce the communication traffic.

G. Suppression

The data that are not anonymized will be present after all the above work. In order to overcome this problem we are integrating the work of suppression with the above work. The data that breaches the privacy of the personal data, and simultaneously which doesn't satisfies the principle of k-anonymity is subjected to suppression. This technique replaces certain values of the attributes by *(asterisk). The suppression is done till the anonymity value is achieved.

IV. ANALYSIS

TABLE1. SENSITIVE VALUE (SV) ANALYSIS TABLE

1000 Data		
	Generalization	**Generalization+Suppression**
Sv(6)	37.5(625)	54.3(457)
Sv(4)	69.6(304)	83.0(169)

1000 data are taken for the analysis. From this we can see that, when the sensitive value count is 6, 625 data is not anonymized by generalization. But by enhancing it with suppression the count of data that is not anonymized is decreased to 457. When the sensitive value count is taken as 4, 304 data is not anonymized even after generalization. But after the enhancement it is considerably reduced to 169. The difference in anonymized data before suppression and after suppression along with the change in count of sensitive value is indicated in the above table.

V. CONCLUSION

In this paper, we have investigated the privacy preserving problem of large-scale data anonymization by MRTDS, and proposed a suppression technique integrating with the Map-Reduce TDS. We have partitioned the data and parallel anonymized, producing the intermediate results. Then we anonymize the merged intermediate results in order to produce consistent k-anonymous datasets. Then we collected the data that doesn't satisfy the k-anonymity value. They are subjected for suppression. Experimental results on real-world adult dataset have demonstrated that with our approach, the amount of anonymized data.ie, the records satisfying the anonymity value is increased in an efficient way over existing approaches

REFERENCES

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

[2] 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.

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

[4] 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.

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

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

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

[8] 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

[9] D.S Deva Kiruba Dafi, C.Saravanan and C.Kanimozhi, "Two-Phase Top-down Specialization for High Scalability and Privacy Concerns" International Journal of Advanced Research in Computer Science & Technology, Vol. 2 Issue 1 Jan-March 2014.

[10] Xuyun Zhang, Laurence T. Yang, Senior Member, IEEE, Chang Liu, and Jinjun Chen, Member, IEEE" A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud". IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 2, FEBRUARY 2014

[11] 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.

[12] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Workload-Aware Anonymization Techniques for Large-Scale Data Sets," ACM Trans. Database Systems, vol. 33, no. 3, pp. 1-47, 2008